Science Assessment Item Collaborative

Assessment Framework

for the

Next Generation Science Standards

September 2015

Developed by WestEd in collaboration with
CCSSO, state members, and content experts.

THE COUNCIL OF CHIEF STATE SCHOOL OFFICERS

The Council of Chief State School Officers (CCSSO) is a nonpartisan, nationwide, nonprofit organization of public officials who head departments of elementary and secondary education in the states, the District of Columbia, the Department of Defense Education Activity, and five U.S. extra-state jurisdictions. CCSSO provides leadership, advocacy, and technical assistance on major educational issues. The Council seeks member consensus on major educational issues and expresses their views to civic and professional organizations, federal agencies, Congress, and the public.

SCIENCE ASSESSMENT ITEM COLLABORATIVE
ASSESSMENT FRAMEWORK
FOR THE NEXT GENERATION SCIENCE STANDARDS

COUNCIL OF CHIEF STATE SCHOOL OFFICERS
June Atkinson (North Carolina), President
Chris Minnich, Executive Director

Council of Chief State School Officers
One Massachusetts Avenue, NW, Suite 700
Washington, DC 20001-1431
Phone (202) 336-7000
Fax (202) 408-8072
www.ccsso.org

TABLE OF CONTENTS

LIST OF TABLES AND FIGURES

PREFACE

The *Science Assessment Item Collaborative (SAIC) Assessment Framework* ("Assessment Framework") provides a range of options and accompanying rationales for the development of Next Generation Science Standards (NGSS)–aligned items and summative assessments. The Assessment Framework is designed to be used in concert with the *Item Specifications Guidelines* to aid state education agencies (SEAs) and other entities in documenting the processes needed to drive the development of NGSS-aligned items and assessments. Due to the interrelated nature of the documents, elements of the Assessment Framework that specifically detail the characteristics of the assessments and associated development considerations may also appear in the *Item Specifications Guidelines*.

The Assessment Framework principally draws on the following three seminal resources:

- the National Research Council (NRC)'s *A Framework for K–12 Science Education: Practices, Crosscutting Concepts, and Core Ideas* (NRC, 2012), hereafter referred to as the "*K–12 Framework*";

- the *Next Generation Science Standards*: *For States, by States* (NGSS Lead States, 2013), hereafter referred to as the "NGSS"; and

- the NRC Board on Testing and Assessment (BOTA)'s report *Developing Assessments for the Next Generation Science Standards* (NRC, 2014), hereafter referred to as the "BOTA report."

The research-supported recommendations and evidence base for practice that are embodied in these reports are foundational to the approach to development of next-generation science assessments (NGSAs) that is endorsed in the Assessment Framework.

Report Organization

The Assessment Framework is organized into a series of chapters. Each chapter begins with a set of Key Questions, which are intended to serve as advance organizers in focusing the reader's attention on specific points while reading the chapter. At the end of each chapter, a set of Key Takeaways is presented; each set is designed to highlight critical information in that chapter and to help the reader make connections within and across chapters. The following chapters are included in the Assessment Framework:

Chapter One: Overview provides background information on the SAIC and the Assessment Framework, including the purpose of the Assessment Framework and the context for its development. This overview includes an introduction to the NGSS and a description of how state stakeholders provided input for the development of the Assessment Framework. The chapter ends with a discussion of an approach to assessment validation.

Chapter Two: Content details the sources of content targeted in the Assessment Framework. Included is background information about the NGSS, the *K–12 Framework*, and the evidence statements (NGSS Network, 2015a, 2015b) for the elementary, middle, and high school grade bands. The processes used to develop these documents are summarized, and the steps that will be used to specify the content to be assessed are discussed.

Chapter Three: Measurement Model presents guidance on measuring proficiency within a specific content domain through the application of an appropriate measurement model. The importance of identifying the measurement model underlying both item and assessment development is explained as a key step in validation of assessment results. Important considerations regarding high-priority elements of evidence for the emerging NGSS-based assessments are discussed.

Chapter Four: Item Types and Clustering focuses on how the proposed architecture and grouping of items into clusters are designed to elicit the necessary evidence from students to support the identified claims. The range of eligible item types for inclusion in the NGSS-aligned assessments is described. Model items are cited and evaluated for the purposes of informing development of the Assessment Framework, and the subsequent evolution of eligible item types is explored.

Chapter Five: Item Specifications Guidelines provides a brief overview of the components included in the stand-alone Item Specifications Guidelines document, which is designed as a separate deliverable to be used by states to support NGSS-aligned assessment development. The considerations underlying the guidelines are highlighted, and SAIC member input on the desired outcomes for the Item Specifications Guidelines is summarized.

Chapter Six: Developing Blueprints articulates the recommended processes to be used for developing blueprints for the overall design of assessments. Guidance is provided on how to determine (a) the appropriate number and types of items per grade; (b) the distribution of items and points across PEs and NGSS dimensions (Science and Engineering Practices [SEPs], Disciplinary Core Ideas [DCIs], and Crosscutting Concepts [CCCs]); (c) the appropriate levels of cognitive demand; and (d) the right balance between breadth and depth with respect to content. Logistical considerations relative to each of these issues (e.g., test length, intended and unintended consequences of the assessment) are included.

Chapter Seven: Item Cluster Development provides information about the plan for developing sufficient numbers and types of items to build the assessments per the blueprint specifications. Steps in this process include identifying science phenomena, detailing criteria for stimuli development, effectively integrating internal and external reviews, piloting items, and designing a feedback loop to inform future development. Best practices related to assessment development with an emphasis on NGSS-specific issues are described.

Chapter Eight: Accessibility Considerations presents guidelines for ensuring accessibility at critical points in the design, development, and administration of the proposed NGSS-aligned assessments. Emphasis is placed on the application of the principles of universal design and

the use of accommodations. Accessibility considerations and recommendations are provided for general-education students as well as for students with disabilities (SWDs) and English learners (ELs). A discussion of accessibility concerns related to specific item types is included.

The **Appendices** include a detailed description of the types of measurement evidence recommended for collection and evaluation during test design and development, field testing, test administration, scoring, and reporting and interpretation of scores (Appendix A); responses from surveys designed to elicit SAIC member feedback (Appendices B, C, and D); annotated resources for the development of assessments for the NGSS (Appendix E); example assessment items from SimScientists and the National Assessment of Educational Progress (NAEP) (Appendix F); and a description of the processes employed during the development of the Assessment Framework, accompanied by a timeline for the development (Appendix G).

CHAPTER ONE: OVERVIEW

Introduction

Key Questions for this Chapter

- What is the purpose of the Assessment Framework?
- What are the elements of an effective assessment framework?
- How was the Assessment Framework developed?
- How will the Assessment Framework support collection of evidence about test validity?

A new approach to K–12 science education was presented in the National Research Council (NRC)'s *A Framework for K-12 Science Education: Practices, Crosscutting Concepts, and Core Ideas* (NRC, 2012). The *K–12 Framework* articulates a broad set of rigorous expectations to support all students in achieving scientific literacy, and provided guidelines on how to prepare students to be able to pursue science, technology, engineering, and mathematics (STEM) careers. The *K–12 Framework* organizes science learning around three main dimensions: Science and Engineering Practices (SEPs), Crosscutting Concepts (CCCs), and Disciplinary Core Ideas (DCIs). It emphasizes that these three dimensions must be interwoven into every aspect of science education—curriculum, instruction, and assessment—in order to fully achieve the vision set forth for science education.

The *K–12 Framework* was subsequently used as the research-supported foundation for the development of the NGSS. The NGSS and the *K–12 Framework* present a holistic approach to science education in maintaining that students must, in both instruction and assessment, "engage in scientific and engineering practices in the context of disciplinary core ideas, and make connections across topics through the crosscutting ideas" (NRC, 2014, p. 4). This multidimensional approach to science education presents both opportunities and challenges for states as they begin to implement the rigorous new standards and develop NGSS-aligned assessments.

As of June 2015, 14 states and the District of Columbia have fully or partially adopted the NGSS. Following adoption of the NGSS, states began the process of incorporating the new science standards into classroom instruction and developing measures for monitoring what students know and can do with respect to these new standards. These states recognized the value of leveraging resources through collaboration with other states and entities that were facing similar challenges related to the design and development of NGSS-aligned assessments. With support from the Council of Chief State School Officers (CCSSO), the SAIC was established, and as of June 2015, 14 states and one territory comprise this collaborative.

An Assessment Framework

A primary purpose of the SAIC is to support states in the development of a pool of high-quality items for large-scale summative assessment. To achieve this goal, the SAIC is initially

developing guidance documents outlining a systematic, methodical, and research-based approach to the design and development of NGSS-aligned summative assessments. This approach begins with the development of an assessment framework, aimed at state science assessment coordinators and assessment developers, and serving as a bridge between the NGSS and methods of assessing those standards.

An assessment framework defines and clarifies what knowledge, skills, and abilities are embodied in the standards to be assessed, and establishes a common understanding of the priorities for assessment at each grade level or grade band. This Assessment Framework is written to inform a collaborative of states. As such, it provides general guidance, and states may use it to craft a state-specific assessment framework.

An effective assessment framework includes the following elements:

- ***Detailed description of the content (knowledge and skills) to be assessed.*** The assessment framework should clearly describe the specific content targeted for assessment. Please see **Chapter Two** for detailed information about the content of NGSS-based assessments.
- ***Specification of the content that is eligible for assessment at each grade level or grade band***. Key stakeholder groups can engage in prioritizing the most important standards as foundational for student understanding at different grade levels or spans. This topic is discussed in **Chapter Two**.
- ***Information about* how *content will be assessed, including a description of the characteristics of item types that will be used to measure the content in the domain***. Teachers place greater value on assessment results when a test is synchronized with what is being taught and how instruction is delivered. Strategies for ensuring linkages among classroom practice, curricula, instructional methods, and assessment are discussed in **Chapters Three, Four,** and **Five**, along with guidance for selecting appropriate item types for NGSS-based assessments and determining the specifications for those items.
- ***A blueprint for developing a test or item pool that meets the specified assessment objectives***. Blueprint developers must ensure that the full depth and breadth of the NGSS are assessed. This will include balancing the need for specific item types (e.g., authentic performance-based tasks) with the constraints of large-scale assessment. Please see **Chapter Six** for an in-depth description of recommended steps in the development of blueprints for NGSS-based assessments.
- ***Guidelines for administering and scoring the assessment***. Developers of NGSS-based assessments will benefit from recent advances in the use of technology for the purposes of scoring and reporting. The appendices of this report include useful information for state staff responsible for these phases of work.

Development of an effective assessment framework for large-scale summative assessments and implementation of the steps called for in such a framework require systematic

documentation of options considered, tradeoffs weighed, and decisions reached during all phases of test design and use. Transparency and replicability are critical for promoting trust with stakeholders and ensuring that the recommended types of evidence are collected to support the validity of inferences drawn from results of the assessment that emerges from this work.

It is anticipated that states will use the Assessment Framework for a number of purposes. Along with the *Item Specifications Guidelines*, the Assessment Framework will be the guiding document to inform the development of requests for proposals (RFPs) that will be used to select and guide assessment vendors in the development of NGSS-aligned assessments. It will also serve as a guiding document for the development of state and local test specifications and blueprints. The Assessment Framework may also be a valuable communication tool, providing information to key stakeholders and professional development providers.

Foundation for the NGSS-Based Assessment Framework

A number of important developments in K–12 science education in the United States have taken place in recent years:

- Publication of the *K–12 Framework* (NRC, 2012). This seminal document presented a new approach to K–12 science education and informed the development of the NGSS.
- Development of the NGSS. Through a state-led process, new K–12 science standards were developed to be "rich in content and practice and arranged in a coherent manner across disciplines and grades to provide all students an internationally benchmarked science education" (NGSS Lead States, 2013).
- Publication of the BOTA report (NRC, 2014). This influential report explores the complexities associated with building an assessment system to support the NGSS and provides recommendations and conclusions to that end.
- Release of the NGSS evidence statements (NGSS Network, 2015a, 2015b). These statements are aimed at providing the necessary detail on what can be used as evidence of what students know and are able to do for each performance expectation (PE).

Additional developments predating the most recent efforts, as described in the *Science Assessment and Item Specifications for the 2009 National Assessment of Educational Progress* (NAGB, 2007), also have helped set the stage for the development of the Assessment Framework:

- Publication, for the first time, of national standards for science literacy in *National Science Education Standards* (NRC, 1996) and *Benchmarks for Science Literacy* (AAAS, 1993). Since their publication, these two national documents have informed state science standards.
- Advances in cognitive research (e.g., on how students learn increasingly complex material over time) that have yielded new insights into how and what students learn

about science (NRC, 1999, 2001, 2005). For example, there is new information about what is appropriate for students to learn at various grades (Catley, Lehrer, & Reiser, 2005; Metz, 1995; Smith, Wiser, Anderson, Krajcik, & Coppola, 2004).

- Growth in the prevalence of science assessments nationally and internationally, including the requirements in the current federal education legislation, commonly known as the No Child Left Behind Act, for science assessment starting in 2007; the ongoing Trends in International Mathematics and Science Study (TIMSS) (Mullis et al., 2001; see http://timss.bc.edu/); and the Programme for International Student Assessment (PISA) (OECD, 2005; see http://www.pisa.oecd.org/). These assessments, including frameworks, curricular analyses, and reports on results, are increasing the visibility of, and interest in comparing and contrasting, student science performances. For example, a recent TIMSS report (Gonzales et al., 2004) lists ten National Center for Education Statistics (NCES) publications related to TIMSS (see http://nces.ed.gov/timss/) and 16 additional publications available through Boston College's TIMSS International Study Center (see http://timss.bc.edu/).
- Growth in innovative assessment approaches that probe students' understanding of science in greater depth than before (e.g., clusters of items tapping students' conceptions of the natural world), sometimes with the use of computer technology (Hilton & Honey, 2011).
- Increased inclusion of formerly excluded groups in science assessments (e.g., SWDs and ELs), requiring a new assessment to be as accessible as possible and also to incorporate accommodations so that the widest possible range of students can be fairly assessed (Lee, 1999).

Finally, it is important to note that these objectives and the overall importance of the development of an assessment framework are emphasized in a key resource for all test developers and consumers, the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014).

Assessment Framework Development Process

Development of the NGSS-based Assessment Framework described in this report was led by the SAIC, with WestEd as the primary author. This effort entailed a comprehensive state survey, multiple rounds of member and expert reviews, and strategic refinement of the emerging recommendations. Members of Achieve Inc.—including Dr. Stephen Pruitt, Senior Vice President of Achieve Inc., who coordinated development of the NGSS—and other experts in assessment design and psychometrics provided valuable feedback on drafts of the Assessment Framework and provided consultation during its development. Further details on the development process can be found in Appendix G.

The members of the SAIC are a diverse group of states and other jurisdictions. Some are members of one of the two major Race to the Top assessment consortia (the Smarter Balanced Assessment Consortium [Smarter Balanced] and the Partnership for Assessment of Readiness

for College and Careers [PARCC]); others do not participate in either consortium. Some members adopted the NGSS by name, while others adopted the full NGSS but with a name change or opted for a partial adoption. Some members plan for a full computer-based administration of NGSS-based assessments, while others plan to use a mix of computer-based delivery and paper-and-pencil delivery. Finally, some members fully embrace the recommendations in the BOTA report (NRC, 2014), while other members support a more state-mediated approach to the transition to the NGSS. Despite these differences, all members of the SAIC have worked together to achieve the goals that they established as a team.

Approach to Validation

Test validation requires the ongoing, systematic collection and evaluation of information to support claims about the trustworthiness of inferences drawn from the results of that assessment (Kane, 2002; Mislevy, 2007; Sireci, 2007). This work is undertaken in accordance with the guidelines for responsible testing practices specified in the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014). The validation process involves the strategic collection and review of a comprehensive body of evidence to support use of a measure for a particular purpose. Validation is intended to help developers build a logical, coherent argument for test use that is supported by particular types of evidence.

The different types of evidence are collected on an ongoing basis, through all phases of design, development, and implementation of the assessment. Specifically, researchers engaged in this work are seeking evidence about test validity (construct, content, predictive, and consequential), reliability (precision and consistency/stability), fairness, and feasibility (resources required, benefit-cost considerations). Information related to these broad categories of evidence can be collected from a number of different sources, including documentation of design decisions (e.g., test purpose, theory of action, target population); steps taken during the various phases of development, including pilot and field testing (e.g., recommendations from bias and sensitivity reviews, findings from alignment studies, inter-rater reliability statistics, test blueprints, item-level statistics); administration and scoring guides; sample score reports; technical manuals; and research reports.

Importantly, a test is not considered “validated” at any particular point in time; rather, it is expected that developers gather information at each stage of design, development, and use in thoughtful and intentional ways, with the gradual accumulation of a body of specific types of evidence that can be presented in defense of test use. As the process of designing the NGSS-based assessments is already underway, states should collect historical documentation that describes the vision for this initiative and the steps taken, to date, to identify appropriate experts, vendors, research organizations, and key stakeholder groups to undertake the various planned steps. As the validation process continues, additional information can be gathered at key points in time, with the goal of purposefully gathering input and feedback from diverse audiences at each stage of work; collecting evidence to verify that appropriate steps were undertaken to ensure the technical quality, fairness, and feasibility of the measures;

documenting tradeoffs weighed and decisions made; and promoting transparency in all steps in order to build stakeholder trust in using the results from the emerging assessments.

The Assessment Framework is intended as a starting point for building a validity argument. It represents the combined thinking of state representatives and content and technical experts as to how to design large-scale summative assessment items and tests that address the full potential for scientific teaching and learning embodied in the NGSS.

Context of the Assessment Framework

The primary focus of this Assessment Framework is to build a basis of item development for NGSS large-scale assessment within the context of overall test design. The Assessment Framework should be considered a starting point for the implementation of a large-scale assessment measuring the NGSS, rather than being considered the final model. It should be noted that the item cluster model presented in the Assessment Framework has not been developed and fully implemented in a state testing system for science, although significant parts of it have. Lessons learned through large-scale development will present opportunities to adjust the model presented and tools recommended. The descriptions and expectations presented in the Assessment Framework should be considered a starting point, rather than the definitive end product. In addition, there are psychometric challenges that will need to be addressed (and limits pushed) for tests built using the item cluster model as the basic building component. These issues include acceptable content coverage, pilot testing, score generalizability, and number of score points to achieve reporting expectations. Matrix sampling is considered an important test design consideration for achieving a reasonable amount of content coverage and for achieving aggregate level reporting at the school, district, and state levels. In addition, reporting for the individual student for anything other than overall science ability will be problematic to support using only item clusters. Even for overall science ability at the individual student level, individual reliability of scores may not be as strong as is achievable with a test composed primarily of individual items. The acceptable limits for the described concerns will need to be addressed and determined by individual states through their development and implementation efforts.

The Assessment Framework's focus on large-scale assessment was an outcome of needs expressed by states to begin the conversation about how to develop such an assessment while still being true to the principles and expectations of the *K–12 Framework* (NRC, 2012), the NGSS (NGSS Network, 2015b), and the BOTA report (NRC, 2014). The presentation of this Assessment Framework in no way espouses the use of a single test in isolation to measure and report on the full NGSS. The BOTA report (NRC, 2014) describes a comprehensive assessment program approach. Concerns previously described in this section can be lessened if the assessment is used within the context of an assessment system that provides (through other assessments) information with greater usability and generalizability at the school level.

This Assessment Framework is not intended to provide a full assessment solution for states. Its intent is to present an acceptable solution for achieving alignment to the NGSS for large-scale assessment. Many lessons remain to be learned as this solution is pursued.

Key Takeaways from Chapter One

- The NGSS, based on the *K–12 Framework*, pose a number of challenges, as well as opportunities, for the development of assessments based on these standards.
- The Assessment Framework is rooted in advances in cognitive research and assessment research and best practices in the assessment industry.
- The Assessment Framework is intended to serve as a bridge between the NGSS and effective assessment of those standards.
- Validation is an ongoing process that requires the systematic gathering of evidence about test validity, reliability, fairness, and feasibility from a broad range of sources.
- The Assessment Framework is a guiding document for states to use in developing their own state-specific assessment frameworks.

CHAPTER TWO: CONTENT

Introduction

Key Questions for this Chapter

- How is the content of the NGSS organized?
- How are assessment boundaries addressed in the design and development of summative assessments?
- How do evidence statements inform assessment development?

To ensure that assessment purpose and intended outcomes are clear to stakeholders, the content to be assessed must be made explicit for item and test design purposes. This important work should communicate what is to be assessed (e.g., eligible content), how tests should be written to measure the content, and how proficiency will be defined relative to that content (e.g., defining assessment targets and constructs). In addition, states or districts will need to identify and disseminate clear learning expectations based on the full depth and breadth of the NGSS to ensure that all functional components, inclusive of both instruction and assessment, remain synchronized and aligned. This work is the focus of this chapter.

Specification of the content of the NGSS for assessment will allow states to achieve clear and focused assessment targets that will be subsequently leveraged to translate the grade-level or grade-band PEs into PE item specifications. Specifications for items, tasks, and test blueprints will be established to satisfy the multidimensional nature of the NGSS. Defining the assessment evidence for each PE in each grade level or grade band will provide the necessary item and task specificity and clarity to make transparent the connections between assessment outcomes and the preceding classroom instructional process. While it is most convenient for item, task, and blueprint specifications to reflect the grade levels and grade bands specified in the NGSS, it should be noted that the Assessment Framework can also be applied with some variability, allowing for flexibility for states. Additionally, the NGSS endpoints provide a set of initial hypotheses about the progression of learning for the SEPs, CCCs, and DCIs (NRC, 2014); these hypotheses can support instruction and assessment within grade bands, depending on how the Assessment Framework is applied.

This chapter also seeks to reinforce the importance of considering the degree to which each state's adopted standards match the NGSS and influence the content specification process. While the Assessment Framework and its supporting documentation were developed assuming its use in relation to the full NGSS, the principles presented within it can be applied to most variations on the standards adopted by individual states.

Organization of the NGSS: Multiple Dimensions

In the *K–12 Framework* (NRC, 2012, p. 1), some of the known limitations facing K–12 science educators in the United States are called out, specifically that science education:

- "is not organized systematically across multiple years of school,"
- "emphasizes discrete facts with a focus on breadth over depth," and
- "does not provide students with engaging opportunities to experience how science is actually done."

These comments suggest that the *K–12 Framework* introduces a paradigm shift in science education in that three explicit dimensions integral to the new science standards and science learning are introduced: Science and Engineering Practices (SEPs), Disciplinary Core Ideas (DCIs), and Crosscutting Concepts (CCCs). Explicit goals for science learning are outlined in the NGSS in the form of performance expectations (PEs). PEs are statements about what students should know and be able to do at the end of instruction (for specific grades in K–5 and for grade bands in middle school and high school), and the PEs prominently incorporate all three dimensions. The associated specific dimensions for each PE are further identified through the NGSS foundation boxes. The sections that follow summarize information about these three dimensions as well as specific considerations related to content specifications for each dimension.

Science and Engineering Practices

The SEPs are statements derived from and grouped by the eight categories detailed in the *K–12 Framework* that explain the science and engineering practices that attempt to engage students in the important practices used by scientists and engineers. More specifically, instructional and assessment tasks must endeavor to engage students in these practices in the context of a core idea when making connections to crosscutting concepts (but should never be taught or assessed in isolation).

The SEPs are:

1. Asking questions (for science) and defining problems (for engineering)
2. Developing and using models
3. Planning and carrying out investigations
4. Analyzing and interpreting data
5. Using mathematics and computational thinking
6. Constructing explanations (for science) and designing solutions (for engineering)
7. Engaging in argument from evidence
8. Obtaining, evaluating, and communicating information

More information about the SEPs is provided in Appendix F of the NGSS, in which guiding principles, rationales, progressions, and further clarity are provided for each of the eight SEPs.

Disciplinary Core Ideas

The organization of the NGSS is structured around core ideas in the major fields of natural science as outlined in the *K–12 Framework*, with the addition of one set of PEs focused on Engineering, Technology, and Applications of Science (ETS). The *K–12 Framework* includes eleven core ideas—four in Physical Sciences, four in Life Sciences, and three in Earth and Space Sciences—and the core ideas are further divided into a total of 39 sub-ideas. Each sub-idea includes a list of what students should understand at the end of instruction for grades 2, 5, 8, and 12; these lists are the DCIs in the NGSS.

The DCIs are organized into the following 12 nodes:

Physical Sciences

- PS1: Matter and its Interactions
- PS2: Motion and Stability: Forces and Interactions
- PS3: Energy
- PS4: Waves and their Applications in Technologies for Information Transfer

Life Sciences

- LS1: From Molecules to Organisms: Structures and Processes
- LS2: Ecosystems: Interactions, Energy, and Dynamics
- LS3: Heredity: Inheritance and Variation of Traits
- LS4: Biological Evolution: Unity and Diversity

Earth and Space Sciences

- ESS1: Earth's Place in the Universe
- ESS2: Earth's Systems
- ESS3: Earth and Human Activity

Engineering, Technology, and Applications of Science

- ETS1: Engineering Design

The NGSS endeavor to increase coherence in K–12 science education, and a summary DCI progression for the grade-band endpoints is provided in Appendix E of the NGSS. (Please note that the full progression can be found in the *K–12 Framework*.) The upper and lower limits of alignment expectations related to the NGSS for specific PEs and evidence statements should be informed by these learning progressions.

Crosscutting Concepts

The CCCs provide a way of linking the different domains of science and have application across all domains of science. The CCCs, as described in Appendix G of the NGSS, are:

1. **Patterns.** Observed patterns of forms and events guide organization and classification, and they prompt questions about relationships and the factors that influence them.
2. **Cause and effect: Mechanism and explanation.** Events have causes, sometimes simple, sometimes multifaceted. A major activity of science is investigating and explaining causal relationships and the mechanisms by which they are mediated. Such mechanisms can then be tested across given contexts and used to predict and explain events in new contexts.

3. **Scale, proportion, and quantity.** In considering phenomena, it is critical to recognize what is relevant at different measures of size, time, and energy and to recognize how changes in scale, proportion, or quantity affect a system's structure or performance.
4. **Systems and system models.** Defining the system under study—specifying its boundaries and making explicit a model of that system—provides tools for understanding and testing ideas that are applicable throughout science and engineering.
5. **Energy and matter: Flows, cycles, and conservation.** Tracking fluxes of energy and matter into, out of, and within systems helps one understand the systems' possibilities and limitations.
6. **Structure and function.** The way in which an object or living thing is shaped and its substructure determine many of its properties and function.
7. **Stability and change.** For natural and built systems alike, conditions of stability and determinants of the rate of change or evolution of a system are critical elements of study.

As previously indicated, core information about the CCCs is provided in Appendix G of the NGSS, in which guiding principles, rationales, progressions, and further clarity are provided for each of the seven CCCs. The *K–12 Framework* emphasizes that much work has been done illustrating the importance of CCCs in science education and that these concepts are featured prominently in many documents (albeit referred to by different names), including *Science for All Americans* (AAAS, 1989), *Benchmarks for Science Literacy* (AAAS, 1993), *National Science Education Standards* (NRC, 1996), and *Science Anchors Project* (NSTA, 2010). Due to the need for each CCC to be revisited numerous times during the learning progression and never assessed separately from the other dimensions, the CCCs may appear to be an elusive assessment target. Appendix G of the NGSS asserts that "students should be assessed on the extent to which they have achieved a coherent scientific worldview by recognizing similarities among core ideas in science or engineering that may at first seem very different, but are united through crosscutting concepts" (p. 3).

Due to the interwoven nature of these three dimensions, assessment developers must determine how best to "provide evidence of students' ability to use the practices, to apply their understanding of the crosscutting concepts, and to draw on their understanding of specific disciplinary ideas, all in the context of addressing specific problems" (NRC, 2014, p. 32). The *K–12 Framework* (NRC, 2012) makes an important distinction in the use of the word "practices" instead of "skills" and how this pertains to the interwoven nature of the three dimensions:

> *We use the term 'practices' instead of a term such as 'skills' to emphasize that engaging in scientific investigation requires not only skill but also knowledge that is specific to each practice.* (p. 30)

Subsequent chapters of this report offer recommendations for accomplishing the goal of assessing all three dimensions of science. Information is shared about traditional item types (selected and constructed response) and building assessments that satisfy the multidimensional aspect of the NGSS. The task of designing assessments to explicitly elicit evidence of multidimensional science learning begins with a purposeful approach to defining the content.

Assessment Boundaries

Use of an evidence-based approach (discussed in more detail in **Chapter Three**) requires exploration of how specific elements that are inherent in the structure of both the NGSS and the *K–12 Framework* provide solid foundations for the following steps in the design and development of assessments:

1. Define the content domain to target for assessment.
2. Specify the content to be assessed (constructs).
3. Develop claims (inferences) about student proficiency.
4. Determine the appropriate observations (evidence) to support the claims.
5. Design the tools (tasks) for eliciting this evidence.

As previously noted, developers of the NGSS have begun the work of defining the content domain (Steps 1 and 2) and developing claims about student proficiency (Step 3) by explicitly organizing the standards around PEs inclusive of all three dimensions. Further, included in the NGSS for each PE are assessment boundary statements that "specify the limits to large scale assessment." The NGSS Lead States (2013) provide the following explanation:

> *The NGSS architecture was designed to provide information to teachers and curriculum and assessment developers beyond the traditional one line standard. The performance expectations are the policy equivalent of what most states have used as their standards.* (p. 2)

Although the NGSS developers made significant strides in defining the content domain and specifying the knowledge and skills to be assessed at grades K–5 and for each of the middle school and high school grade bands, states may want to engage in further discussion about the content to be assessed on the summative assessment. What are the state's priorities for assessment at each grade? Factors that will contribute to this decision include the following:

- the overall architecture of the assessment system (e.g., formative, interim, summative, large scale, classroom based),
- decisions made at state and local levels as to the priorities for curriculum and instruction,
- the scope and purpose of the assessment system,
- fiscal implications, and
- the vision of the *K–12 Framework* and the NGSS, of ensuring that all students are supported in reaching all standards, K–12.

Both the BOTA report (NRC, 2014) and *Systems for State Science Assessment* (NRC, 2006) emphasize the need for a full system of assessments that includes multiple approaches (e.g., large-scale and classroom-based) to meet a range of purposes (e.g., to guide instruction, for program evaluation, or to test achievement) in a cohesive manner. Depending on a state's assessment infrastructure, information from various approaches may prove effective in assessing the depth and breadth of the NGSS.

A number of topics that are summarized in this chapter are discussed in greater detail in later chapters. For example, the role of grouping PEs is further explored in **Chapters Four** and **Seven**; **Chapter Seven** also provides more information on the use of evidence statements in item cluster design. **Chapter Five** provides guidelines for item specifications.

Key Takeaways from Chapter Two

- The goals of the NGSS are outlined as performance expectations that must be translated into specifications documents to support the development of items, tasks, and test blueprints.
- The challenge posed to assessment developers is how to "provide evidence of students' ability to use the practices, to apply their understanding of the crosscutting concepts, and to draw on their understanding of specific disciplinary ideas, all in the context of addressing specific problems."
- Priorities for assessment at each grade level or grade band will be based on the overall architecture of the assessment system (formative, interim, and summative), state priorities for curriculum and instruction, the scope and purpose of the assessment system, and fiscal implications, as well as on the need to adhere as closely as possible to the vision of the NGSS and the *K–12 Framework* in supporting all students in achieving all standards.
- Evidence statements have been released for all grade levels and grade bands, and are expected to be useful in designing tools for eliciting evidence.

CHAPTER THREE: MEASUREMENT MODEL

Introduction

Key Questions for this Chapter

- What activities guide the pairing of evidence with claims about student proficiency?
- What are the high-priority elements of evidence?

An evidence-based approach provides a rigorous process for collecting the evidence needed to support the validity of inferences intended to be drawn from results from the emerging NGSS-aligned science assessment. Over the past decade, this approach has been used to guide both traditional item development activities and the design of innovative item types that build on technology-enhanced features. Further, an evidence-based approach is an extension of evidence-centered design, which is one of two research-based approaches recommended in the BOTA report on developing assessments for the NGSS (NRC, 2014).

Important goals of the evidence collection process are to define the explicit claims that developers seek to make and to match these claims with evidence of learning; in this way, a system of claim-evidence pairs guide the development of an assessment instrument. This is accomplished through five distinct yet interrelated activities: domain analysis, domain modeling, conceptual assessment validity framework, implementation, and delivery. A brief description of the process involved in each of these activities is provided in Table 1 (adapted for this context from Gorin & Mislevy, 2013).

Table 1. Five Activities Guiding Pairing of Evidence with Claims

Processes in Evidence-Based Approach	
Activity	***Description***
Domain Analysis	Determine the specific content to be measured, as set forth in the NGSS
Domain Modeling	Determine, at a high level, the components of the assessment system
Conceptual Assessment Validity Framework	Determine the claim-evidence pairs to be assessed (constructs), to be defined in the output of the Content Specifications and Item Specifications
Implementation	Develop the assessment items/tasks
Delivery	Determine the processes for assessment administration and reporting

The initial analysis of all NGSS PEs in connection with their associated dimensional elements (SEPs, DCIs, and CCCs) constitutes a part of the domain analysis. The domain analysis comprises a comprehensive evaluation of assessable content and identifies the appropriate item and task types for each PE item cluster and the targeted depth of knowledge appropriate for the measure.

The specific outputs of these targeted activities include science item specifications documentation, sample items, style guidelines, stimulus specifications, technology-enhanced item (TEI) specifications, functional HTML prototypes, performance task specifications, bias and sensitivity guidelines, and accessibility and accommodations guidelines. As recommended by the BOTA report (NRC, 2014), a system of assessments, from formative classroom assessment through summative standardized assessment, will best support the approach of the NGSS. A "from the bottom up" approach that connects test developers with teachers and curriculum specialists to provide a consistent and continuous path from classroom instruction through summative testing will yield the desired results. Ideally, this begins with the process of designing assessments for the classroom, perhaps integrated into instructional units, and moves toward assessments for monitoring.

State education agency staff, content specialists, and teachers are needed to support this approach and to provide continuity across all components of the assessment system. Stakeholder input, including outreach during the public review of content specification and the draft blueprint, will be integral to this effort. Multiple review periods and an open public review period may be required during the pilot development phase, to allow for ample opportunity to elicit input and feedback during the iterative development process. Processes used during the test development process should be transparent, and decisions made should be inclusive of the full range of stakeholders.

Attending to lessons that have been learned through similar initiatives can be valuable. Development of the College Board's Advanced Placement (AP) Insight product, released in response to the revised assessments for AP Biology, AP Chemistry, and AP World History (College Board, 2011), parallels many of the challenges faced by developers of NGSS-based assessments. The College Board emphasizes the integration of core scientific ideas with science practices in the development of both the summative science exams and the formative AP Insight products. In collaboration with WestEd, it recruited subject-matter experts and AP teachers to design item specifications that support the development of scaffolded assessment of multidimensional science learning (both content and science practices) in the context of real-world applications of science. This work is cited in Appendix C of the NGSS as one of the first clear examples of successful implementation of this type of integration. **Chapter Seven** provides guidance on how to approach use of scaffolding within the context of developing item clusters.

Key Component of Measurement Model: Evidence of Technical Quality

Developers of the new NGSS-based assessments are likely to benefit from guidance about key questions to consider about the technical quality of the emerging measures. As described previously, to support developers' claims that an assessment is valid for a particular purpose, specific types of evidence must be collected in relation to the assessment's technical quality (validity, reliability), fairness for all students, and feasibility in the given context. This need is reinforced by the BOTA report (NRC, 2014):

> *Recommendation 3-1 To ensure that assessments of a given performance expectation in the Next Generation Science Standards provide the evidence necessary to support the intended inference, assessment designers should follow a systematic and principled approach to assessment design, such as evidence centered design or construct modeling. In so doing, multiple forms of evidence need to be assembled to support the validity argument for an assessment's intended interpretive use and to insure equity and fairness.* (p. 81)

Appendix A provides a detailed description of the types of measurement evidence recommended for collection and evaluation during test design and development, field testing, test administration, scoring, and reporting of scores. It also provides operational definitions and guidelines for data collection and documentation processes. This information will be useful for states seeking a summary of the types of technical evidence that they may want to consider collecting during each phase of work.

In the context of the measurement models for NGSS-based assessments, the following types of information are viewed as "first-tier" evidence, or those elements of evidence that are of the highest priority:

- *Construct validity*: Documentation of the purpose of the assessment, target population, and how results will be used; evidence that suggests that the assessment measures what it is intended to measure.
- *Content validity*: Information about the nature and degree of alignment to the full depth and breadth of the NGSS at the item and assessment levels. Because of the nature of the NGSS, this evidence must link item clusters to measurement claims for PEs and the three dimensions of the NGSS.
- *Consequential validity*: Documentation of intended and unintended consequences (positive and negative) of testing.
- *Reliability*: Evidence that the assessment meets industry standards for precision and stability/consistency; information about inter-rater reliability of hand scorers.
- *Fairness*: Information about the appropriateness of the assessments for the target population, especially for identified student subgroups.
- *Feasibility*: Information about the human and financial resources required to administer the assessments; documentation of testing time and estimated burden to students, teachers, and schools.

It should also be noted that the BOTA report (NRC, 2014) recommends that states consider using a system of assessments in achieving their measurement goals for an NGSS-aligned assessment:

> *CONCLUSION 6-1 A coherently designed multilevel assessment system is necessary to assess science learning as envisioned in the framework and the Next Generation Science Standards and provide useful and usable information to multiple audiences. An assessment system intended to serve accountability purposes and also support learning will need to include multiple components: (1) assessments designed for use in the classroom as part of day-to-day instruction, (2) assessments designed for monitoring purposes that include both on-demand and classroom-embedded components, and (3) a set of indicators designed to monitor the quality of instruction to ensure that students have the opportunity to learn science as envisioned in the framework. The design of the system and its individual components will depend on multiple decisions, such as choice of content and practices to be assessed, locus of control over administration and scoring decisions, specification of local assessment requirements, and the level and types of auditing and monitoring. These components and choices can lead to the design of multiple types of assessment systems*. (pp. 212–213)

The large-scale summative assessment will be limited in the breadth of NGSS PEs that can be assessed. As previously described, this is due, in large part, to the nature of the PEs and the depth at which the assessed standards need to be measured. As a result, there is a limit to the number of score points that can be produced by any single assessment, and caution is warranted in interpreting test scores, both for individual students and for aggregate populations. Many of the strains on the summative assessment can be mitigated if the summative assessment is supported through use of the type of multimeasure, multilevel assessment system recommended in this report.

Key Takeaways from Chapter Three

- An evidence-based approach builds validity by matching claims/inferences with evidence of learning through five interrelated activities: domain analysis, domain modeling, conceptual assessment validity framework, implementation, and delivery.
- As recommended by the BOTA report, a system of assessments, from formative classroom assessment through summative standardized assessment, using a “bottom-up” approach, will best support the assessment of the NGSS.
- The highest-priority categories of evidence are construct validity, content validity, consequential validity, reliability, fairness, and feasibility.

CHAPTER FOUR: ITEM TYPES AND CLUSTERING

Introduction

This chapter seeks to address the concept of item clustering and to discuss its usefulness in ensuring the three-dimensional alignment called for by NGSS developers. Promising item types will be considered in this context, with guidance about the development of templates (also called item shells or prototypes) for each item type to help guide future item development. It is important to note that this chapter will explore item types and item clusters as individual components of an assessment; contexts for use of these components (e.g., static test, adaptive test) are discussed in **Chapters Six** and **Seven**.

Key Questions for this Chapter

- What types of items should be included on an NGSS-aligned assessment?
- What is the purpose of item clustering in the context of NGSS-based assessment?
- What item types can be used to elicit the types of evidence needed to make claims about student performance in addressing NGSS PEs?

Context of Item Clustering

SAIC members have come to an agreement on common terminology used to describe two components of this emerging assessment. First, an *item cluster* is a set of items (usually between four and six items, with some items having more than one part) that are based on at least one common stimulus (e.g., text, audio, video, animation, simulation, experiment). Individual items that are part of an item cluster are not intended to be separated and used independently from one another. Second, because the term *performance-based task* can be used to describe a broad family of assessment activities, the SAIC has adopted the definition outlined by Smarter Balanced: "[a] performance task involves significant interaction of students with stimulus materials and/or engagement in a problem solution, ultimately leading to an exhibition of the students' application of knowledge and skills" (Smarter Balanced, 2012, p. 1).

The BOTA report (NRC, 2014) recommends the use of assessment tasks with multiple components, rather than more traditional, discrete, stand-alone items:

> *CONCLUSION 2-1 Measuring the three-dimensional science learning called for in the framework and the Next Generation Science Standards requires assessment tasks that examine students' performance of scientific and engineering practices in the context of crosscutting concepts and disciplinary core ideas. To adequately cover the three dimensions, assessment tasks will generally need to contain multiple components (e.g., a set of interrelated questions). It may be useful to focus on individual practices, core ideas, or crosscutting concepts in the various components of an assessment task, but, together, the components need to support inferences about students' three-dimensional science learning as described in a given performance expectation.* (p. 44)

As presented in the Assessment Framework, item clusters are the large-scale summative assessment fulfillment of the assessment tasks recommended in the BOTA report. Item clustering will be needed in order to fully and accurately assess the NGSS. Additionally, each item within an item cluster must be aligned to at least two dimensions of the NGSS, with a strong preference that every effort be made, when feasible, to develop items aligned to all three dimensions of the NGSS. The overall item cluster must demonstrate alignment to all three dimensions.

One concern with an item-cluster-only approach for item development is that if the item cluster is appropriately developed, extracting individual items for stand-alone use will not be possible due to the scaffolded and intertwined nature of the items. A final consideration is that, at present, there are no known extant NGSS items developed that are fully aligned to the NGSS. For this reason, one planned outcome of the SAIC work is two prototype item clusters. These prototypes, which will be made available as separate documents, will offer examples of the item-cluster and NGSS alignment expectations.

The Assessment Framework presents an approach to item development that takes into consideration the following premises:

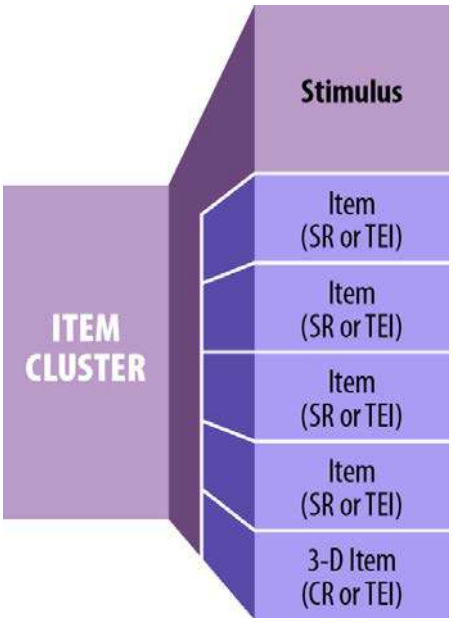
- Item clusters, not individual items, are the base unit for the SAIC test development. That is, individual items are intentionally developed to be situated within the context of an item cluster and not to be used as stand-alone items.
- Item clusters are the primary focus for developers in terms of alignment to the NGSS. That is, each item cluster must demonstrate strong three-dimensional alignment to the NGSS.
- To qualify as NGSS-aligned, item clusters must be aligned to one or more PEs and must be inclusive of all of the dimensions associated with the PE(s) (i.e., DCI, SEP, CCC).
- Each individual item within the cluster must align with at least two dimensions of the NGSS (e.g., DCI, SEP, and/or CCC) to qualify for inclusion in an item cluster.

Item clusters as described in this chapter and in the Item Specifications Guidelines fulfill these expectations.

Architecture of Item Clusters

As shown in Figure 1, the basic structure of an item cluster includes a common stimulus with an associated set of items.

Figure 1. Sample representation of the relationship of an item cluster to its component items



Each item is inextricably linked to the stimulus and to the other items within the item cluster, and the stimulus may be interspersed among the items to add information as needed. This means that student exposure to the stimulus is considered essential in order to respond correctly to any individual item, and that the item cluster must be constructed in such a way that individual performance on each item is adversely affected if an item is responded to without the context of the other items in the cluster. Testing time for each item cluster will be content dependent, but an approximation of 20 minutes of testing time per item cluster should be assumed. This time limit will allow for a reasonable overall test length while still providing an acceptable coverage of NGSS standards (i.e., PEs).

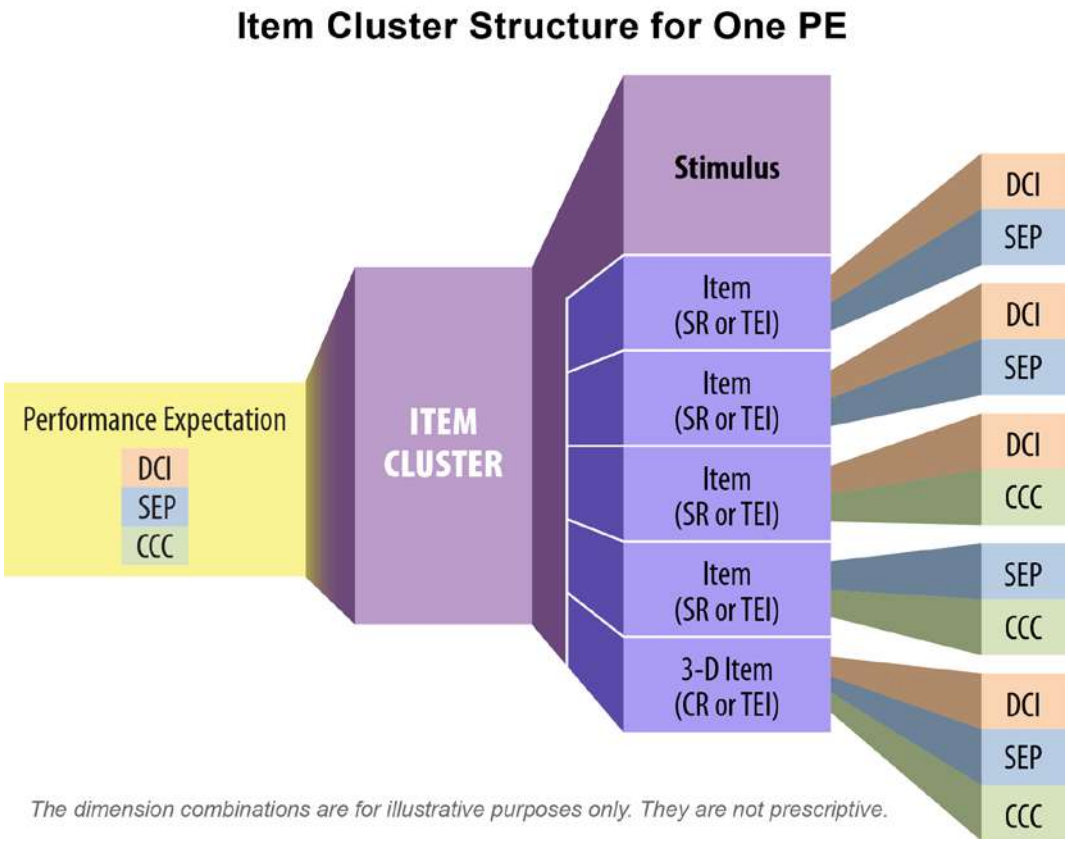
A *stimulus* is defined as a component of the cluster that does not directly require a student response. A stimulus can include one or more of the following:

- text;
- audio;
- video;
- animation/simulation;
- experimentation;
- discussion;
- activity; and/or
- demonstration.

Initially, stimuli will be identified or developed with the intent of inclusion on a large-scale summative assessment. States and developers should pursue creative solutions and should not allow current challenges of administration to constrain their thinking. The item cluster model is designed to allow for gradual evolution of stimuli, but still maintain NGSS alignment expectations.

An example of an item cluster's overall three-dimensional alignment is shown in Figure 2, with the dimensions of each item in a simplified single-PE cluster included.

Figure 2. Sample representation of the relationship of an item cluster aligned to a single PE to its component items, with item-aligned dimension combinations shown



- It should be noted that all items will exhibit some degree of alignment to the disciplinary context of the DCI, as all items are inextricably linked to the context, which was selected to align to the discipline(s) associated with the PEs. Therefore, every item in an item cluster will naturally fall within the content limits of the DCI, but not every item may truly call for the assessment of understanding of the content put forth in the DCI. Thus, items that only align to SEPs/CCCs are not intended to be viewed as devoid of a disciplinary context, but, rather, are intended to be viewed as items that place relatively greater emphasis on assessing an associated SEP and/or CCC than they do on assessing the underlying DCI content. Each SEP and CCC has its own knowledge that is most relevant in context of a DCI.
- If an evidence statement appears to align to a single SEP or CCC dimension, it is recommended that the evidence statement be grouped with the DCI, in order to prevent an item writer from developing an item to a single dimension in isolation (e.g., attempting to assess a science practice in isolation, without tying the item to the context and/or the DCI).

- At least one item should be aligned to all three dimensions, as shown in Figure 2, as this is the overall vision of the NGSS.
- Each item is inextricably linked to the stimulus and to the other items within the item cluster. This means that student exposure to the stimulus is considered essential in order for the student to respond correctly to any individual item, and that the cluster of items must be constructed in such a way that individual performance on each item is adversely affected if an item is responded to without the context of the other items in the cluster. (See the Item Specifications Guidelines for more information on stimuli for item clusters.)
- Testing time for each item cluster will be content dependent, but an estimate of 20 minutes of testing time per item cluster is assumed for summative assessment purposes. This estimate will be further refined as prototypes are completed.
- Each item cluster will have items tied to evidence statement selections for one or more PEs. These evidence statement selections are the fundamental component of item alignment with scientific content. Item clusters aligned to more than one PE could be from the same domain (i.e., Physical Sciences, Life Sciences, Earth and Space Sciences), but could also be from related, but different, content areas (e.g., photosynthesis and chemical reactions). PEs can also be from different domains. PEs from the domain of Engineering, Technology, and Applications of Science should always be bundled with PEs from one of the science disciplines.

  The rationale for correlating the parts of a PE evidence statement with two or more of the PE's dimensions is that such a correlation provides a building block for item construction when the PE is bundled with one or more other PEs in an item cluster. Looking at the entirety of the dimensions and evidence statements for two or more PEs in an item cluster can be somewhat overwhelming in terms of the amount of information provided in relation to assessment goals. By structuring the PE and evidence statement components into natural dimensional/evidence-statement relationships that might form the basis of an item in an item cluster, the item cluster developer can better perceive how all of these PE elements fit together and how they might be used, along with the multidimensional alignment groupings for other PEs in an item cluster, to form a balanced, conceptually cohesive item cluster.
- While it may be possible to develop items within a single cluster that are collectively sufficient to assess the entirety of a single PE, this is not preferable and will not be possible in many, if not most, cases. For item clusters inclusive of more than one PE, it is not expected that a single item cluster will be able to provide the opportunity for a student to generate evidence of every aspect of each PE in the item PE bundle.

More detailed explanations of item clusters are provided in the Item Specifications Guidelines.

Eligible Item Types

This section discusses the types of items that can be used to populate each item cluster developed for online administration. More specific examples of item types are presented in the Item Specifications Guidelines. The typical item cluster will consist of four to six single-part or multipart items. As previously stated, for an item cluster to be judged as fully aligned to the NGSS, it is recommended that the item cluster contain constructed-response formats, within a single item and/or interspersed throughout the cluster (e.g., create a model and provide an explanation). Table 2 summarizes the known item types that are frequently used on large-scale state and multistate summative assessments.

Table 2. Frequently Used Item Types

Item Type	*Item Subtype and Structure*	*Response Behavior*	*Sample Task/Purpose*
Selected response (SR)	Multiple choice, single correct response (MC)	Select an option by clicking on a radio button or anywhere in the text; generally four options	Identify an appropriate rationale to explain a scientific phenomenon; select an appropriate solution to an engineering design problem
	Multiple choice, multiple correct responses (multiple select) (MS)	Select among multiple options by marking a checkbox or clicking anywhere in the text; generally five or more options	Identify a plausible explanation for a phenomenon and the appropriate rationale; select statements that support a claim of a scientific phenomenon
	Matching tables (with True/False or Yes/No) (MT)	Select among multiple statements by marking an option in a table cell for each row	Identify appropriate data; identify appropriate statements given constraints
	Inline choice (IC)	Select an option by clicking on a drop-down menu; four options	Identify evidence that would support a claim or explanation
	Hot spot (HS)	Select text or objects in a response area; may include more than four options; each option should be a salient feature	Identify aspects of a model that support a given claim

Table 2. *(continued)*

Item Type	*Item Subtype and Structure*	*Response Behavior*	*Sample Task/Purpose*
Constructed response (CR)	Short text (ST)	Enter text into a multiline text box.	Generate a hypothesis; describe a possible engineering problem
	Equation or numeric entry; edit equations (EQ)	Enter mathematical symbols and/or numbers; may include selecting special symbols from an on-screen table or menu	Use a mathematical model to represent a scientific phenomenon; determine a solution to an engineering problem
	Cloze text (CT)	Enter text into a text box within a sentence	Construct a description or simple explanation of a scientific phenomenon or solution to an engineering problem
	Table text (TT)	Enter text into a table or chart	Design an investigation or make predictions
	Constructed response (essay) (CR)	Use keyboard to enter text into a multiline text box; may include text formatting tools	Construct a detailed explanation of a scientific phenomenon or solution to an engineering problem
Technology-enhanced items (TEIs): Data selection	Slider (SL)	Select a value on a scale by clicking on a slider and dragging it to the appropriate location on the scale	Select values for variables to design and conduct an investigation
	Data inspector (DI)	Select a value on a graph by clicking on a slider attached to a vertical line and dragging it to the appropriate location on the *x*-axis	Select data as evidence to support an explanation or the solution to an engineering problem

Table 2. *(continued)*

Item Type	*Item Subtype and Structure*	*Response Behavior*	*Sample Task/Purpose*
Technology-enhanced items (TEIs): Data display	Graphing: plot points or line graphs (G)	Click in the question response area to create a point or start a line; click and drag to complete the line and to add additional data points	Create a model; analyze data
	Function graph (FG)	Click on an icon to select the type of graph; drag two points to the correct position	Create a model; analyze data
	Composite graph (CMG)	Click in the question response area to create composite displays including two or more of the following: points, lines, curves, or shaded areas	Analyze data by fitting a line or curve to a set of points; represent possible solutions to an engineering problem, using shaded areas
	Bar graph; histogram (BG)	Drag bars to display data in a bar graph or histogram	Create a model; analyze data
	Fraction model (circle graph) (FM)	Click on the edge of a circle to create a new division, and/or drag division lines to the appropriate location within a circle	Create a model; analyze data
	Interactive number line (INL)	Click in the question response area to create a point or start a line; click and drag to complete the line	Create a model; analyze data
	Zoom number line (ZNL)	Present graphical data by zooming in on a number line to graph one point (often used for fractions)	Analyze data; demonstrate how an outcome may be affected by a unique context

Table 2. *(continued)*

Item Type	*Item Subtype and Structure*	*Response Behavior*	*Sample Task/Purpose*
Technology-enhanced items (TEIs): Drag and drop	Drag and drop single or multiple elements (DD)	Select an object by clicking on it; then drag and drop it into an appropriate location within the response area (including tables and art)	Modify a model to better fit a new constraint
	Hot text: select and order text (HT)	Select text by clicking on it; then drag and drop it into an appropriate location within the response area (including tables and art)	Reorder steps/stages into an appropriate sequence, given a context or scenario
	Text extraction (EXT)	Select text from a sentence or equation by clicking on it; then drag and drop it into an appropriate location within the response area (including tables and art)	Select parts of a description of a scientific phenomenon that support an explanation or solution to a problem
Multi-component	Two-part multiple choice, with evidence-based response (EB MC)	Part 1: Select an option by clicking on a radio button or anywhere in the text Part 2: Select a response to support the response to Part 1	Identify a response/claim and the appropriate rationale to support the response/claim
	Other	Any combination of two functionalities within a single item	Generate and test models; display, analyze, and interpret data; design and conduct investigations or solutions and explain results

Source: Based on Smarter Balanced (2014) and PARCC (2013).

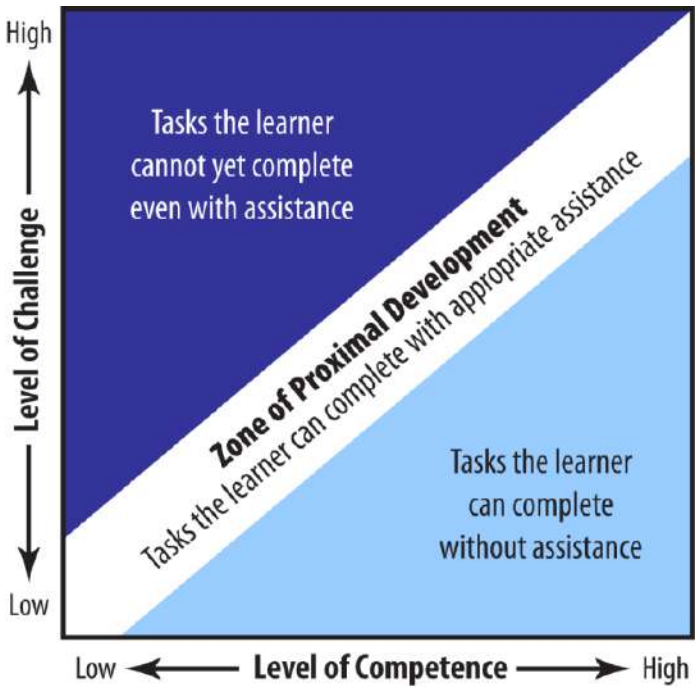
Use of essay or constructed-response item formats can present specific challenges when these formats are included in large-scale assessments. As previously noted, hand-scoring of constructed-response items can be costly and time consuming. Automated scoring protocols for constructed-response items are not well established and are often not as reliable when compared to human scoring results under most circumstances. Currently, constructed-response formats are most likely machine-scored using key words, and this scoring does not often take into consideration the contextual use of the key words, thus introducing scoring errors when

compared with human scorers. As new machine-scoring algorithms are developed, however, new tools that are more trustworthy and efficient may become available. Because of the emerging nature of automated scoring of CR items, the assumption for the item cluster model in the Assessment Framework is that CR items are necessary to fully access the NGSS, but that their use should be limited due to constraints of summative testing and related costs (either in hand scoring or in development of automated scoring engines).

Cognitive Demand and Scaffolding

By gradually increasing the level of cognitive demand, scaffolding potentially can be very useful during assessment. That is, scaffolding within an item cluster can be used to guide students through a series of progressively more challenging interrelated questions to elicit evidence of what students know and are able to do in relation to a specific PE or PE bundle and the associated dimensions. This gives students the opportunity to first experience stimuli-associated items that are relatively less cognitively demanding before being exposed to items that ask them to demonstrate understanding of deeper, more complex concepts. Scaffolding can be added strategically within an item cluster to lead students through a series of items that build upon one another, each subsequent item requiring a deeper understanding of the stimulus than did the prior item and encouraging students to consider responses to previous items when tackling the next item. This approach is based on a research-supported theory of learning that highlights the importance of understanding what each student currently knows and can do with and without appropriate assistance, as illustrated in Figure 3 (Vygotsky, 1978). Scaffolding allows for students to answer items occurring early in a sequence incorrectly, but not have inaccurate information or responses carry through to later items in the cluster.

Figure 3. The relationship between cognitive demand (challenge) and student competence



A wide variety of item types readily lend themselves to the application of scaffolding strategies that gradually increase the level of cognitive demand required to respond to successive items. In addition, item clusters provide an ideal environment for scaffolding, as items within a cluster are linked to a common purpose (i.e., all items are measuring certain aspects of a PE or multiple PEs) and can be delivered strategically so as to nudge students to attempt and correctly respond to progressively more challenging items.

State education agencies and other developers of NGSS-based measures will want to ensure that the types of items chosen to populate each item cluster represent a range of cognitive demands and the scaffolding strategies applied are linked to the common purpose. One option to help ensure a range of cognitive demand that is appropriate to the PE or PEs to which an item cluster is aligned is to utilize the evidence statements released by Achieve, Inc. (NGSS Network, 2015a, 2015b). Per the executive summary of the high school evidence statements' front matter (NGSS Network, 2015a),

> *The evidence statements were designed to articulate how students can use the practices to demonstrate their understanding of the DCIs through the lens of the CCCs, and thus, demonstrate proficiency on each PE.* (p. 1)

The evidence statements are arranged into numerical categories (1, 2, 3, and sometimes 4) and represent a range in terms of cognitive demand. It is important to note that the categories do not equate with any established cognitive scale (e.g., Webb's Depth of Knowledge or Bloom's Revised Taxonomy), and that the numbered categories do not correlate with or imply degrees of cognitive complexity (e.g., category label 1 does not necessarily imply lower cognitive complexity than category label 2). Developing item clusters that include items aligned to the full

range of evidence statements will result in items covering the range of cognitive complexity intended for the performance expectations..

Extant Science Items to Build Upon

The BOTA report (NRC, 2014) supports the development of item clusters:

> *To adequately cover the three dimensions, assessment tasks will need to contain multiple components, such as a set of interrelated questions. It may be useful to focus on individual practices, core ideas, or crosscutting concepts in a specific component of an assessment task, but, together, the components need to support inferences about students' three-dimensional science learning as described in a given performance expectation (Conclusion 2-1).* (p. 63)

Model items are needed to support developers of NGSS-based assessments. Several recent efforts have leveraged technology in creative ways to push thinking in new ways in terms of model items. Based on these efforts, this section discusses an exemplary set of items (with examples included in **Appendix F**) that have clear strengths and limitations, which might be useful to developers of NGSS-aligned item clusters.

These examples incorporate technology-based enhancements that make possible the kinds of on-demand performances that simulate real-world investigations and authentic contexts necessary to accurately assess the SEPs in the NGSS. These simulations are becoming more common features of classroom instruction in science and, as such, represent avenues for assessment that is synchronized with science teaching and learning and that is more authentic than traditional items. Recent research suggests that simulations can have the technical quality and feasibility to be credible components of state assessment systems (Quellmalz, Silberglitt, & Timms, 2011). However, students need opportunities to experience science simulations in the classroom and to become more familiar with technology-supported tools and interfaces before these model items are incorporated in large-scale measures of science learning (NRC, 2014).

The BOTA report (NRC, 2014) describes simulation-based modules for ecosystems, developed by the SimScientists program (http://simscientists.org), that interweave information about ecosystems with the NGSS SEPs of Building and Using Models, Planning and Conducting Investigations, and Interpreting Patterns, and with the NGSS CCC of Systems and System Modeling. As shown in Figures F-1 through F-3 in **Appendix F**, these modules leverage technology to present a model of the characteristics of and changes in an ecosystem, features that would be difficult to include with printed materials.

New task formats developed for the National Assessment of Educational Progress (NAEP) also show promise for measuring SEPs in authentic contexts, although they are not specifically aligned to the NGSS. With support from WestEd, several interactive computer tasks (ICTs) were developed and tested on the 2009 NAEP Science Assessment. For the 2014 NAEP Technology and Engineering Literacy Assessment, novel interactive tasks engaged students in engineering practices related to problem solving. Although these tasks are not aligned to the SEPs in the

NGSS, they demonstrate how technology-based enhancements can provide dynamic, interactive stimuli that invite students to engage in engineering practices during instruction and assessment. Examples from NAEP are shown in Figures F-4 through F-6 in **Appendix F**.

As stated earlier, one outcome of the SAIC work is two prototype item clusters. These prototypes, which will be made available as separate documents, offer examples of the item-cluster and NGSS alignment expectations.

Evolution of Eligible Item Types

Novel, interactive tasks that push the limits of current technology in schools bring unique challenges for schools, in terms of infrastructure and capacity, and for psychometricians, who must monitor technical quality, fairness, and consequences (intended and unintended, positive and negative) of testing with new item types. Further, even as systems for state science assessments evolve to include promising innovative approaches, emerging technologies continue to extend the boundaries of what is possible in education, from games that build upon the interactivity in science simulations to create immersive environments with embedded "stealth" assessments (Shute & Ventura, 2013) to augmented reality, three-dimensional immersive worlds, and three-dimensional printing, which show promise for evaluating learning across virtual and physical media (Metcalf, Kamarainen, Tutwiler, Grotzer, & Dede, 2011; Davenport, Silberglitt, & Olson, 2013).

Assessment developers also recognize that science teaching practices are changing at all levels of the educational system, from classrooms to states. Implementation plans that acknowledge these transitions allow for gradual incorporation of innovative item types on assessments, in tandem with instructional changes. To ensure that progress is made, however, states need to set benchmarks for when they anticipate that particularly promising item types might be introduced. The BOTA report recognized the need for a gradual implementation from traditional approaches to a "fully integrated, technologically enhanced, coherent system of assessments" (NRC, 2014, p. 230).

The item cluster model is designed to allow for gradual evolution of items used on NGSS-based assessments, while maintaining NGSS alignment expectations.

Key Takeaways from Chapter Four

- Item clusters consisting of a stimulus and a variety of item types are the proposed key components of a large-scale summative assessment that measures all three dimensions of the NGSS.
- The different item types from the Race to the Top assessment consortia (PARCC and Smarter Balanced) will be included in the scope of the eligible item types. These item types are described in this chapter, along with a brief explanation of what types of evidence each provides in a science assessment.
- Scaffolding within and across assessment item clusters helps guide students through a series of progressively more challenging interrelated questions, to better provide evidence of the knowledge and skills of students across a wide range of ability and understanding.
- Online assessments provide opportunities for more interactive stimuli and a greater range of item types that provide better evidence of science and engineering practices and crosscutting concepts than printed materials do.
- Novel, interactive tasks bring unique challenges and opportunities for schools, in terms of infrastructure and capacity, and for psychometricians, who must monitor technical quality, fairness, and consequences (intended and unintended, positive and negative) of testing with new item types.

CHAPTER FIVE: ITEM SPECIFICATIONS GUIDELINES

Introduction

Key Questions for this Chapter

- How and by whom were the item specifications developed?
- What guidance is provided to support SAIC states with the development of item specifications?

Item specifications guidelines were developed separately from the Assessment Framework and are available as the *Science Assessment Item Collaborative Item Specifications Guidelines for the Next Generation Science Standards*. These guidelines provide states with the necessary documentation to enlist the support of test vendors in designing and developing NGSS-aligned item clusters. The guidelines serve to aid in the development of state-specific item specifications, and item cluster specifications, that yield item clusters that meet the specified alignment and structure criteria as described in this Assessment Framework, and they describe the characteristics that are needed to effectively measure different PEs or clusters of PEs.

The Item Specifications Guidelines focus specifically on how the content of the NGSS will be assessed, by articulating the NGSS–to–item cluster correlations that are necessary for the development of NGSS-aligned items, item clusters, and assessments. In particular, emphasis is placed on developing item pools for summative assessment (i.e., large-scale, evaluative testing done at the end of academic years) of the NGSS.

The following sections are included in the Item Specifications Guidelines:

- Introduction
    - Basic Terminology
- General Item Specifications Guidelines
    - Cognitive Complexity
    - Universal Design/Vocabulary and Language
    - Scoring Considerations
    - Achievement Level Descriptors and Special Student Populations
- Item Cluster Alignments
    - PE Item Specifications
    - Multi-PE Item Cluster Alignments
        - PE Bundling Considerations
        - Example of a Two-PE Item Cluster Alignment
        - From Item Cluster Alignment to Item Cluster

- Guidelines for Item Clusters
    - Item Cluster Stimuli
- Item Types and Subtypes for Item Clusters
    - Selected-Response Items
    - Constructed-Response Items
    - Technology-Enhanced Items
    - Hybrid Approach to TEIs

Key Takeaways from Chapter Five

- The Item Specifications Guidelines provide states with the necessary documentation to enlist the support of test vendors in designing and developing items to be used in the development of NGSS-based assessments.
- The Item Specifications Guidelines provide suggested procedures and strategies to aid in the development of state-specific item specifications, and item cluster specifications that yield items that meet the criteria specified in the Assessment Framework.
- The Item Specifications Guidelines focus specifically on how the content of the NGSS will be assessed, by articulating the NGSS–to–item cluster correlations that are necessary for the development of NGSS-aligned items, item clusters, and assessments.

CHAPTER SIX: DEVELOPING BLUEPRINTS

Introduction

Key Questions for this Chapter

- What are the guiding principles for blueprint development for a next-generation science assessment?
- How should the Assessment Framework inform each state's development of achievement level descriptors?

Effective blueprints guide developers in building a test that meets the criteria for technical quality and fully assesses a body of standards. Blueprints developed for NGSS-based assessments (next-generation science assessments, or NGSAs) should describe the overall design for the tests and provide specific directions on how item clusters and their expected metadata characteristics (e.g., alignment, depth of knowledge, difficulty, content) can be used to evaluate what students know and can do in relation to the new science standards. State-specific blueprints should specify the numbers of items for each item type, the balance of representation (e.g., percentages of items for each PE or dimension), and how to achieve the balance of cognitive demand that is required for assessment.

This chapter seeks to provide general guidance for SAIC test blueprint developers at the state level. It acknowledges that states will own this phase of work so that the resulting outcomes can be customized for their unique contexts.

NGSS Considerations When Developing Test Blueprints

The accurate measurement of a student's three-dimensional learning of science is a critical consideration for NGSS-based assessments. No longer can a one-to-one association be made between an item and a standard. Instead, the complex nature of the NGSS requires a more progressive blueprint and test design in order to meet the challenge of measuring three-dimensional science learning, as emphasized in Conclusions 2-1 and 2-4 of the BOTA report (NRC, 2014):

> *CONCLUSION 2-1 Measuring the three-dimensional science learning called for in the framework and the Next Generation Science Standards requires assessment tasks that examine students' performance of scientific and engineering practices in the context of crosscutting concepts and disciplinary core ideas. To adequately cover the three dimensions, assessment tasks will generally need to contain multiple components (e.g., a set of interrelated questions). It may be useful to focus on individual practices, core ideas, or crosscutting concepts in the various components of an assessment task, but, together, the components need to support inferences about students' three-dimensional science learning as described in a given performance expectation.* (p. 44)
>
> *CONCLUSION 2-4 Effective evaluation of three-dimensional science learning requires more than a one-to-one mapping between the Next Generation Science Standards (NGSS) performance expectations and assessment tasks. More than one assessment*

> *task may be needed to adequately assess students' mastery of some performance expectations, and any given assessment task may assess aspects of more than one performance expectation. In addition, to assess both understanding of core knowledge and facility with a practice, assessments may need to probe students' use of a given practice in more than one disciplinary context. Assessment tasks that attempt to test practices in strict isolation from one another may not be meaningful as assessments of the three-dimensional science learning called for by the NGSS.* (p. 46)

The item cluster is presented as the base unit of an NGSS-based assessment. As such, test specifications and blueprints should first specify the characteristics of the item clusters to be included in the test (e.g., number, DCI domains represented) and how they may vary for common test sections and matrix test sections. The stated expectations (i.e., characteristics) of item clusters (e.g., time period, PE bundling, individual items being two- or three-dimensional, the range of evidence statement categories) are intended to support a test blueprint that achieves the expected coverage of the NGSS. Additionally, states should determine how best to specify expectations of the item cluster metadata and how those expectations should be addressed in test blueprints.

Guiding Questions

Given the widely held belief that assessments should measure what is taught, careful consideration of the assessment system and of the content to be assessed is critical. Further, the structure of the NGSS requires that instruction and, consequently, assessment reflect "the inherent complexity in scientific understanding and reasoning as it exists in the real world" (Gorin & Mislevy, 2013, p. 2). The following questions are key to the goal of developing and designing an instructionally sensitive assessment or assessment system, one characterized by test blueprints that reflect real-world application of scientific understanding and reasoning, the three-dimensional structure of the NGSS, and the need for student-level and state-level reporting:

- Will there be a single summative assessment, or will the state implement a "bottom up" assessment system as outlined in the BOTA report?
- What amount of time will be devoted to testing (summative and interim)?
- If the decision is made to administer a summative-only assessment, will it be administered at each grade or at each grade band?
- Who are the target audiences for this assessment system? Will diverse stakeholder groups be served?
- What are the goals for reporting?
- What inferences from test results—e.g., claims about student performance—do developers seek to draw?
- What reporting categories are envisioned? Should the DCIs, CCCs, and SEPs be reported separately?

- Other than student-level reports, what other levels of reporting are of interest? Will the assessment be adaptive, a static common item design in which all students answer the same set of questions, or a combination of both common and matrix-sampled items?
- Is the state committed to supporting an online assessment delivery, or will content be delivered in paper-and-pencil format?

Careful thought and consideration need to be given to each of these questions, and the decision-making process should include the breadth of stakeholders who have an interest in the structure and reporting of the assessments. There is no single correct approach to developing a test blueprint or assessment system, but the test blueprints should be consistent with the standards, should reflect fidelity to instruction, and should provide meaningful data to evaluate how well the standards are being implemented.

Achievement Level Descriptors

This Assessment Framework should also help inform each state's development of a set of state-specific initial achievement level descriptors (ALDs) that will be aligned with the NGSS for each achievement level and for all tested grades. It is recommended that the development of ALDs occur concurrently with the test development cycle, following the approach implemented by Smarter Balanced (2012). This shift will allow the ALDs to specifically address student performance expectations that should ultimately inform the way a state's NGSS-aligned science assessments are conceived and developed.

Effective ALDs clarify and make transparent the knowledge, skills, and processes that students are being asked to demonstrate at predetermined levels of achievement (for example, Basic, Proficient, and Advanced, or Levels 1–4). ALDs are often included in student-level score reports as well as on state aggregate reports, and in order to be effective, ALDs must be able to clearly explain the differences among the discrete proficiency levels (that is, what students know and should be able to do at each level) to all stakeholders, from parents to teachers to state policymakers. The limited number of score points that will be achievable with an assessment structured on item clusters should be considered in the determination of the number of ALD levels. The limit is a result of limited testing time and the expectation of greater depth of student evidence, which is essential in an NGSS-based assessment.

The NGSS evidence statements (NGSS Network, 2015a, 2015b) include a single proficiency level, and do not include information on determining levels of achievement beyond a single "proficient" level relevant to the PEs. Developing ALDs for multiple levels of achievement at the beginning of the test development cycle will help to ensure that the assessment supports the distinctions among the levels and that it will ultimately translate into transparent and valuable descriptors for all stakeholders. As shown in Table 3, policy, range, threshold, and reporting ALDs should be considered.

Table 4. ALDs by Use, Purpose, and Intended Audience

ALD Type	*Use*	*Purpose*	*Intended Audience*
Policy	Test development and conceptualization	Set tone for the rigor of performance standards expected by sponsoring agency	Policymakers
Range	Item-writing guidance	Define content range and limits	Item writers and test developers
Threshold	Cut-score recommendation and standard-setting guidance	Define threshold performance at each achievement level	Standard-setting panelists
Reporting	Test score interpretation	Describe the knowledge, skills, and processes that test-takers demonstrate and indicate the knowledge and skills that must be developed to attain the next level of achievement	Stakeholders, such as parents, students, teachers, K–12 leaders, and higher education officials

Source: Smarter Balanced (2012).

Key Takeaways from Chapter Six

- State-developed test blueprints should specify the numbers and types of test components needed (e.g., item clusters, item types), the balance of representation (e.g., percentages of items for each PE or dimension), and the levels of cognitive demand to be assessed.
- The test blueprints should be consistent with the standards, and should provide meaningful data to evaluate how well the standards are being implemented.
- Developing ALDs for multiple levels of achievement at the beginning of the test development cycle will verify alignment among the assessment targets and the descriptors and will ensure that the assessment content supports the distinctions among the levels and that it will ultimately translate into transparent and valuable descriptors for all stakeholders.

CHAPTER SEVEN: ITEM CLUSTER DEVELOPMENT

Introduction

Key Questions for this Chapter

- How should PEs be unpacked and bundled to support the proposed item cluster architecture?
- How should PEs be bundled together in order to support rich and robust item clusters?
- What will the typical development process for designing, authoring, and testing item clusters entail?

The NGSS emphasize that students must have the opportunity to learn science content in ways that more closely resemble the ways in which real scientists think and work (NRC, 2014). Per the front matter of the NGSS (NGSS Lead States, 2013):

> *The real innovation in the NGSS is the requirement that students are required to operate at the intersection of practice, content, and connection. Performance expectations are the right way to integrate the three dimensions. It provides specificity for educators, but it also sets the tone for how science instruction should look in classrooms. If implemented properly, the NGSS will result in coherent, rigorous instruction that will result in students being able to acquire and apply scientific knowledge to unique situations as well as have the ability to think and reason scientifically.* (pp. 3–4)

Accordingly, NGSAs should mirror how science content is taught and tested in the classroom, and should reflect the "bottom-up" approach to designing an assessment system, as recommended in the BOTA report (NRC, 2014). Formative assessment in the classroom checks for understanding throughout the process of instruction and may utilize summative tasks in order to elicit evidence about student understanding. For the purposes of the Assessment Framework, discussion of item development will focus on NGSS-aligned large-scale summative assessments.

The first step in item cluster development is to bundle PEs. PEs should be bundled in a way that mirrors how they are presented (in terms of grouping) and taught in science classrooms at the state and/or local level. Sample groupings of PEs and the rationale behind each grouping are provided in the Item Specifications Guidelines. When designed effectively, the grouping of PEs will facilitate (1) the assessment of CCCs that require the purposeful intersection of two or more PEs during assessment (e.g., PEs that share a common CCC); (2) the application of a real-world stimulus that addresses an overarching science phenomenon; and (3) increased breadth achieved by the assessment. Through effective PE groupings and proper item development processes, the item clusters will be structured to gather the evidence defined in the previously detailed steps, in support of an evidence-based approach.

To identify the assessable aspects of the PEs and their associated dimensions for assessment purposes (see **Chapter Two**), the second step is to organize the relevant information for each PE and its associated dimensions by synchronizing the following: (1) information from the assessment boundaries identified in the NGSS; (2) the evidence statements; (3) Appendices E, F, and G (progressions for DCIs, SEPs, and CCCs) of the NGSS; and (4) the *K–12 Framework*.

Evidence Statements

Evidence statement development, led by Achieve Inc., has been completed for all grade levels and grade bands. These evidence statements are intended to be used by educators and assessors to ascertain what evidence can be gathered from an assessment in order to determine a student's level of mastery for a given PE. As noted in the evidence statements' front matter (NGSS Network, 2015a):

> *The evidence statements can serve as supporting material for the design of curriculum and assessments. In the NGSS, each PE is accompanied by a foundation box with associated practice, core idea, and crosscutting concept. The evidence statements expand this initial structure to include specific, observable components of student performance that would demonstrate integrated proficiency by using all of the necessary tenets of the practice to demonstrate understanding of the disciplinary core ideas (DCIs) through the lens of crosscutting concepts (CCC). We hope that by providing these links among the practice, DCI, and CCC for each PE, educators and assessors will have a clearer idea about 1) how these dimensions could be assessed together, rather than in independent units or sections; 2) the underlying knowledge required for each DCI; 3) the detailed approaches to science and engineering practices; and 4) how crosscutting concepts might be used to deepen content- and practice-driven learning.* (p. 3)

It is important to note that the evidence statements have been written to allow for multiple methods and contexts of assessment and for the assessment of multiple related PEs.

Progressions

The *K–12 Framework* specifies grade-band endpoint targets at grades 2, 5, 8, and 12 for each component of each core idea. Appendices E, F, and G of the NGSS provide summaries of how each dimension increases in complexity and sophistication across the grades, as envisioned in the *K–12 Framework.* The learning progressions in these documents describe the upper and lower anchors of these progressions, beginning with the end of second grade and ending with what students are expected to know about science when they leave high school. The BOTA report (NRC, 2014) states the following:

> *Assessment developers will need to draw on the idea of developing understanding as they structure tasks for different levels and purposes and build this idea into the scoring rubrics for the task. The target knowledge at a given grade level may well involve an incomplete or intermediate understanding of the topic or practice.* (p. 37)

The progressions inform item cluster developers on how to interpret upper and lower limits for expectations of alignment expectations and appropriate content for items and item clusters.

Bundling of Performance Expectations

Once PEs are unpacked, they are bundled together in order to support rich and robust item clusters. Developers of NGSS-based assessments can use a number of strategies to decide which PEs should be bundled together for assessment. For example, the PEs may be bundled across domains (e.g., Life Science and Earth and Space Science PEs) or grades, or may be domain- or grade-specific. PEs may be bundled with dimensional overlap (PEs with the same dimension in each) or bundled to include a range of dimensions (i.e., to limit dimensional overlap). Importantly, however, these decisions should be intentional, and should demonstrate understanding of each option and careful weighing of the implications of each approach for the resulting item clusters and test designs. For instance, two PEs that are bundled within a grade and have dimensional overlap may produce item clusters that can obtain a greater depth of the SEP content and grade-level content, but may also result in an item cluster that provides less overall breadth at that grade level or grade band (that is, assessing only one SEP in an item cluster instead of two or even three SEPs).

Regardless of the approach or approaches used, PEs should be bundled in a way that enables assessment via a single natural phenomenon that is presented within a stimulus. To do so, it is recommended that a maximum of three PEs are bundled together. Some PEs may be robust enough to support an entire item cluster on their own, but it is likely that most bundles will consist of two or three PEs. An individual PE may be used in more than one bundle, depending on the approach or approaches used, so care must be taken to ensure that an individual PE is not unintentionally overrepresented among the bundles. It is anticipated that many different item clusters can and will be created from any one bundle in order to assess the full breadth of the bundled PEs and their associated dimensions through a myriad of analogous natural phenomena.

Development of Item Clusters

Different approaches may be used to develop item clusters. One approach is to hire an outside vendor with relevant assessment experience; another is to engage state educators in completing the work. Regardless of the approach selected, the interrelated and complex nature of assessing the three dimensions requires that certain elements are present in an NGSA.

As noted in **Chapter Four**, the recommended item cluster architecture could include a range of four to six single and/or multipart items representing different item types, including machine-scored TEIs and selected-response (SR) items (e.g., multiple choice, multiple select, fill-in-the-blank), to satisfy the full breadth and depth of the PEs (especially for hard-to-assess SEPs) and to allow for an appropriate range of cognitive demands and scaffolding of items. An item cluster must measure all three dimensions of the NGSS, but individual items need only align to two of the dimensions. Due to the interconnected nature of the dimensions, an item that aligns to only one dimension is not appropriate for inclusion in an NGSA. (Aspects of at least two dimensions must be present in all items because knowledge or skills cannot be effectively assessed in isolation.)

Use of evidence statements as a basis for item alignment will ensure that important content-validity expectations are considered. NGSA developers will want to be careful to align an item to all parts of a complete evidence statement, in order for the item to be considered fully aligned to the NGSS. Finally, all evidence statement categories should be addressed within an item cluster, in order for the cluster to be considered appropriately aligned to the intent of the NGSS. Alignment should be evaluated both at the item level and at the item-cluster level.

Criteria for Stimuli Development

Each item cluster has a common stimulus upon which all items in the cluster are dependent. Items that can be answered without referring to a stimulus are not appropriate for an NGSA. To address the full depth and breadth of one or more PEs, the stimulus may be a text-based description that includes a description of experimental data or an experimental setup. Alternatively, the stimulus may be a fully interactive computer simulation in which students can control for different variables, run multiple trials, and collect and analyze data.

The stimulus is an integral part of the item cluster and, therefore, should meet specific criteria for inclusion in an NGSA. To mirror the work of real scientists, the stimulus should be based on a real-world science phenomenon that is representative of how students learn in the classroom. The ideal stimulus also should have a rich, grade-appropriate context that can support a variety of robust item types used to gather evidence of what students know and are able to do. These contexts should be:

- creative;
- based on real and verifiable sourced data;
- as specific and concrete as possible;
- appealing to students, but not commonly used in school textbooks or classroom experiments; and
- free of bias/sensitivity issues.

A stimulus may be composed of several different components—e.g., animations, text, tables, graphs, and images—and may include an interactive experience that students must undertake before items can be answered. The stimulus may be accessible after the student has answered a few questions, in order to provide information that is necessary for the student to answer subsequent questions in the cluster.

To ensure a high standard for selection, as well as efficiency in the selection process, a completed stimulus, along with item ideas, should be submitted to a state education agency (SEA) for approval before development work begins in earnest. It is important to note that item ideas are not complete items with all parts present, but are, instead, the rough basis upon which items will be developed. Item ideas should include a description of each item, designation of item type(s), and documentation of alignment, so that reviewers have a clear understanding of the entire item cluster being proposed.

In traditional item development, the number of items developed exceeds the minimum number of items needed, so as to account for attrition that occurs during each round of review. Due to the very integrated nature of an item cluster, an overage of items cannot be developed in this manner. One possible solution is to develop two different complete item clusters using a single stimulus, with the understanding that the items likely will not be interchangeable across clusters.

Once stimuli are approved, item ideas should be developed into fully functioning items by well-qualified item writers and/or editors familiar with high-stakes item development processes, industry best practices, and the cognitive demands inherent in each of the three dimensions of the NGSS. Individual items should be authored to work together as an integrated set and should provide the appropriate degree of scaffolding needed to determine levels of mastery. For online administration, students should only be allowed to work in the forward direction as they answer items, but they should be allowed to navigate in both directions between the parts within an item. This allows items later in the set to cue or clang with previous items, but not to be cued by or clang with a following item or part within the same item.[1] A variety of item types with a range of cognitive complexities that align to different dimensional combinations (e.g., SEP and CCC, or CCC and DCI) should be used to elicit the evidence necessary to determine student proficiency.

Item Cluster Reviews: Considerations

Fully developed item clusters should be reviewed by review committees with diverse membership that may include SEA staff, state educators, trained assessment specialists (e.g., district administrators or test coordinators), content specialists, and curriculum developers. Review panels will want to consider cluster length, readability, format/style, typography, content, vocabulary, sentence complexity, concept load or density, and cohesiveness. Typical review topics include:

- Within each item, is it clear what the students are being asked to do?
- Do the individual items match the standard(s) being assessed?
- Does the set of items in the item cluster measure an appropriate breadth and depth of the PE(s)?
- Do the items adhere to any available item specifications?
- Does each item have only the intended number of correct answers?
- Is the content and vocabulary in the cluster of items grade-level appropriate?
- Does each item in the cluster assess unique content?
- Is there any cueing or clanging as students move from start to finish throughout the item cluster? (For any item in a set, could a student get the right answer for the wrong reason

[1] *Cueing* occurs when a student is steered toward a correct or incorrect answer option based on information that appears in the stimulus or stem and in only that answer option. *Clanging* is when a student is able to correctly answer an item based on information provided in the stem or answer options of another item in the set.

or the wrong answer, based on a valid strategy? Do the stems or answer options for any items that appear early in the set provide information about the correct answers to any items appearing later in the set?)

- Do all items in the cluster make use of the stimulus?
- Is an appropriate degree of scaffolding present within and among items?
- Is each item in a cluster fully aligned to the specified evidence statement(s) within the context of the PE?

Recommendations for improvements to items, stimuli, or any other elements (e.g., difficulty, item sequence) should be discussed so that group consensus can be reached, and final decisions should be documented. Dissenting comments, concerns, and any other remaining issues should be noted on the items/stimuli and used to drive subsequent discussions among all parties.

Items or stimuli may require significant post-review editing. All item clusters, especially ones for which consensus was not reached or for which significant edits were required, should be reconciled with SEA staff, to ensure that the recommended changes meet SEA approval before they are implemented. Feedback from the item-review process should be captured and used to inform future item development.

Piloting and Feedback Loop

Item clusters will be piloted at state and local levels, either as embedded field-test item sets on operational forms or on separate pilot forms. Student-level data will be collected, and, once these data are collected, item clusters should be reviewed again. It is important to note that data for individual items within item clusters need to be reviewed and evaluated with respect to the full item cluster, not individually, due to dependencies inherent in the structure of item clusters (versus stand-alone items). Items with questionable statistics should be reviewed by a committee of educators, psychometricians, SEA staff, and content specialists to determine whether an item cluster can be used on an operational form.

The item developers should collect feedback from students, teachers, and reviewers at each stage of the item development process. Especially in the first few years of administering the new NGSS-aligned assessment, prior decisions made about item development may need to be amended as evidence emerges about the assessable content, targets, or evidence needed to support claims of what students are able to do.

Key Takeaways from Chapter Seven

- A large-scale NGSA should attempt to mirror how science content is taught and tested in the classroom and should reflect a “bottom-up” approach to designing an assessment system.
- The first step in item cluster development entails making inferences about what students should know and be able to do at the end of each grade band. This process should leverage the NGSS evidence statements, and may also include further clarification of the PEs and their associated SEPs, DCIs, and CCCs, in order to clearly define the targets and boundaries for assessment.
- In keeping with the recommendations from the BOTA report, states will want to consider a variety of approaches for bundling PEs to support development of rich item clusters that address the full depth and breadth of the NGSS.
- The stimulus is an integral part of the item cluster and should be based on real-world science phenomena and representative of how students learn in the classroom.

CHAPTER EIGHT: ACCESSIBILITY CONSIDERATIONS

Introduction

> *How school systems evaluate the learning derived from educational standards—through high-stakes tests, formative classroom assessments, and informal evaluations of learning during instruction—has a driving influence on educational pathways and equity. Exemplary assessment practice recognizes that there are multiple ways in which students might express their developing understanding, although not all forms of assessment allow for such multiple modes of expression.* (NRC, 2012, p. 289)

Key Questions for this Chapter

- How and when can the principles of universal design for assessment be applied to the development of an NGSA?
- What steps can help ensure the fair and accurate assessment of students in special populations?
- Which extant documents on accessibility and accommodations can be leveraged when designing an NGSA?

While the primary goal of the SAIC is not to identify root causes or provide resources to ensure equal access to quality science curricula and instruction, the SAIC can play a key role in ensuring that students' "true" levels of scientific understanding are accurately assessed. The term "accessible" has come to represent a core goal of the assessment process—namely, to provide all students with equal opportunity to show what they know and can do on an assessment. Sensitivity to student diversity has been a hallmark of this initiative from its outset, resulting in recommendations for monitoring access issues during all phases of assessment design, development, administration, scoring, and reporting.

The SAIC envisions an assessment design that promotes accessibility using different item and text formats, technologies, designs, and accommodations, in order for the design to be as inclusive as possible. It must be clear from the outset that the emerging assessments need to measure the achievement of all students against the same content and achievement standards.

The following sections of this chapter review some of the underlying assumptions of next-generation assessment design that maximizes accessibility, and provide short summaries of seminal documents on accessibility and accommodations for special student subgroups.

NGSS: All Standards, All Students

An effective starting point for an overview of NGSS accessibility issues is Appendix D of the NGSS, entitled "All Standards, All Students." The appendix identifies scientifically underserved populations ("non-dominant" student subgroups) and the challenges that must be addressed to enable science achievement in all student subgroups. These target subgroups include the four student groups defined in the Elementary and Secondary Education Act (ESEA) of 2001 (also known as the No Child Left Behind Act) and the reauthorized ESEA, Section 1111(b)(2)(C)(v):

- economically disadvantaged students,
- students from major racial and ethnic groups,
- students with disabilities, and
- students with limited English proficiency.

Appendix D also examines three student subgroups that have not been previously identified as reporting subgroups for accountability purposes but with historical rationales for being classified as scientifically underserved:

- girls,
- students in alternative education programs, and
- gifted and talented students.

While the appendix does not directly address large-scale summative assessment, it does provide a starting place for exploring opportunities to ensure that emerging measures are developed and implemented using strategies that value student diversity and demonstrate adherence to best-practice recommendations from the science research and education communities. In particular, it is important to note the following:

- The learning expectations of the NGSS are set high and will require shifts in science teaching to ensure that all students have the opportunity to fully engage with tested content—including in high school. Therefore, states will want to support elementary and secondary science teachers as they work together to identify the most effective strategies for including all students—regardless of racial, ethnic, cultural, linguistic, socioeconomic, and gender backgrounds—in NGSS-based instruction and assessment.

- The NGSS call for exploration of new and innovative ways to connect school and home/community, especially for non-dominant student groups. Doing so may require evaluation of economic priorities to ensure that they are fully aligned with the NGSS commitment to “all standards, all students.” To support this mission, states may want to reconsider how school resources are allocated and used, in order to ensure that access to material resources, human capital, and social capital in relation to the NGSS is equally distributed across classrooms.

- Effective implementation of the NGSS for all students, including non-dominant student groups, will require shifts in the education support system. Effective implementation of the NGSS includes support for all students to benefit from high school courses that prepare them for NGSS-aligned assessments. Key components of the education support system include teacher preparation and professional development, administrator support and leadership, public/private/community partnerships, formal and informal classroom experiences that require considerable coordination among community stakeholders, technological capabilities, network infrastructure, cyber-learning opportunities, access to

digital resources, online learning communities, and virtual laboratories. These efforts will enable the seamless integration of equitable, efficient measures of the NGSS.

Ensuring Student Access

The BOTA report (NRC, 2014) notes the following:

> *Achieving equity in the opportunity to learn science will be the responsibility of the entire system, but the assessment system can play a critical role by providing fair and accurate measures of the learning of all students . . . As we have noted, however, it will be challenging to strike the optimal balance in assessing students who are disadvantaged and students whose cultural and linguistic backgrounds may significantly influence their learning experiences in schools.* (p. 221)

The importance of focusing on student diversity issues throughout the assessment development cycle is underscored by one of the report's recommendations:

> *RECOMMENDATION 7-5 Policy makers and other officials who are responsible for the design and development of science assessments should consider the multiple dimensions of diversity--including, but not limited to, culture, language, ethnicity, gender, and disability--so that the formats and presentation of tasks are as accessible and fair to diverse student populations as possible. Individuals with expertise in these areas should be integral participants in assessment development and in the interpretation and reporting of results.* (p. 231)

The authors of the report note that the use of technology in testing may work to mitigate some of the testing challenges faced by certain student subgroups by "translating, defining, or reading aloud words or phrases used in the assessment prompt or offering variable print size that allow students to more readily demonstrate their knowledge of the science being tested" (p. 224).

For developers of NGSS-based assessments, implementing this recommendation will necessitate development of a research agenda that calls for the systematic collection of evidence to support test use with diverse student populations. Developers will want to document the steps taken to prevent the introduction of construct-irrelevant variance (i.e., knowledge or skills not related to the construct of interest) due to test format or presentation of tasks. States implementing these assessments will also want to encourage teachers to expose students to these innovative formats and item types during daily instruction, and to monitor short- and long-term effects of test use to ensure that emerging scores can be used to draw valid inferences about all students.

Applying Principles of Universal Design for Assessment

A research-supported approach to building equity and fairness into the assessment process applies principles from universal design for assessment (UDA). Also known as universal test design (Johnstone, Altman, & Thurlow, 2006), UDA is broadly defined as a set of applied

principles that assist in the design of assessments that minimize and/or mitigate physical, linguistic, cultural, and other barriers to accessibility and threats to test validity. As described in Johnstone et al. (2006), UDA includes seven basic elements:

1. Inclusive assessment population
2. Precisely defined constructs
3. Accessible, non-biased items
4. Amenable to accommodations
5. Simple, clear, and intuitive instructions and procedures
6. Maximum readability and comprehensibility
7. Maximum legibility

For emerging NGSAs, states will want to consider ways in which the principles of UDA can be woven into all phases of development, including solicitation of vendor support via RFPs. As stated in Johnstone et al. (2006),

> *Universally designed assessments are designed and developed from the beginning to allow participation of the widest possible range of students, and to result in valid inferences about performance for all students who participate in the assessment. Universally designed assessments add a dimension of fairness to the testing process.* (p. 81)

Details and examples of these UDA elements can be found in Johnstone et al. (2006), and additional information about application of UDA principles to testing is available on the National Center on Educational Outcomes's Universally Designed Assessments webpage (http://www.cehd.umn.edu/NCEO/TopicAreas/UnivDesign/UnivDesignTopic.htm). Because accessibility, fairness, and opportunity to learn are fundamental to the validity of inferences drawn from test results, developers of NGSS-based assessments may benefit from an intentional focus on *all* students, instead of focusing on particular student subgroups. This focus will be evidenced in fundamental characteristics of the test design, inclusive testing practices, documentation of activities such as bias and sensitivity reviews, adherence to research-supported readability and legibility guidelines, policies that allow for appropriate accommodations, and findings from post-administration analyses, such as differential item functioning.

Inclusion and Accommodations

The principles of UDA—including large-scale, standardized assessments—do not call for every student to be administered a test under the same conditions. Instead, they call for each student to be measured against the same standards, even if the format and structure of the test vary. Thus, all students should be tested in ways that are defensible, intentional, developmentally appropriate, free from bias, and equitable in terms of the standards that are measured and the level of rigor at which each standard is measured.

Smarter Balanced (2013) has addressed the issue of accessibility and the need for accommodation in great depth in the frameworks underlying its assessments. The consortium notes that accommodations are sometimes necessary to ensure that tests are accessible and equitable:

> *Accommodations . . . should be identified that will provide access for students who still need assistance getting around the barriers created by their disabilities or their level of English language proficiency after the assessments themselves are as accessible as possible. For example, where it is appropriate, items may be prepared at different levels of linguistic complexity so that students can have the opportunity to respond to the items that are more relevant for them based on their needs, ensuring that the focal constructs are not altered when making assessments more linguistically accessible.* (p. 13)

In addition, PARCC has implemented a strong accessibility focus during assessment development. Both consortia have produced accessibility and accommodations guidelines (PARCC, 2014; Smarter Balanced, 2015) that have been widely reviewed by state educators as well as by experts in assessing students with disabilities and English learners. NGSA developers will want to review these recently developed guidelines as they move forward with test design, development, and implementation.

Following are additional recommendations specific to particular student subgroups.

Students with Disabilities. The National Assessment Governing Board (hereafter "Governing Board") has published guidelines for the 2009 NAEP Science assessment (NAGB, 2007) that draw specific attention to the needs of students with disabilities (SWDs). In these guidelines, the Governing Board describes a number of specific assessment issues associated with this student subgroup and offers these general principles when crafting test items:

- Avoid layout and design features that could interfere with the ability of the student to understand the requirements and expectations of the item.
- Use plain language.
- Develop items so that they can be used with allowed accommodations.
- Address alternatives for students who are not able to use the equipment and materials necessary for responding to an item.

Smarter Balanced has listed a number of assessment accommodations and designated supports that can be offered for SWDs, including braille, American Sign Language (ASL) video translations, closed captioning, screen masking, color contrast, and speech-to-text dictation devices (Smarter Balanced, 2015). A comparable list of accommodations and designated supports is identified in the *PARCC Accessibility Features and Accommodations Manual* (PARCC, 2014). The manual identifies accommodations for students testing on the computer, while Appendix A of this document identifies comparable accommodations for students using a paper-pencil form. The manual also describes the accessibility features that are intended for optional use by all students (many of which are the same as for the Smarter Balanced assessment; others include blank paper provisions, highlighting capabilities, and pop-up

glossary), as well as accommodations that are made available to SWDs (Presentation, Response, Timing, and Scheduling), including ELs with disabilities.

English Learners. The Governing Board also outlined assessment and accommodation issues for ELs (NAGB, 2007, pp. 211–212). These issues include the following:

- Item reading level should be below grade level, so as not to confound the performance expectation being measured and reading ability.
- Item writers should be aware that words that pose linguistic difficulty in science may not only be technical terms but also logical connectives (e.g., *simultaneously, essentially, in addition to*), common terms that have more than one meaning (e.g., *bond*), and terms that have subtly different meanings in science and in everyday life (e.g., *mass*). Additional issues that are particularly relevant to EL testing include the use of cognates and false cognates.
- Item writers need to be able to determine when language is part of the construct that an item is intended to assess and when language is a source of bias and measurement error. Only where the use of scientific language is at the core of the performance expectation tested should technical words, scientific discourse, scientific notation, and other aspects of scientific language be included in an item.

The Smarter Balanced guidelines for mathematics assessment (2013) make note of the ways in which linguistic complexity can cloud an assessment's ability to measure knowledge and skills:

> *In particular, research has demonstrated that several linguistic features unrelated to mathematics content could slow the reader down, increase the possibility of misinterpretation of mathematics items, and add to the [EL] student's cognitive load, thus interfering with understanding the assessment questions and explaining the outcomes of assessments. Indices of language difficulty that may be unrelated to the mathematics content include unfamiliar (or less commonly used) vocabulary, complex grammatical structures, and styles of discourse that include extra material, conditional clauses, abstractions, and passive voice construction.* (p. 12)

Both the Smarter Balanced guidelines (Smarter Balanced, 2013) and the PARCC manual (PARCC, 2014) provide examples of assessment accommodations and designated supports for ELs, including translated test directions, translated glossaries, and bilingual dictionaries. The Smarter Balanced guidelines caution that students whose English language skills require them to make extensive use of a bilingual dictionary may require additional time to complete their assessment tasks.

Accessibility and Accommodations for Specific Item Types

When developing item or item-cluster templates and prototypes for state NGSS-based assessments, accessibility considerations for special student populations should be considered and described at the outset. These considerations will naturally be informed by the particular

item type. According to the Governing Board (NAGB, 2007), particular stimuli may be more difficult for some SWDs:

> *Certain physical disabilities such as limited vision, hearing, or physical dexterity present challenges in fully assessing the science performance expectations for SWDs. This may be especially the case when special item formats are used to assess students' ability to conduct a scientific investigation (hands-on performance tasks) or to carry out a computer simulation (ICTs). To the extent possible, the assessment items should be designed in such a way as to allow permissible accommodations.*
>
> *In some cases, ICTs may serve in lieu of hands-on performance tasks. This is especially the case with performance assessments where there appears to be an equivalence of hands-on and computer-generated science investigations although not all hands-on tasks can be done on the computer.* (pp. 208–209)

Because of the added linguistic demands for students, extended constructed-response items will require special attention when scoring for ELs. Code switching, native-language phonetics, and native-language sentence structure are possible ways in which EL responses may differ from those of native English language speakers. As a result, the Governing Board recommends the involvement of bilingual teachers and EL experts in all aspects of the benchmarking and scoring process (NAGB, 2007).

Key Takeaways from Chapter Eight

- The term "accessible" has come to represent a core goal of the assessment process—namely, to provide all students with the opportunity to show what they know and can do on an assessment—and student diversity should be taken into consideration during all phases of assessment design, development, administration, scoring, and reporting.
- Universal design, designated supports, and accommodations serve as tools to open assessment to broader audiences.
- Accessibility considerations need to be consistent with guidelines from Smarter Balanced and PARCC, and the entire assessment system, not just an individual test, should be adaptable.
- Item reading level should be below grade level so as not to confound the performance expectation being measured and reading ability.
- Only where the use of scientific language is at the core of the performance expectation tested should technical words, scientific discourse, scientific notation, and other aspects of scientific language be included in an item.

CLOSING COMMENTS

The development of assessments aligned to the NGSS will provide an unprecedented opportunity for state stakeholders to actualize the potential embedded in these standards. When an evidence-based approach is fully implemented, decisions about content, construct, and evidence provide critical information to teachers to guide instruction and promote student learning. The release of guidelines developed by states for states seeking to develop and implement next-generation science assessments (NGSS Lead States, 2013) and the recommendations that have emerged from the National Research Council (NRC, 2012, 2014) have further enabled this effort. In addition, the focus on accessibility that has characterized the Race to the Top assessment consortia has yielded new strategies for promoting fair, inclusive assessments from which valid inferences can be drawn about what students know and can do.

Just as the consortia assessments have presented challenges as well as opportunities, the NGSS-based assessments will be a catalyst for reevaluating the standards for practice in terms of both science instruction and assessment. Lessons learned and systems improved as a result of these earlier initiatives will be invaluable to NGSA developers. For example, many synergies are evident among mathematics tasks, ELA research tasks, and science tasks, and these synergies can be tapped to promote optimal leveraging of resources. It is hoped that, with attention to the recommendations provided in this report for the design, development, and implementation of technically sound, fair, and feasible science assessments, states can better envision a coherent approach, including all content areas, during decision-making about student assessment.

REFERENCES

American Association for the Advancement of Science (AAAS). (1989). *Science for all Americans: A Project 2061 report*. Washington, DC: Author.

American Association for the Advancement of Science (AAAS). (1993). *Benchmarks for science literacy*. New York, NY: Oxford University Press.

American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME). (2014). *Standards for educational and psychological testing* (4th ed.). Washington, DC: APA. Retrieved from http://www.apa.org/science/programs/testing/standards.aspx

Catley, K., Lehrer, R., & Reiser, B. (2005). *Tracing a prospective learning progression for developing understanding of evolution*. Paper commissioned by the Committee on Test Design for K–12 Science Achievement, Board on Testing and Assessment, Center for Education, National Research Council, Washington, DC.

College Board. (2011). *AP Insight*. New York, NY: Author. Retrieved from https://apinsight.collegeboard.org/

Davenport, J., Silberglitt, M., & Olson, A. (2013). *In touch with molecules: Improving student learning with innovative molecular models.* Retrieved from http://www.rcsb.org/pdb/general_information/news_publications/newsletters/2013q3/corner.html

English Language Proficiency Assessment for the 21st Century (ELPA21). (2014). *ELPA21 field test accessibility and accommodations manual*. Retrieved from http://www.arkansased.org/public/userfiles/Learning_Services/English%20Language%20Learners/ELPA21/ELPA21_Accessibility_and_Accommodations_Manual.pdf

Fraenkel, J. R., & Wallen, N. E. (2000). *How to design and evaluate research in education* (4th ed.). San Francisco, CA: McGraw-Hill.

Gonzales, P., Guzmán, J. C., Partelow, L., Pahlke, E., Jocelyn, L., Kastberg, D., & Williams, T. (2004). *Highlights from the Trends in International Mathematics and Science Study (TIMSS) 2003* (NCES 2005-005). U.S. Department of Education, National Center for Education Statistics. Washington, DC: U.S. Government Printing Office.

Gorin, J. S., & Mislevy, R. J. (2013, September). *Inherent measurement challenges in the Next Generation Science Standards for both formative and summative assessment.* Paper presented at the Invitational Research Symposium on Science Assessment, Washington, DC. Retrieved from http://www.k12center.org/rsc/pdf/gorin-mislevy.pdf

Hilton, M., & Honey, M. A. (Eds.). (2011). *Learning science through computer games and simulations.* Chicago, IL: National Academies Press.

Johnstone, C. J., Altman, J., & Thurlow, M. (2006). *A state guide to the development of universally designed assessments.* Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Retrieved from http://www.cehd.umn.edu/nceo/OnlinePubs/StateGuideUD/default.htm

Kane, M. (2002, Spring). Validating high stakes testing programs. *Educational Measurement: Issues & Practice*, *21*(1), 31–41.

Lee, O. (1999). Equity implications based on the conceptions of science achievement in major reform documents. *Review of Educational Research*, *69*, 83–115.

McDonald, R. P. (1999). *Test theory: A unified treatment*. Mahwah, NJ: Erlbaum.

Metcalf, S., Kamarainen, A., Tutwiler, M. S., Grotzer, T., & Dede, C. (2011). Ecosystem science learning via multi-user virtual environments. *International Journal of Gaming and Computer-Mediated Simulations*, *3*(1), 86–90.

Metz, K. E. (1995). Reassessment of developmental constraints on children's science instruction. *Review of Educational Research*, *65*(2), 93–127.

Mislevy, R. (2007). Implications for evidence-centered assessment design for educational assessment. *Educational Measurement: Issues & Practice, 25*, 6–20.

Mullis, I. V. S., Martin, M. O., Smith, T. A., Garden, R. A., Gregory, K. D., Gonzalez, E. J., Chrostowski, S. J., & O'Connor, K. M. (2001). *TIMSS assessment frameworks and specifications 2003*. Boston, MA: Boston College, International Study Center, Lynch School of Education.

National Assessment Governing Board (NAGB). (2007). *Science assessment and item specifications for the 2009 National Assessment of Educational Progress*. Retrieved from http://www.nagb.org/content/nagb/assets/documents/publications/frameworks/science/2009-science-specification.pdf

National Research Council (NRC). (1996). *National science education standards*. National Committee on Science Education Standards and Assessment. Coordinating Council for Education. Washington, DC: National Academy Press.

National Research Council (NRC). (1999). *How people learn: Brain, mind, experience, and school*. Committee on Developments in the Science of Learning. J. D. Bransford, A. L. Brown, & R. R. Cocking (Eds.). Commission on Behavioral and Social Sciences and Education. Washington, DC: National Academy Press.

National Research Council (NRC). (2001). *Knowing what students know: The science and design of educational assessment*. Committee on the Foundations of Assessment. J. Pellegrino, N. Chudowsky, & R. Glaser (Eds.). Board on Testing and Assessment, Center for Education, Division of Behavioral and Social Sciences and Education. Washington, DC: National Academy Press.

National Research Council (NRC). (2005). *How students learn: History, mathematics, and science in the classroom*. Committee on How People Learn, a Targeted Report for Teachers. M. S. Donovan & J. D. Bransford (Eds.). Division of Behavioral and Social Sciences and Education. Washington, DC: National Academy Press.

National Research Council (NRC). (2006). *Systems for state science assessment.* Committee on Test Design for K–12 Science Achievement. M. R. Wilson & M. W. Bertenthal (Eds.). Board on Testing and Assessment, Center for Education, Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

National Research Council (NRC). (2012). *A framework for K–12 science education: Practices, crosscutting concepts, and core ideas.* Committee on Conceptual Framework for the 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 (NRC). (2014). *Developing assessments for the Next Generation Science Standards*. Committee on Developing Assessments of Science Proficiency in K–12. Board on Testing and Assessment and Board on Science Education, J. W. Pellegrino, M. R. Wilson, J. A. Koenig, & A. S. Beatty (Eds.). Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

National Science Teachers Association (NSTA). (2010). *Science anchors project*. Retrieved from http://www.nsta.org/involved/cse/scienceanchors.aspx

Next Generation Science Standards (NGSS) Network. (2015a). *NGSS evidence statements front matter.* Retrieved from http://nextgenscience.org/sites/ngss/files/Front%20Matter%20Evidence%20Statements%20PDF%20Jan%202015.pdf

Next Generation Science Standards (NGSS) Network. (2015b). *NGSS high school evidence statements.* Retrieved from http://www.nextgenscience.org/ngss-high-school-evidence-statements

NGSS Lead States. (2013). *Next generation science standards: For states, by states.* Washington, DC: Achieve, Inc.

Organisation for Economic Cooperation and Development (OECD). (2005). *PISA 2006 scientific literacy framework.* Paris: OECD/PISA.

Partnership for Assessment of Readiness for College and Careers (PARCC). (2013). *PARCC item development technical guide (Table 5.3)*. Retrieved from https://parccsharepoint.org/Public_Access/Diagnostic%20and%20K-1%20Assessment%20ITN%20-%20June%202013%20-%20Reference%20Documentation/June%202013%20DRAFT_PARCC%20Item%20Development%20Technical%20Guide%2020130627.pdf

Partnership for Assessment of Readiness for College and Careers (PARCC). (2014). *PARCC accessibility features and accommodations manual*. Retrieved from http://www.parcconline.org/sites/parcc/files/parcc-accessibility-features-accommodations-manual-11-14_final.pdf

Quellmalz, E. S., Silberglitt, M. D., & Timms, M. (2011). *How can simulations be components of balanced state science assessment Systems?* (Policy Brief). San Francisco, CA: WestEd.

Shute, V., & Ventura, M. (2013). *Stealth assessment: Measuring and supporting learning in video games*. Cambridge, MA, & London, England: The MIT Press. Retrieved from http://myweb.fsu.edu/vshute/pdf/Stealth_Assessment.pdf

Sireci, S. (2007). On validity theory and test validation. *Educational Researcher, 36*(8), 477–481.

Smarter Balanced Assessment Consortium (Smarter Balanced). (2012). *Smarter Balanced Assessment Consortium: Performance task specifications.* Retrieved from http://www.smarterbalanced.org/wordpress/wp-content/uploads/2012/05/TaskItemSpecifications/PerformanceTasks/PerformanceTasksSpecifications.pdf

Smarter Balanced Assessment Consortium (Smarter Balanced). (2013). *Content specifications for the summative assessment of the Common Core State Standards for Mathematics.* Retrieved from http://www.smarterbalanced.org/wordpress/wp-content/uploads/2011/12/Math-Content-Specifications.pdf

Smarter Balanced Assessment Consortium (Smarter Balanced). (2014). *Getting students ready for the field test: Information on item and response types*. Retrieved from http://sbac.portal.airast.org/wp-content/uploads/2014/03/Student-Item-and-Response-Types.pdf

Smarter Balanced Assessment Consortium (Smarter Balanced). (2015). *Usability, accessibility, and accommodations guidelines*. Retrieved from http://www.smarterbalanced.org/wordpress/wp-content/uploads/2014/08/SmarterBalanced_Guidelines.pdf

Smith, C., Wiser, M., Anderson, C. W., Krajcik, J., & Coppola, B. (2004). *Implications of research on children's learning for assessment: Matter and atomic molecular theory*. Paper commissioned by the Committee on Test Design for K–12 Science Achievement, Board on Testing and Assessment, Center for Education, National Research Council, Washington, DC.

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes.* Cambridge, MA: Harvard University Press.

APPENDICES

APPENDIX A: Framework for Collecting Evidence for Test Validation, by Work Phase

APPENDIX B: SAIC State Survey #1 Summary Report

APPENDIX C: SAIC State Survey #2 Summary Report

APPENDIX D: SAIC Assessment Framework Survey (#3) and SAIC Assessment Framework Workbook Feedback Summary Report

APPENDIX E: Annotated Resources for the Development of Assessments for the NGSS

APPENDIX F: Examples from SimScientists and NAEP

APPENDIX G: Documentation and Timeline of Key Activities

APPENDIX A: Framework for Collecting Evidence for Test Validation, by Work Phase[1]

States may benefit from a framework that provides guidance about the specific types of evidence that should be collected during each of the following phases: (I) Test Design and Development; (II) Field Testing; (III) Test Administration; (IV) Scoring; and (V) Reporting and Interpretation of Scores.

Table A-1. Framework for Collecting Evidence for Test Validation, by Work Phase

PHASE I: TEST DESIGN AND DEVELOPMENT[2]

A. Validity: Item Level

Category of Evidence	Operational Definition	Comments About Documentation
Validity: Construct	A construct is the concept or the characteristic that a test is designed to measure. Construct validity indicates that the test scores reflect the examinee's standing on the psychological construct measured by the test.	Ensure that test captures all elements of construct as intended
Test purpose	The reason or object for which an assessment is designed, developed, and intended to be used.	Clearly stated purpose related to range of appropriate purposes for testing (e.g., placement, classification, measuring growth/achievement)
Population/Classification	The set of examinees for whom the test is intended for the purpose(s) stated.	Clearly defined population; geographical location
Theoretical foundation/framework	The underlying framework, model, or perspective that defines the domain being measured and how best to measure it.	Clearly stated, coherent, current/accepted theories
Universal Design (UD)	Incorporating considerations and features into an instrument to promote its accessibility and validity for the widest range of examinees, including examinees with disabilities and examinees with limited English proficiency.	Specific and/or explicit evidence of application of UD principles during design and development phase

[1] Unless otherwise noted, guidelines provided in this framework are drawn from *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014).

[2] Documentation is required at both the item and test levels, so evidence collection strategies are provided separately.

Category of Evidence	Operational Definition	Comments About Documentation
Readability	The measure of the complexity of the language in the text and directions.	Expert judgment; documentation; number; statement that text is grade appropriate and appropriate for the population and purpose; protocol; readability formulae (e.g., Lexile, Dale-Chall)
Validity: Content	Content is the set of behaviors, knowledge, skills, abilities, attitudes, or other characteristics to be measured by a test. Content validity indicates the degree to which the items measure the content (i.e., knowledge/skills/abilities).	Ensure that test captures all elements of content as intended
Alignment (items-to-standards)	A procedure for ensuring that test items under development are aligned to existing content standards. Ex post facto alignment is a process—usually a formal study—for evaluating whether existing items are aligned to existing content standards.	Alignment studies completed using appropriate unit(s) of analysis and appropriate model; explanation of process or results (including limitations). In-process alignment may be done by writers, editors, or other developers and expert reviewers during the item development process. Ex post facto alignment should be done by independent experts in assessment, standards, and relevant content areas. Alignment procedures and studies should look for appropriateness of item content and cognitive level as described in individual standards, and coverage (breadth and depth) of the set of standards.
Expert judgment	The use of individuals with relevant knowledge and background for verifying the degree to which the test's questions are representative of the content that the test questions are intended to assess.	Credible experts; methodology/protocol described; explanation of findings; distracter analysis
p-values/point biserials	*p*-values are the probability of correctly answering an item. Point biserials are correlations between the total test score and item score.	High *p*-value reflects an "easy" item; a range of difficulty appropriate to test purpose is needed. Discussion of how *p*-values relate to the items' ability to discriminate among the target (sub)groups of examinees
IRT/Item fit	The probability of a correct response to an examinee's ability level on the construct (latent trait).	Description of model; explanation of results; Item Characteristic Curve (ICC) One-, two-, or three-parameter IRTs are okay.

Category of Evidence	Operational Definition	Comments About Documentation
Structural equation modeling	The relationship between the construct and the measurable factors that affect it and traces the relationships within a network of variables.	Report on the relative contribution of each factor examined; support/verification of predictions of the relationship of the construct to the measurable factors
t-tests	The statistical hypothesis tests that examine the equality of the means of two variables or two groups on the same variable (Fraenkel & Wallen, 2000).	Value for *t* statistic and its significance level; explanation of results
ANOVA	A statistical procedure that examines the equality of differences between the means of more than two groups and the interaction among effects (Fraenkel & Wallen, 2000).	Value for *f* statistic and its significance level; explanation of results
Factor analysis	A statistical technique to determine if multiple variables can be described by a few factors (unidimensionality) (Fraenkel & Wallen, 2000).	Correlations (factor loadings); explanation of results
Bias and Sensitivity (Linguistic, Ethnicity/Race, Cultural/Religious, Geographic, SES, Disability, Gender)	Bias is the presence of construct-irrelevant elements that potentially advantage or disadvantage any examinee subgroup. Sensitivity is the presence of content that evokes an emotional response that inhibits examinees' ability to demonstrate what they know and can do.	Expert review of item content and wording as well as associated stimuli
Expert review	A method in which individuals with knowledge of (and often, membership in) a subgroup evaluate the items in a test or item pool to ensure that the items do not give unfair advantage or disadvantage to any examinee subgroup.	Review by representative experts and/or members of the community/target population(s)

B. Validity: Test Level

Category of Evidence	Operational Definition	Comments About Documentation
Validity: Construct	A construct is the concept or the characteristic that a test is designed to measure. Construct validity indicates that the test scores reflect the examinee's standing on the psychological construct measured by the test.	Ensure that test captures all elements of construct as intended

Category of Evidence	Operational Definition	Comments About Documentation
Equivalence/ Comparability	Two or more tests/test forms that measure the same construct and/or are interchangeable.	Description of method of analysis (typically involves expert judgment; unit of analysis reflects the entire construct); subtest intercorrelations
Multitrait/Multimethod (MTMM)/Subtest intercorrelation	MTMM display evidence of the relationships/factors (convergence or divergence) related to examinee performance that can be compared so that the validity of the assessment can be determined/evaluated. Subtest intercorrelation is evidence that the pieces of the test are measuring the same construct (e.g., subtests within the reading section). Note: Subtest intercorrelation may appear as evidence of internal consistency. However, we believe that there is other, stronger evidence for internal consistency. Therefore, the recommendation is that subtest intercorrelation be presented as evidence of construct validity.	Correlation table or MTMM matrix
Validity: Content	Content is the set of behaviors, knowledge, skills, abilities, attitudes, or other characteristics to be measured by a test. Content validity indicates the degree to which the items measure the content (i.e., knowledge/skills/ abilities).	Ensure that test captures all elements of content as intended
Test blueprint	The structure and contents of a test, including the relative weighting or distribution of strands of content.	Table or chart showing the content distribution, item type, etc.
Alignment (test form–to–blueprint)	The degree to which test form reflects the intended breadth, depth, and emphasis of content specified in the test blueprint.	Alignment studies done (independent); appropriate unit(s) of analysis and model/appropriate dimensions evaluated; explanation of results (including limitations)
Descriptive statistics	The summary measures of a distribution of scores, providing information about central tendency, location, and variability.	Mean; standard deviation; *N*; explanation of results (e.g., evidence that field test results were used to select appropriate items)
IRT/Test fit	The proportion of correct responses to an examinee's ability level on the construct (latent trait).	Description of model; explanation of results Test Characteristic Curve (TCC) One-, two-, or three-parameter IRTs are okay.

Category of Evidence	Operational Definition	Comments About Documentation
Linking/Equating	Linking is putting two or more tests on a common scale to show that the scores can be compared. Equating is the term used to define a special case of linking when the two tests are essentially parallel.	Report of linking/equating error; description of linking/equating methods (including assumptions, feasibility); reference to dimensionality; factor analysis; correlations; DIF; structural equation modeling
Validity: Criterion (Predictive/Concurrent)	Criterion validity is the extent of the relationship of a test score to an external criterion. The extent to which a score can predict the value of a criterion measure is predictive validity (McDonald, 1999). Concurrent validity compares scores of two instruments administered at about the same time (Fraenkel & Wallen, 2000).	
Cross tabulations	The tabular representations of the relationships (categorical or continuous) among two or more different measures.	Description of relationships; explanation of results; description of measures; includes expectancy tables
Pearson correlation	The number between -1 and 1 that indicates the degree to which two quantitative variables are related (shows strength and direction of relationship).	Correlation coefficient; description of measures; explanation of results
Validity: Consequential	Consequential validity is the degree to which results are used in a manner consistent with the intended purpose and uses of the assessment.	Ensure that stated test purpose matches test use and anticipate plausible unintended outcomes
Use of results	The intended and unintended ways in which test scores are analyzed, reported, and/or brought into service to inform and facilitate decision-making (i.e., diagnosis, evaluation, classification, selection, promotion, placement, and entry/exit).	Proficiency level descriptors; description of range of levels of performance; fidelity between stated purpose of assessment and how results are reported/guidelines for use of results—look at stated purpose of the assessment along with, for example, sample reports, scoring outcomes/results; includes item release strategy

C. Reliability: Item Level

Category of Evidence	Operational Definition	Comments About Documentation
Reliability: Internal Consistency	Internal consistency is the extent to which items on a test measure a construct consistently.	
Coefficient alpha	An internal consistency reliability coefficient based on the number of parts into which the test is partitioned (e.g., items, subtests, or raters), the interrelationships of the parts, and the total test score variance (AERA, APA, & NCME, 2014).	
KR-21	A reliability formula based on the number of items on a test, the mean, and the standard deviation (between 0 and 1). It should be interpreted like a correlation coefficient.	
Test length/Power estimates	The statistical measures that indicate the probability that the null hypothesis will be rejected when there is a true difference (no Type II error).	Probability that the test will correctly lead to the conclusion that there is a difference in performance when an alternative hypothesis is specified (t-test, ANOVA, chi-square) Number of items for entire test as well as reporting category (not format or number of pages)
Split-half	An internal consistency reliability coefficient obtained by using half of the items on a test to yield one score and the other half of the items to yield a second, independent score.	Correlation coefficient with Spearman-Brown

D. Reliability: Test Level

Category of Evidence	Operational Definition	Comments About Documentation
Reliability: Stability and Consistency	Stability is the extent to which scores on a test are essentially invariant over time. Consistency is the extent to which multiple forms of a test measure a construct consistently.	
Standard Error of Measurement (SEM)/Confidence Intervals	SEM indicates the dispersion of measurement errors when estimating examinees' true scores from their observed test scores. Confidence intervals are bands defining score zones in which the true scores are believed to lie, with a given level of confidence.	

Category of Evidence	Operational Definition	Comments About Documentation
Test-retest	A correlational measure based on the administration of the same test twice to the same group of examinees after a (brief) time interval has elapsed.	Time between administrations; correlation coefficient
Alternate Form	Two or more tests are designed to measure the same construct (McDonald, 1999).	Correlation; explanation of results
Reliability: Generalizability	Generalizability is the dependability of an observed score (of an individual or group of individuals) and the accuracy with which this observed score generalizes (to an individual's overall performance or to a larger group).	
G coefficient	A reliability index encompassing one or more independent sources of error. It is formed as the ratio of (a) the sum of variances that are considered components of test score variance in the setting under study to (b) the aforementioned sum plus the weighted sum of variances attributable to various error sources in this setting.	Includes SEM, confidence intervals
Reliability: Classification Consistency	Classification consistency is the property of an instrument whereby classification decisions based on the instrument's scores are accurate and consistent. At the system level, classification consistency implies that decisions about performance drawn across measures/processes are consistent.	Percentage of agreement; rationale; must include explanation of how data are used; discriminant analysis; mean scores and standard deviations for each performance level; correlation (kappa)
Correlation coefficient	A statistical measure that compares the strength and degree of agreement between two binding classification decisions.	Correlation
Percent correspondence	The degree of agreement between two binding classification determinants.	Percent of agreement, classification error
Classification error	The likelihood that an examinee is classified correctly or incorrectly.	Probability of classification or misclassification

PHASE II: FIELD TESTING

Category of Evidence	Operational Definition	Comments About Documentation
Validity: Content	Content validity is the degree to which the items on an instrument are representative of the questions that could be asked about the content.	Not embedded: the degree to which the items are representative of the questions that could be asked about the content; the degree to which the pool of items contains the breadth of depth of the content/standards that are assessed Embedded: the degree to which the forms reflect the requirements of the test blueprint (this may occur over time)
Blueprint	The structure and contents of a field test, including the relative weighting or distribution of strands of content.	Could occur over multiple administrations if specified. Table or chart showing the content distribution and item type, etc. Make transparent any changes in content assessed.
Sampling	The process of selecting a number of examinees from a population in such a way that the selected examinees are representative of the population intended to be tested.	Method (random sampling, as opposed to convenience sampling, is preferred); description of sample; characteristics; the quality of sampling is that it shows fidelity to the assessment's intended purpose (fidelity is the degree to which the norming population is representative of an instrument's identified target population); sample size (*N*) is large enough to cover the range of examinees/population characteristics targeted (e.g., 30 examinees per "cell")
Norming	The use of field test results to make decisions about test performance with respect to a reference group that permits meaningful comparisons to other individuals or generalizations to the population.	Descriptive statistics or IRT statistics; how the items performed for the range of examinees (degree to which items performed with respect to the purpose of the test and the population tested); should have a purposive sample that shows oversampling of target subgroups tending to have low numbers and include calibration for these subgroups

PHASE III: TEST ADMINISTRATION

Category of Evidence	Operational Definition	Comments About Documentation
Validity: Construct	A construct is the concept or the characteristic that a test is designed to measure. Construct validity indicates that the test scores reflect the examinee's standing on the psychological construct measured by the test.	Test administration (e.g., accommodations provided, fidelity to standard protocol) does not alter the construct being tested—for example, reading aloud the reading comprehension section of the assessment alters the construct
Accommodations	The changes made to the test itself or its administration procedures in order to accommodate examinees who require such changes in order to be able to show what they know and are able to do. In theory, changes do not alter the construct and are intended to minimize the influence of construct-irrelevant factors.	Theoretically, allowed accommodations do not alter the construct assessed and do not affect the reliability of the measure
Fidelity	The degree to which the protocol for standardized test administration is followed.	Test administration conditions/procedures do not alter the construct; make transparent any changes in administration guidelines
Standardization	The rules and specifications for testing procedures, which are intended to ensure that testing conditions are the same for all examinees.	Level of detail and degree to which rules ensure standardized testing conditions
Validity: Consequential (Test Security)	Consequential validity is the degree to which results are used in a manner consistent with the intended purpose and uses of the assessment. In terms of security, scores can be used/interpreted in a manner consistent with the test's purpose.	Security protocol for development, administration, scoring, and reporting (nondisclosure, confidentiality, erasure analysis)
Protocols	The systems established to prevent viewing, publication, or unauthorized copying of test materials.	Systematic; clear; adequate/appropriate for ensuring security (including limiting access/distribution)

PHASE IV: SCORING

Category of Evidence	Operational Definition	Comments About Documentation
Validity: Content	Content validity is the degree to which the items on an instrument are representative of the questions that could be asked about the content.	For scoring, content validity is the degree to which the test content is meaningfully measured quantitatively or qualitatively.

Category of Evidence	Operational Definition	Comments About Documentation
Rubric	The established criteria, including rules, principles, and illustrations, used in scoring responses.	Rubric standardizes the scoring process; levels/elements within a rubric are discernible and real. Make transparent any changes in scoring procedures.
Scale	The means for comparing scores across performance/examinees in which scores are arrayed on a numerical scale with the intention of quantifying examinee performances.	Meaningful differentiation of examinee performance; appropriate range; lends itself to evaluation of examinee performance
Standard setting (cut score and proficiency levels)	The method/process for establishing points on a scale such that scores at or above a point are interpreted differently from scores below that point (NCES).	Defensible; cut scores are neither arbitrary nor capricious; method(s)/experts used; SEM; number of participants
Training of scorers/Scoring protocol	The established system with materials for training scorers.	Clear protocol; evidence of calibration; anchor papers, etc. (as appropriate); monitoring/auditing procedure
Reliability: Inter-rater Reliability	Inter-rater reliability is an approach to reliability in which a researcher compares the scores generated by two (or more) raters.	Level of agreement; stated rating process, and degree of fidelity to rating process
Correlation (kappa)	A statistical measure that compares the strength and degree of agreement between two (or more) different raters.	Coefficient
Percent correspondence	A measure of inter-rater agreement, usually reported at the item level, defined as the share of examinee responses on which multiple raters agree.	Percent of agreement, classification error, rationale. Agreement can also be defined as within one rating category, within two, etc.
Bias and Sensitivity	Bias is the presence of construct-irrelevant elements that potentially advantage or disadvantage any examinee subgroup. Sensitivity is the presence of content that evokes an emotional response that inhibits examinees' ability to demonstrate what they know and can do.	DIF analyses at subgroup level (e.g., linguistic, ethnicity/race, cultural/religious, geographic, SES, disability, gender)
DIF	A statistical property of a test item in which different but otherwise comparable groups of examinees who have the same total test score have different average item scores or, in some cases, different response patterns.	Significance level and discussion of interpretation

PHASE V: REPORTING AND INTERPRETATION OF SCORES

Category of Evidence	Operational Definition	Comments About Documentation
Validity: Consequential	Consequential validity is the degree to which results are used in a manner consistent with the intended purpose and uses of the assessment.	Monitor intended and unintended outcomes and/or test misuse
Reporting category	The categories/labels associated with scores (e.g., standard, objective, examinee-level expectation, examinee level, school level, state level, performance level).	Score reports have an appropriate level of granularity/detail (unit of analysis); consistent with purpose of assessment and intended use of results; clarity and coherence of presentation
N	The number of examinees tested.	Subgroup numbers; minimum *N* (which examinees/groups are excluded)
Central tendency/Variation	The average or typical score attained by a group of subjects.	Means (averages)/medians (middle scores); standard deviation (variability from the mean); range; shape of distribution; frequencies
Effect size	A statistic representing the magnitude of an effect and its practical significance so that outcomes of the assessment(s) can be compared to other measures for validation (*N* taking ELP tests tends to be small; therefore, effect size is a means for examining practical significance for the population of examinees even with an absence of statistical significance).	Method/formula
Use of results	The intended and unintended ways in which test scores are analyzed, reported, and/or brought into service to inform and facilitate decision-making (i.e., diagnosis, evaluation, classification, selection, promotion, placement, and entry/exit).	Fidelity between stated purpose of assessment and the actual use of the assessment; guidelines for how results should be interpreted, reported, and used—look at, for example, sample reports, scoring outcomes/results; includes item release strategy

High-priority evidence categories are listed in Table A-2, with a set of guiding questions that states might consider asking as they seek to collect these key pieces of evidence and examples of appropriate evidence for that category. The intent is to ensure that states have collected sufficient information to clearly communicate their measurement model or approach to educators, parents, policymakers, and other stakeholders.

Table A-5. Tier 1, High-Priority Elements of Evidence for Emerging NGSS-Based Assessments, by Evidence Category

Category of Evidence	Guiding Questions	Examples of Evidence
Construct Validity	Are the test purpose, target audience, and intended uses clearly specified? What evidence suggests that the test captures all elements of the intended science constructs as intended? What theory of action or theoretical foundation describes the connection between test results and intended claims? What evidence supports the intended interpretations of scores from these measures? What information was considered during identification of the most appropriate item types for use on this assessment? What measurement theory of action or theoretical foundation supports those decisions? Did evidence of construct underrepresentation or introduction of construct-irrelevant variance emerge? What efforts were taken to ensure standardized administration conditions? What steps were taken to maintain test security?	Communiqués to stakeholders that communicate test purpose, target audience, and intended uses Theory-of-action diagrams Relevant research citations in documentation Documentation of steps taken during item development Evidence of teacher participation during all steps of work Findings from alignment studies that verify depth of knowledge measured Findings from dimensionality studies Administration guides Test security procedures and protocols Technical reports

Category of Evidence	Guiding Questions	Examples of Evidence
Content Validity	What evidence suggests that test items (or item clusters) are measuring the full depth and breadth of the NGSS? In what ways was matrix sampling used, if at all, to promote full coverage? In what ways were state educators at key grades involved in decision-making about the appropriateness of content assessed? How are states ensuring that students have the opportunity to learn tested content? Were items appropriately field tested prior to operational testing, and was feedback collected from educators and students? If multiple test forms are used, how are forms linked or equated?	Findings from studies that verify alignment to intended content (will require a novel approach due to multidimensionality of the standards) Documentation of content sampling plan Recommendations from content reviews Findings from expert reviews Item-level statistics collected during field testing, such as p-values, point-biserials, and item characteristic curves Test-level descriptive statistics, such as mean, standard deviation, *N*-counts, and test characteristic curves Reports on equating methodology and findings Findings from tests of fit, structural equation modeling, ANOVAs, or factor analyses Test blueprints Field-test sampling plans and findings

Category of Evidence	Guiding Questions	Examples of Evidence
Consequential Validity	In what ways were plausible unintended outcomes considered? How were achievement levels and cut scores determined? How are results being used? Was the test susceptible to misuse? Have any issues related to ethics or equity emerged in relation to this assessment? To what extent are results meeting the needs of or benefiting different stakeholder groups? Has new evidence emerged that calls into question the current interpretation of test scores?	Findings from surveys of students, teachers, or parents Documentation from standard setting (participants, methodology, outcomes) Documentation of test security violations
Reliability	What evidence suggests that the tests meet industry standards for reliability as a measure of students' annual achievement in relation to the NGSS? Were items appropriately field tested prior to operational testing?	Documentation of scoring methods Estimates of internal consistency Documentation of test length Information about scale used and range of student performance Scoring guides and rubrics Manuals for training hand-scorers Findings from inter-rater reliability, split-half, test-retest, alternate forms, or generalizability studies Reporting of standard error of measurement or use of confidence intervals

Category of Evidence	Guiding Questions	Examples of Evidence
Fairness	What evidence was collected to support use of this assessment with the target population? What efforts were taken to examine the appropriateness of this measure for students from diverse geographic, cultural, linguistic, religious, or socioeconomic backgrounds? What evidence suggests that this measure is appropriate for SWDs or ELs?	Documentation of application of principles of universal design for assessment during development Recommendations from bias/sensitivity reviews Results from differential item functioning (DIF) analyses Findings from expert review Documentation of population included for field testing Description of allowable test accommodations
Feasibility	Are the assessments administered, responses scored, and results reported in cost-efficient and responsible ways? How much time is required to administer this test at each grade? What special skills are required for test administrators and scorers? In what ways was technology used to promote administration, scoring, or reporting efficiencies?	Findings from surveys of students, teachers, or parents Administration windows and testing time estimates Findings from impact analyses, utilization studies, or cost-benefit analyses

APPENDIX B: SAIC State Survey #1 Summary Report

Context and Stakeholder Activities in Support of the Assessment Framework

State input—through a series of targeted online surveys, biweekly meetings, and in-person meetings—was a deliberate and driving factor and served to inform and shape the structure and content of the emerging Assessment Framework. While care was taken to address the priorities expressed by the SAIC members, development of the Assessment Framework is an iterative process, and further input from the SAIC members will continue to shape subsequent iterations. Ultimately, this Assessment Framework and the development of prototype item clusters will help to guide the development of the Item Specifications Guidelines and a diverse and robust item pool of NGSS-aligned items (specifically, item clusters, as explained in more detail in **Chapter Four**).

In order to fully understand the needs of each member of the Collaborative, WestEd developed and distributed a survey to the individual SAIC members that comprise the Science Assessment Item Collaborative (SAIC). A total of nine SAIC members and one U.S. territory responded to SAIC State Survey #1. (Within this appendix, "state" is used to mean state or territory.) WestEd received responses from assessment personnel from the members of the Collaborative. This report aggregates these responses and presents the results and findings from the first SAIC survey.

Question 1: Please provide any feedback that you may have on the outline for the Assessment Framework presented during the meeting on 2/13/15. Survey results are provided by Assessment Framework section.

Assessable Content

- One SAIC member inquired as to whether this section would present all of the standards or simply provide a description of the layout of the standards, including the architecture and a description of the appendices.
- One SAIC member desired a confirmation as to whether the Framework will, by the time it is released, contain all evidence statements for all of the standards or will contain a description of the evidence statements and a few examples.
- One SAIC member asked what consideration would be given to individual state adaptations.
- One SAIC member suggested that the Collaborative should consider drafting and presenting key assessment principles, as is reflected by the need for this new Framework. These principles can be adapted from those of the NRC framework, the NGSS, and other sources, as guiding principles for the implementation of this Framework.
- One SAIC member noted that the NGSS and the K–12 Science Framework were well vetted and appropriate. There is concern that the quick timeline and the recent release of a handful of evidence statements do not lend confidence.
- One SAIC member felt that evidence statements work well as specifications for item writers in understanding what students should know and be able to do.
- One SAIC member felt that the Framework should take into consideration that the Collaborative is not designing an end-of-course biology assessment (all high school PEs need to be addressed, and the full NGSS PEs—three dimensions—need to be addressed, including DCIs, CCCs, and SEPs).

Measurement Model

- One SAIC member desired greater clarification on the levels of assessment (e.g., interim vs. summative, classroom vs. large-scale).
- One SAIC member inquired as to whether this Framework would present measurement claims for all standards or just a description of a few claims, including a few examples.
- One SAIC member inquired as to whether the Framework would focus on all assessments at all levels or only large-scale summative assessments.
- One SAIC member expressed interest in a measurement model in order of importance: summative, interim, then classroom.
- One SAIC member wants the Collaborative to define what is meant by "classroom assessment"—there can be formative, interim, and/or summative at the classroom level; classrooms should not equal formative only; what is possible to report out will be very important, and how it is framed will be equally so.

Item Types and Clustering

- One SAIC member expressed an interest in having scenario-based (clustered) and enhanced problem-solving (MSCR) items replace generic MCs.
- One SAIC member expressed interest in whether the model supports scenario-based only or whether some items will need to be stand-alone in order to meet validity concerns.
- One SAIC member stressed that the item types should be performance-based, CR, SR, and technology-enhanced (not multiple-choice).
- One SAIC member commented that the item types and clustering made sense in light of the content, the science standards, and the work that Smarter Balanced has done in mathematics.
- One SAIC member urged the Collaborative to balance the need for a high percentage of automated scoring, but still allow for some hand-scored items.
- One SAIC member inquired as to what specifically would be presented in this section. Will this section provide descriptions of how individual items would look like and how clustering will be and/or should be done? What will be the difference between the individual-items section and model items?

Item Specifications Guidelines

- One SAIC member stressed the importance of evidence-centered design.
- One SAIC member hoped that the guidelines would provide a level of detail similar to that of the item specifications developed by Smarter Balanced.
- One SAIC member acknowledged that item coding will be a challenge—especially if coding for all three aspects of the NGSS in each "standard."

Developing Blueprints

- One SAIC member acknowledged that three-dimensional blueprints are unknown territory—making collaboration a valued approach.
- One SAIC member stressed that the timeline should allow for members of the Collaborative to gather input from essential state stakeholders.

Item Development

- One SAIC member agreed that item development that includes a feedback loop with state teams is the best approach and will produce the best results fastest.
- One SAIC member noted the importance of having item developers with specific content background (e.g., life science).

Accessibility Considerations

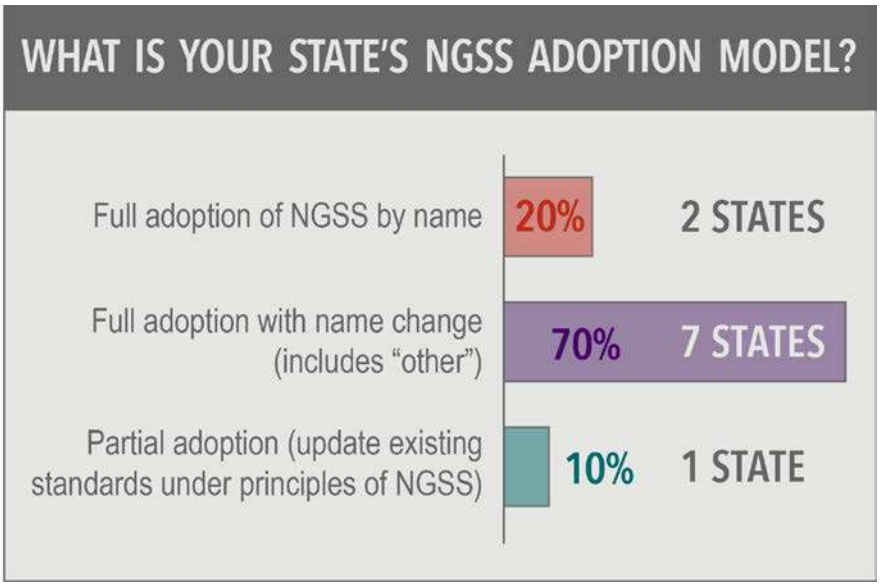
- One SAIC member stressed the importance of accessibility considerations, and acknowledged that accessibility must be a consideration from the beginning and throughout all activities embedded.
- One SAIC member urged the Collaborative to consider providing accessibility supports similar to those provided by Smarter Balanced.
- One SAIC member felt that the Collaborative should address Appendix D of the NGSS.
- One SAIC member reminded the Collaborative that students with disabilities and English language learners should always be considered when designing assessments.

Documentation

- One SAIC member was pleased with the plan for documentation.
- One SAIC member acknowledged that documentation is an essential tool, providing stakeholders and decision-makers insight into the progress of the Collaborative.
- One SAIC member questioned whether CCSSO will house documentation/record-keeping on the project (e.g., as Achieve did for standards).

Question 2: What is your state's NGSS adoption model?

Figure B-1.



- Three SAIC members communicated that they plan to fully adopt the NGSS with a name change. These SAIC members are included in the "Full adoption with name change (includes 'other')" section of Figure 1.
- Two SAIC members communicated that they plan for full adoption of the NGSS with no name change. These responses are identified in Figure 1 as "Full adoption of NGSS by name."
- One SAIC member communicated that plan for a partial adoption of the existing standards (an update to existing standards under the principles of the NGSS). This state is identified in Figure 1 as "Partial adoption (update existing standards under principles of NGSS)."
- Four SAIC members selected "Other," and the majority of these SAIC members communicated that they plan to fully adopt the NGSS with a name change. One of the SAIC members that selected "Other" is planning to fully adopt the NGSS with modifications to the middle school

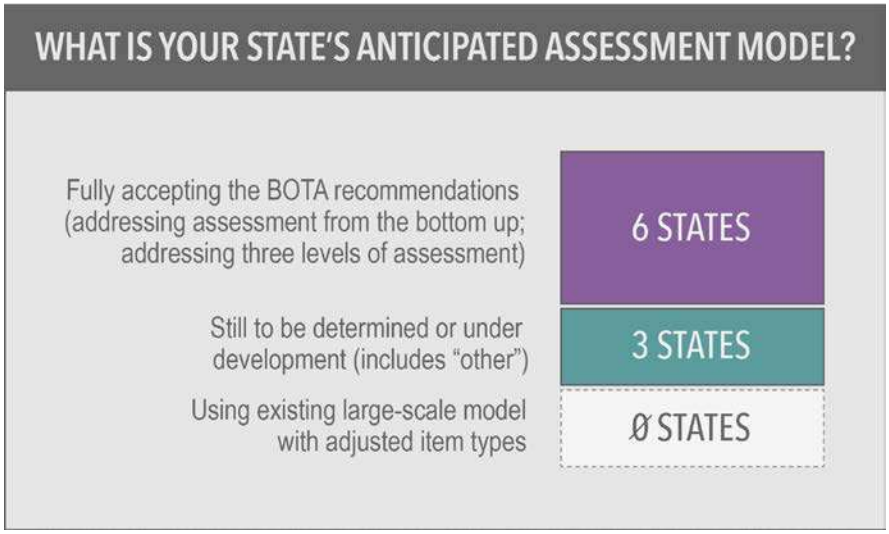
standards to reflect two models and/or pathways: the Preferred Integrated Learning Progression model and the Alternative Discipline Specific model. These SAIC members are included in the "Full adoption with name change (includes 'other')" section of Figure 1.

- One SAIC member is interested in developing a science alternate assessment that is aligned to the NGSS. This SAIC member inquired as to whether the Collaborative is planning to support alternate assessments.

Question 3: What is your state's anticipated assessment model?

Figure B-2.

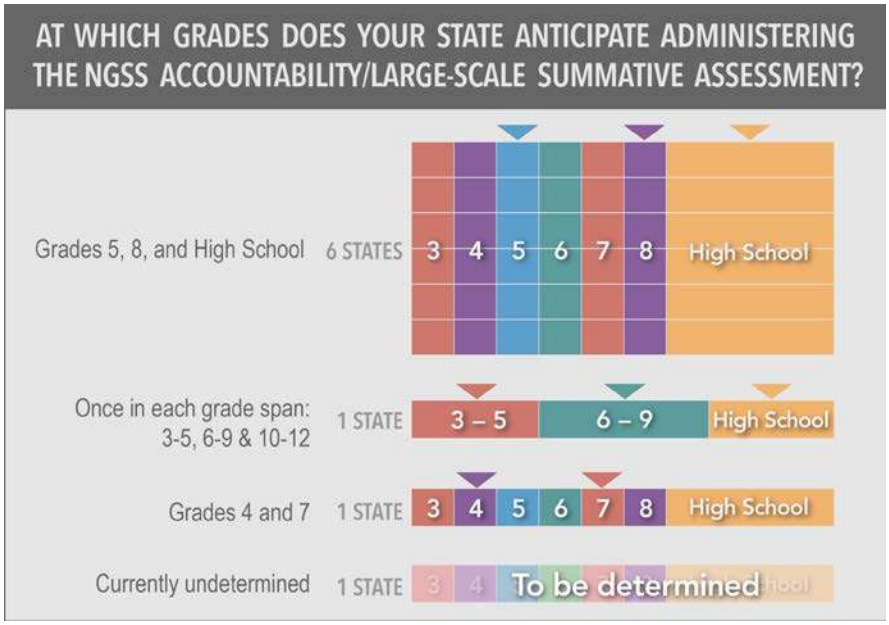
- The majority of SAIC members communicated that they plan to fully adopt the BOTA recommendations (addressing assessment from the bottom up; addressing three levels of assessment).
- Three SAIC members selected "Still to be determined or under development (includes 'Other')" for this question.



- One SAIC member that selected "Other" communicated that it was uncertain of what the model would look like. This member desires a local component, in which schools could "check off" that they completed an assessment, but scores would not be collected locally. This model would also have a SAIC member component that would have adjusted items and a performance-based component that would go in depth.

Question 4: At which grades does your state anticipate administering the NGSS accountability/large-scale summative assessment?

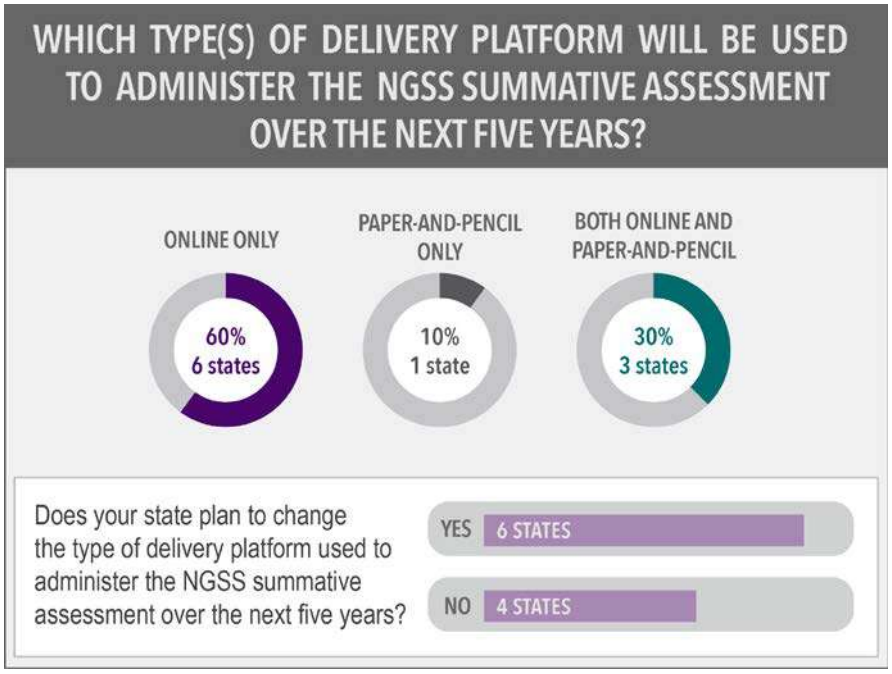
Figure B-3.



- The majority of SAIC members responded they will test in grade 5, grade 8, and high school (grade 9, 10, or 11).
- One SAIC member reported they will administer assessments in each grade span: 3 through 5, 6 through 9, and 10 through 12.
- One SAIC member communicated that they will administer a summative assessment for grades 4 and 7 and a through-course assessment for grades 3–8 and possibly high school, and it is exploring options at high school, either for a summative assessment or to continue with end-of-course; formative items are being developed by the state's teachers.

Question 5: Which type(s) of delivery platform will be used to administer the NGSS summative assessment over the next five years?

Figure B-4.



Question 6: Does your state plan to change the type of delivery platform used to administer the NGSS summative assessment over the next five years?

- Most SAIC members communicated that they are planning to transition from a paper-and-pencil assessment to a computer-based science assessment.

Question 7: How does your state envision clustering or grouping items on the summative assessment to support NGSS assessment?

Survey responses included:

- "Similar to Smarter Balanced math and ELA assessment targets."
- "Variety of different approaches—one stimuli (with multiple items)—similar to Smarter Balanced."
- "Scenario-based clusters similar to NAEP items would be ideal. Provide phenomena/situation/data/etc., and ask standards-based questions associated with the provided information, whether it be language based, data based, or both. NAEP-type items are beautifully scaffolded, and HOTs items would be welcomed as well."
- "Having scenario-based items with different types of items. Ideally, it would change from year to year (if psychometrics allowed for this). We have lots of psychometric questions around these. We also plan to have a state performance-based assessment."
- "By horizontally and vertically aligned standards and PEs."

- Two SAIC members replied that this is to be determined. One would like the Collaborative to consider how the performance expectations may be clustered, and to use this consideration to inform how items may also be clustered or grouped.
- One SAIC member will be working with a vendor to create an assessment blueprint and address the issue of clustering/grouping items.

Question 8: Does your state feel that any of the following item types should NOT be considered for an NGSS summative item bank (Multiple-Choice, single correct response (4 answer options), Multiple-Choice, multiple correct response (multiple answer options), Short Text (multi-line text box, no formatting), Drag-and-Drop (select object by clicking and dragging to appropriate area), Hot Spot (select targeted areas in response area), Table Fill in (enter text into table cells, or drag and drop into table), Graphing (plot points on a grid), Slider (a bar is dragged to increase or decrease a value on the graphs), Equation/Numeric (select buttons representing numbers and symbols to create numeric response or equation), Two-Part Multiple-Choice (part 1 selects option, part 2 selects evidence to support response to part 1), Hot Text (select and move/order text))?

- One SAIC member questioned the rigor available in multiple-choice items, and stated that MC items MUST be of high-quality, non-rote, and limited in number.
- It is imperative that the use of technology items be grade appropriate. Some of the technology item types may not be appropriate, depending on the items themselves and the rigor to which they are written. For example, the two-part MC, if used for argumentation/explanation, leaves out a key component in science—specifically the reasoning. This is an important science practice that should not be ignored. Therefore, two-part MCs would be incomplete.
- One SAIC member did not think that MCs with a single correct response (four answer options) align with instruction.
- There is some interest in adapting technology-enhanced items to be paper-and-pencil.
- One SAIC member did not think that having 150 multiple-choice items (attempting to measure every PE and DCI, CCC and SEP) is feasible to produce and/or administer.

Question 9: Please provide any known sources of present, publicly available item sources for model items (even if these could be critiqued for alignment to NGSS).

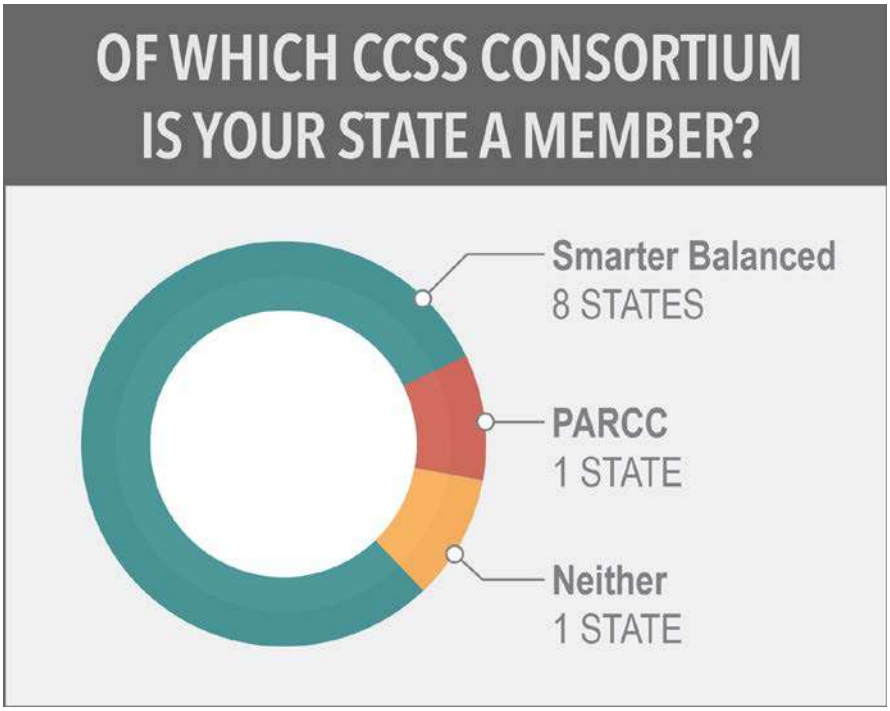
- Survey responses included:
- "TIMSS."
- "BOTA report (NRC)."
- "NAEP 2009 Science ICT."
- "Released NAEP and PISA items are the closest to types of items that the Collaborative should pursue."
- "WestEd has scenario-based items worth reviewing."
- "California released multiple-choice test questions aligned to 1998 standards. These standards are publicly available. These items may be revised to align to the NGSS."
- "Massachusetts has biology test modules that can be revised to align to the NGSS. A draft of the item specs can also be shared."
- "Washington's current sample and released items are posted on Washington's Science Assessment webpage: Example of Released items; Examples of Sample items (see Science Assessment Update)."

Question 10: What are the main challenges that you see for the Collaborative, Assessment Framework, and/or item pool to overcome?

- The timeline; time limitations to administer summative assessments.
- Developing items that assess the three-dimensional NGSS performance expectations.
- The development of items for alternate assessment, which some SAIC members may need to implement.
- Aligning items to specific claims or targets (reporting categories).
- Determining the grade levels that items will be developed for.
- Use of technology-enhanced items (e.g., technology requirements, bandwidth).
- Costs associated with science assessment items (e.g., development, administration, scoring).
- Creating items that are useful to all SAIC members, and deciding on the use of items among the SAIC members, including maintaining security.
- Political environment: public sentiment around assessments; political disputes (e.g., issues with global warming and evolution).
- Developing a really good assessment without creating an unnecessary burden on student workload (balance with CCSS).
- Protocols for development (including content and bias reviews), state-managed item development opportunities, and consistency of metadata associated with items.
- Management of item pool (development of item sharing agreement) and agreements for acceptable use.

Question 11: Of which CCSS consortium is your state a member?

Figure B-5.



APPENDIX C: SAIC State Survey #2 Summary Report

In order to fully understand the needs of each member of the Collaborative, WestEd developed and distributed a second survey to the individual states that comprise the Science Assessment Item Collaborative (SAIC). A total of 11 states and one U.S. territory responded to SAIC Survey #2. WestEd received responses from assessment personnel from the members of the Collaborative. This report aggregates these responses and presents the results and findings from the second SAIC survey.

Question 3: What should be the Collaborative's commitments for the NGSS-aligned Assessment Framework?

In response to this survey question, SAIC members responded by suggesting that the Assessment Framework should offer enough flexibility for customization while still providing focus, purpose, and clear item specification guidelines for assessment of all of the three dimensions of the NGSS. Members shared that they would be looking to the Assessment Framework for guidance on whether the three dimensions would be most successfully assessed in single PEs or in bundles of PEs. These members emphasized that the Assessment Framework should address issues regarding test items for states that have integrated standards pathways and discipline-based pathways, and that differentiated item development for alternate assessments should also receive a place in the Assessment Framework.

The SAIC members stressed the need for thoughtful consideration of the grade levels, grade bands, and/or subject-specific courses that would be the focus of NGSS-aligned assessments. Several members voiced the opinion that the Assessment Framework should offer accessibility and accommodations guidelines and a comprehensive and valid measurement model. In addition, several members expressed interest in an Assessment Framework that addressed the recommendations in the BOTA report. Finally, members suggested that the Assessment Framework should be compatible with existing consortia systems (for formative, interim, and summative assessments) for ease of implementation where applicable.

Question 4: What should be the Collaborative's commitments in terms of NGSS-aligned items?

Members responded that they seek clarity on the links between item types and valid, performance-based assessment of the three NGSS dimensions. Members would like to see item prototypes for all grade levels and more information about how the dimensions can effectively be assessed by items and clusters. They also requested specifications that "detail the development of a variety of item types, including computer-based and paper-pencil items, performance tasks, animations, simulations, accessibility supports, alternate assessment items, etc." One member requested "examples of task stimuli that can be linked to a series of items to measure each aspect of an NGSS PE or bundle of PEs" and examples of cognitively challenging or complex items and tasks that would not require hand scoring. Some members emphasized that the item specification should exclude one-dimensional items, focusing instead on three-dimensional items and bundles consisting only of two- and three-dimensional items.

Question 5: In addition to the K–12 Framework, NGSS, and Evidence Statements, what other key documents need to be referenced in the Content section of the Assessment Framework?

There were many varying ideas/suggestions around this question. Responses included:
- BOTA Report
- Stanford group's work on performance tasks
- Documents by named authors: Joe Krajcik; Mark Wilson; Derek Briggs
- Psychometric research on multidimensional items
- Smarter Balanced, PARCC, College Board, NAEP, TIMSS, and PISA item frameworks
- Work that the NECAP states have done around performance-based tasks and the NGSS
- UAAG manual
- WestEd SimScientists examples
- Taking Science to School
- AAAS Atlases
- Standards for educational psychology
- A rubric for performance assessment
- Common Core State Standards for Literacy in Science
- KY's assessment system plan (once developed)
- DE, CT implementation plans
- PD plans from other states on assessments connected to instruction

Question 6: To what extent should the Content section of the Assessment Framework address states' different models for implementing the NGSS?

When asked "To what extent should the Content section of the Assessment Framework address states' different models for implementing the NGSS?" some SAIC members remarked on the value of including different states' implementation models in the Assessment Framework. Others felt that state models should not be highlighted: "Development of the Framework should be guided by the final version of the NGSS as published, not by the different models of implementation." Still others favored a more moderate position in which the NGSS as published are honored but the strengths of various state models are also highlighted, with attention to the sharing of lessons learned from different approaches.

Question 7: What additional information should be collected from each state in the Collaborative to inform development of the Content section?

There were many varying ideas/suggestions around this question. Responses included:
- State budgets and implementation timelines
- Indication of grade levels/grade bands and/or high school courses for science assessment
- State-specific bias and sensitivity concerns, including issues of cultural competency and relevancy
- Resources available and plans for hand scoring and/or machine scoring (AI)
- Preference for paper/pencil tests, computer-based testing, or both
- List of assessments administered at the state level, including consortium assessments, and plans for development of future state assessments
- Willingness to participate in Phase II
- State reporting categories

Question 8: Which aspects of the Content section of the Assessment Framework warrant further discussion? What are the most important questions or considerations for the Content section of the Assessment Framework?

The most common aspect of the Content section mentioned as warranting further discussion is bundling of Performance Expectations (PEs). More broadly, one member wished for elaboration on how groups of items can be bundled with a similar stimulus for purposes of developing performance-based assessments. Other points included: test design validity; grades at which results would best inform instruction; the time window for NGSS-aligned testing within the school year; reporting of student achievement; a system for deciding whether items are most appropriate or are inappropriate for formative, interim, or summative assessment; and concern about states adopting different levels of the NGSS and potential consequences for implementation.

Question 9: Please provide feedback as requested regarding Phase II.

Of the eight SAIC members to this question, opinions varied across states. Half of the SAIC members (*n* = 4; 50%) voiced the need to better define item-sharing agreements and, by extension, vendor agreements. SAIC members wondered who will monitor the shared items across vendors with states in the SAIC.

Some SAIC members (*n* = 2; 25%) took issue with the timing and release of Phase II. For these SAIC members, a 2016–2017 operational assessment is preferable to the 2017–2018 timeline described by Scott Norton (CCSSO). Acknowledging the difficulties in aligning the assessment window, one member noted that Phase II development could still prove useful to the development of formative tools.

One member reported having extensive experience with online computer adaptive testing (CAT), as well as with Smarter Balanced–based item development (CAT items and performance tasks). This member wondered what the final cost of Phase II will be.

Some SAIC members (*n* = 5; 62%) skipped this question.

Question 10: Which claims (or claim language) should the Assessment Framework address? (Claims should be thought of in terms of reporting categories.)

Many of the 12 SAIC members to this question (*n* = 5; 42%) indicated that, at minimum, there should be three reporting claims. Four of these SAIC members (33%) said the claims should be Science and Engineering Practices (SEPs), DCIs, and Crosscutting Concepts (CCs), while the fifth (8%) could not decide between that set of claims or Physical Science, Life Science, and Earth/Space Science and Engineering. Two SAIC members (17%) thought that suggesting possible claims may be premature without psychometrician input regarding what claims are defensible. One member (8%) indicated that SEPs should be reported across the four science domains. One member (8%) said the Assessment Framework should address claims modeled after those of NAEP and PISA, with one overall score and subscores as appropriate.

Question 11: What level(s) (discipline, core idea, sub-idea) of Disciplinary Core Ideas (DCIs) should be the focus of reporting for a summative system?

Opinions varied widely across SAIC members, and some SAIC members shared multiple opinions. Of the eleven SAIC members to this question, six SAIC members mentioned the usefulness of providing more detail in the reports for end-of-course assessments, perhaps at the subdomain and content-area levels. Three SAIC members explicitly stated the importance of providing a three-dimensional reporting structure for summative assessment reports. Others suggested that summative reports should focus on discipline and core ideas, with one member specifying that this is particularly key within a "matrix summative test." Two SAIC members suggested that a summative system focus its reports on overall composite scores and discipline scores.

Question 12: How should Crosscutting Concepts (CCs) and Science and Engineering Practices (SEPs) be included in reports?

Seven of the 11 SAIC members to this question (64%) mentioned the importance of maintaining the integrity of the three NGSS dimensions by not separating out the DCIs, CCs, or SEPs, but by reporting at the subdomain levels (or using some other three-dimensional reporting structure). One of these SAIC members (9%) suggested, "It should be assumed that that all three dimensions are included in whatever reporting categories are used. This assumption should be made explicit." Three of the 11 SAIC members (27%) indicated that CCs, SEPs, and DCIs could each be reporting claims, and could be "crossed" such that items could reported in multiple categories.

Question 13: Should reporting categories be different for different levels of the assessment system (e.g., monitoring; large-scale summative accountability assessments)?

With the exception of one "TBD" answer and three "No" answers, SAIC members answered "Yes" to this question. One "No" member wrote that "in order to monitor student progress, educators need similar data so as to plan for instruction," and another member said that reporting categories should be "[the s]ame for all test[s so that] teachers become use[d] to the system. Different for interims, summative . . . [fewer] Performance Expectations cover[ed] on each interim[;] could provide more information on interims." One member suggested that categories should not be different for alignment purposes but that the formative and interim [e.g., grade-level] assessments could allow for more subclaim levels to better inform the instruction needed at the classroom levels. One member suggested that reporting categories be more specific for formative use of the assessment and more general for accountability purposes.

"Yes" SAIC members provided the following rationales:

- "[T]he goal of a large-scale summative assessment is for monitoring vs. an assessment that is used in the classroom (interim, formative) and so more detail is needed at the classroom level (individual or bundling of PEs)."
- Large-scale summative assessments require less-detailed reporting.

Another suggestion was that reporting categories should be different, but that perhaps they could be mapped to each other, while another member suggested that the rubrics should be aligned. "While there may be MORE information at a finer scale on class[room] or monitoring assessments, other reporting information and rubric types should be similar to create consistency. This helps teachers build capacity for high-quality assessment as they move between professional development of one assessment type to another while also preparing students AND teachers for the types of expectations they are held to at all levels of science."

Question 14: Are the Evidence Statements sufficient documentation for evidence of success at each grade? If not, what additional information should be presented in the Assessment Framework?

Four of the ten SAIC members to this question (40%) indicated that the Evidence Statements provided sufficient documentation for evidence of success at each grade level, but each member provided a caveat. One said that these Evidence Statements were fine for now and for the current work, but that, later, individual states will want further delineations based on their specific levels. Another member indicated that some of the Evidence Statements were sufficient, but others were not sufficient because they were unclear and may require additional explanation. A third mentioned that the Evidence Statements would provide sufficient documentation of success if paired with learning-progression documents, while another indicated that the Achieve Evidence Statements would most likely be sufficient and have been released.

Another four SAIC members (40%) said that the Evidence Statements were not sufficient in showing evidence of grade-level success. Of these, two SAIC members (20%) mentioned that discussions of performance level descriptors (PLDs) and how the Evidence Statements are or are not different from them may be necessary. One member indicated that the Evidence Statements, as is, are too broad to develop consistent items, and another suggested that the Assessment Framework is not the right place for this kind of documentation, particularly since ALDs will be needed anyway.

Two SAIC members (20%) responded that they were not sure whether the Evidence Statements provided sufficient evidence of success at each grade.

Question 15: Levels of a measured characteristic: presently, the Evidence Statements address only a single proficient level. An assessment system (with measures that serve different purposes) generally includes multiple proficiency levels. Consensus on number of levels for the Assessment Framework is advisable. How many levels should be specified for the Assessment Framework?

All eleven SAIC members to this question (100%) indicated that there should be four levels specified for the Assessment Framework. Five SAIC members (45%) mentioned that scores are easier to read and interpret when there is consistency in the scores' meaning and in reporting formats across assessments. Of these, two SAIC members (18%) mentioned consistency with NAEP, PARCC, and/or Smarter Balanced (while noting that these assessments have differing scales).

Question 16: Which aspects of the Measurement Model section of the Assessment Framework warrant further discussion? What are the most important questions or considerations for the Measurement Model section of the Assessment Framework?

Five of the ten SAIC members to this question (50%) thought it most important to gather input from psychometricians about the measurement model being used. Three SAIC members (30%) mentioned that further discussion about how the three-dimensionality of the NGSS will be measured and reported is warranted. One member (10%) indicated that, when considering the Measurement Model section of the Assessment Framework, more clarification is needed about two

aspects: assessment design ("Are the PEs the assessment targets? Are we attempting to develop tasks or an assessment that measures every PE for each grade level or grade band?") and how the assessment will be scored ("Machine-scored, AI-scored, and hand-scoring options"). Another member (10%) wanted more discussion of the actual measurement model being used, and whether it will be an evidence-based design or a more traditional approach. A third member (10%) expressed interest in further discussion about the scalability of the model and whether the model could scale up or down to accommodate the needs of states, districts, and schools. This member also felt that further discussion about the framework's alignment to the math and ELA assessments' framework and design was needed.

Question 17: Can alignment of items be discussed/presented for items both individually and within the context of clusters of items? Or should alignment be examined only at the cluster level?

There were many varying ideas/suggestions around this question. Responses included:

- Perhaps both. Some experience with clustering (or bundling) is needed to see how alignment could be presented at this level.
- Bundles of items should be aligned in regard to CAT items. As for a performance-based task or simulation, one or multiple items might be involved.
- Both need to be discussed further, so that guidance can be given in order to develop a blueprint and items that will be able to be used by all partners in the Collaborative.
- Individually at Performance Expectation.
- Suggestion: This discussion may need to consider all options: alignment of stand-alone items, alignment of stand-alone items with clusters, and alignment of clusters.
- Potentially both—some robust performance tasks may be sufficient in generating evidence of attainment of the entire PE or combo of PEs, but many will require that there be clusters or linked sets of items in order to have sufficient evidence.
- Individually, for sure. Some experience with clustering is needed to see how alignment could be presented at this level at the cluster level.
- Will clusters alone work? What if you have a great bank of cluster items, but you are unable to put them together to build a viable assessment? How many clusters in your bank will you need to support an assessment if that is the only thing you have?
- This completely depends on whether the clusters are attached to a scenario or if they are stand alone items clustered by another characteristic.
- Cluster.
- Primary alignment can be at the individual level to make it easier during the writing process. It can have multiple ranges in multiple areas, yet have only one primary area of focus, without limiting the item and the student.

Question 18: To what elements of the NGSS should each individual item align?

The majority of SAIC members to this question said that each item should align to a minimum of two elements; however, there was a strong preference for maximizing the use of three-element items. One member indicated that a balance of two-element and three-element items can be recommended in the Assessment Framework and/or in states' test blueprints, while another suggested that once the Collaborative has attempted the writing process, we will know more about this limitation and can ensure that development is balanced, which can be demanded by the blueprint. One member felt that we should be focusing on common stimuli with a bundle of items,

rather than on the development of individual multiple-choice questions, and that it would be very difficult to develop an individual item that measures all of a PE or a bundle of PEs. Another suggestion was that items may need to align to individual PEs as well as clusters/bundles of PEs.

Question 19: In terms of development, should items be developed within the context of a cluster and then be extracted as individual items for use, or should items be discussed/described as individual items, with guidance on how to cluster?

The majority of SAIC members to this question suggested that items will have to be clustered in order to be aligned and to accurately assess the NGSS. Some SAIC members suggested that discussions may need to focus on both options or on one or the other. A couple of SAIC members believed that items should be described as individual performance expectations and not clustered.

Responses indicated that there is a lot of room for discussion around this question, as follows:

- Some SAIC members felt that if items are developed in a cluster, they cannot be extracted without destroying the context of the item, while others felt that items can be developed in a cluster and extracted for individual use.
- "Using bundled standards to provide a novel context but providing enough data/information for students to 'solve the problem/phenomenon' would be the goal of [the] NGSS. If you do this, though, you cannot 'extract' the items from one another without destroying the context of the item in most cases."
- "Develop items for either one or the other. If items are developed in a cluster, within a common context, it is very difficult to extract those items for individual or stand-alone use."
- Grouping items developed to stand alone.
- Some felt this would give states the most leverage. It allows the cluster to stay intact and yet to be open for other methods. A focus on both options would also offer flexibility to test developers in terms of items available for use.
- Develop items for either one or the other.
- "Grouping items developed to stand alone into clusters after the fact will not allow for the use of rich contexts and common stimuli for a set of items."

Question 20: Describe your state's perception of a performance-based task. To what extent are performance-embedded tasks clusters of items by nature?

Since perceptions of performance-based task vary across states, all responses to this question are described in the following bullets:

- Performance tasks require higher-level thinking skills and reasoning. Smarter Balanced performance tasks would be considered an example of performance tasks as part of a summative assessment.
- A performance-based assessment (similar to Smarter Balanced) entails a common stimulus and a set of items that measure different standards or parts of standards that together assist with meeting a PE or set of PEs.
- Performance-based tasks incorporate the students making claims, conducting investigations, doing research, and using simulations to gather information, which leads to an explanation from reasoning, in order to develop an argument regarding the task.
- Aligned by standard. However, would be problematic when navigating between DCI vs. Topic Arrangement. Use AP model. Multiple labs that must be done during the year.
- Suggestion: The types or nature of performance tasks will most likely depend on the test delivery mode: computer-based, paper-pencil, or hands-on. Performance tasks may need to be multidimensional, with several components that may be integrated (computer-based, paper-pencil, and hands-on).
- Performance tasks are multidimensional, requiring the synthesis of knowledge, skill, reasoning, and possibly product or performance generation; they may or may not have multiple components or "questions"/steps/procedures prescribed within them.
- Looking at a hybrid system with a local component and a state performance-based summative assessment that is supported by the local assessment. The two would be independent assessments—one would not rely on the other. However, administering the local assessment would benefit students when they take the state component.
- One state considers NAEP HOTs items and NAEP scenario-based items to be large-scale performance-based tasks. Beyond these performance assessments, the state also considers PISA-type items with prompts of content/data and banks of associated questions (often MC or MSCR) to work as performance-based tasks. If the item is an explanation or argumentation question, a free-response essay or short-answer writing prompt is appropriate.
- Not cluster of items by nature. A cluster of items would not grow on a concept or have a progression of difficulty. Student understanding should be shown through the completion of a performance task.
- One state currently uses scenarios in its tests, but not to the level of a performance task. It would consider the use of performance tasks based on the legislatively approved budget for the science assessment. It would like the option of variety in test item types and is open to that idea. The majority of performance tasks that the state has been exposed to appear to be clusters of items.

Question 21: Should the Collaborative address interoperability concerns in item development, or should they be external to the Assessment Framework development? Should the Collaborative seek to provide recommendations for interoperability within the Item Specifications Guidelines?

Most members responded that the SAIC should address all concerns that impact item development and should seek recommendations for addressing each. Reasons for this included:

- Keeping the discussion and dialogue transparent should keep the major concerns of interoperability to a minimum.
- Keeping in mind the great variety of options that states may need to explore is paramount in maintaining the collaborative team.
- The items may need to be compatible with Smarter Balanced and/or PARCC test delivery and item banking systems.
- If there are members of the paid Collaborative that are not online, it will be important to address their needs.
- One member felt that it should be addressed prior to moving into Phase 2, but that it was not a high priority for the initial stages of Phase 1.

Question 22: What model(s) should be the expectation of item interoperability (e.g., the CCSS-based assessment consortia models, QTI/IMS Global)?

SAIC members responded that CCSS-based assessment consortia models and QTI/IMS/APIP should be considered. One said, "It would be ideal to keep consistent with the consortia models for the ease of use by the field in both administration and reporting of results. One major concern would be the impact on the states' accountability system. Does it fit and how do comparisons across exams translate? How this is communicated to the various stakeholders in the field is going to be the 'elephant in the room' discussion that still needs to surface." Another suggested that several models should be considered and presented to the SAIC partners for consensus.

Question 23: Should item types be described based on the assumption of online test delivery? What guidance will be needed to accommodate paper-and-pencil administration and scoring? Should a particular delivery mode be required for certain subsets of item types or clusters? How should this concern translate into a Collaborative commitment?

The majority of members responded that online delivery should be assumed, but also saw the need for some guidance for paper/pencil administration and scoring (e.g., when called for by an IEP). One stated that both online and paper/pencil delivery should be able to be accommodated. Another noted that Smarter Balanced has developed guidelines for both online and paper/pencil items. A third member suggested that considerations may need to include details regarding items with embedded supports (consistent with Smarter Balanced and PARCC) and items intended to be used in alternate assessments.

Question 24: Which aspects of the Item Types and Clustering section of the Assessment Framework warrant further discussion? What are the most important questions or considerations for the Item Types and Clustering section of the Assessment Framework?

SAIC members indicated that item types and possible clustering are a major topic for discussion. They had the following suggestions and questions:

- Item type should depend on the evidence needed to demonstrate understanding of the PE.
- Without interoperability, this doesn't become an efficient and operationally beneficial process; the [item] type should be selected based on its ability to elicit evidence of student attainment of the PE, not limited just for the sake of limiting—i.e., suggesting that there can be no MC would be a mistake if a MC can generate defensible evidence, especially as part of a linked set or cluster of items.
- Explore all of the possible item types, knowing that each state will select the best fit based on the restraints presented in the state's test design and mode of delivery.
- How to bundle PEs?
- How to bundle different item types (CR, SR, PT, TEI) in order to measure the full intent/construct of a PE or bundle of PEs?
- If items are to be clustered, how will the clusters be determined?
- Which item types will best align to the NGSS and assess the NGSS?

With regard to psychometric considerations, one member cited them as the most important consideration, while another said, "I think that saying we need to address psychometrics differently is all well and good, but without understanding of how that would work, it is hard for some in the room to be brave enough to look 'outside of the box.'"

Question 25: Would states find draft test blueprints useful? Alternatively, would general guidance about blueprint development be useful?

Responses to this question were mixed. Some SAIC members said that both draft blueprints and general guidance would be helpful, but if nothing else is possible, guidance on a potential blueprint and development ideas would be acceptable. Other SAIC members said that they feel the need to get through the Framework first and prioritize guidance on item specifications. One member indicated that [general guidance and/or draft test blueprints] should be one of the last things on the to-do list for Phase 1 and that "[i]ndividual states may have many different criteria that need to be met, so [general guidance about test blueprints and/or draft test blueprints] may need to be done individually."

Question 26: What assumptions are most important for guiding blueprint development (e.g., test length, administration format, amount of hand scoring)?

Responses were as follows (in decreasing order of importance):

1. Test length
2. Test format (CAT vs. fixed form, CAT vs. scoring model(s), CAT + performance task–based assessment)
   Some SAIC members suggested: Considerations for how and what to test (one dimension, two dimensions, all three dimensions in one question): computer-based test, paper-pencil test, alternate assessments, item types, and reporting categories. Discussions may also need to focus on whether to test all PEs for each grade level in one test form or whether the PEs should be samples, and on what will determine PE sampling per test form.
3. Scoring and reporting
4. The Framework and budget (equal mention)

One member also indicated that [these decisions are] a "state-level decision."

Question 27: What does your state view as the most important expectations for item development (e.g., developers must have content expertise; items must be reviewed by teacher teams)?

SAIC members responded that developers must have content expertise and that teachers should be included as an essential part of all phases of item development, including writing, content review, range finding, and content review with data. They also indicated that item developers should have an in-depth knowledge of the NGSS, as well as development expertise with regard to accessibility and universal design, bias/sensitivity, and psychometrics and test design. Members indicated a need for innovative writers who can produce quality items that are as three-dimensional as possible and that items should have the possibility of being used in multiple settings. It was widely requested that the vendor be a collaborator with state agencies' science content specialists in the development process and that opportunities for state content review exist (using knowledgeable, representative reviewers from a variety of backgrounds) for the potential item samples for the field. One member mentioned that timelines for delivery will be a major element.

Question 28: To what accessibility considerations/standards should the Collaborative commit?

Several specific ideas and imperatives regarding accessibility and NGSS-based assessment emerged from the SAIC member survey. In response to the question "*To what accessibility considerations/standards should the Collaborative commit?*" members responded that the considerations need to be consistent with Smarter Balanced and PARCC accessibility considerations, specifically referring to the Smarter Balanced *Usability, Accessibility, and Accommodations Guidelines* (Smarter Balanced, 2015), which contains sound general and assessment-specific guidelines. Another suggestion was to look at the English Language Proficiency Assessment for the 21st Century (ELPA21) consortium's accessibility guidelines (ELPA21, 2014). One member noted, "The principles of universal design must be evident." Universal design was emphasized by multiple members, who commented that the entire assessment system, not just an individual test, should be accessible.

Question 29: What additional guiding documents are recommended for use as resources during this process?

When SAIC members were asked to respond to the question "What additional guiding documents are recommended for use as resources during this process?" the majority of members requested that the guidelines and lessons learned from Smarter Balanced and PARCC be considered and that item development documents from both consortia may be useful. Members specifically referred to Smarter Balanced stimuli and the Usability, Accessibility, and Accommodations Guidelines (Smarter Balanced, 2015). Other suggestions for use as resources included NAEP, College Board, and PISA item development documents; PE bundling examples; Evidence Statements (for high school, and for middle and elementary grades once available); the CCSS for ELA Appendices A, B, and C; the CCSS Grades 6–12 Standards for Literacy in History/Social Studies, Science, and Technical Subjects; the BOTA report; and any other research that has been done during and since the development and final release of the NGSS. One suggestion was to use documents from the SAIC states that are willing and able to share their work (e.g., Kentucky's assessment system plan, once developed; documents and implementation plans from Delaware; Connecticut's implementation plans).

APPENDIX D: SAIC Assessment Framework Survey (#3) and SAIC Assessment Framework Workbook Feedback Summary Report

SAIC Assessment Framework Survey (#3)

Combined States' Responses (all comments are verbatim)

Question 3. Please provide comments on the overall structure, level of specificity, clarity, etc., of the Assessment Framework. Both positive and negative comments are encouraged.

- It is a document that gives clarity to our work, specifically keeping in focus the recommendations of BOTA, NGSS, and A Framework for K-12 Science Education.
- I believe it outlines the steps we need to take toward development of this new assessment.
- I particularly like the reference at the end for a task based test to cover more than one discipline.
- The provision of Key Questions and Takeaways at the beginning and end of each chapter, respectively, helps the reader to focus and refocus on the key details addressed in the chapter.
- Please provide a list of appendices with the Table of Contents.
- The Assessment Framework seems to serve two, related purposes – one is to help guide state development of their large-scale summative assessments and the other is to guide the work of the collaborative in developing a "pool of high-quality summative assessment items (pg. 4)". While there is considerable overlap in these purposes, they are not the same and it causes confusion at various points. The document should be clearer about the dual purpose up front and throughout the document.
- The framework states that it is supporting the development of summative assessments (pg. 4) but "summative assessments" is not defined. Are we talking about large-scale statewide assessments that satisfy ESEA and/or local classroom summative assessments given at the end of a unit and/or year?
- The terms items and tasks (pg. 5) and items clusters and performance tasks are used (and even defined on pg 23) but it isn't clear if these are intended as separate types of assessments or meant to be interchangeable terms.
- I thought the general outline and structure of the framework was okay. I suppose that assuming that the readers will not have read the referred documents is fair, though it was difficult to read so much "redundant" information as someone who HAS read them.
- There must be a decision regarding the audience for the document, and clarity regarding who is "speaking" and who the "listener" is. For example sometimes things are stated so it feels like the document is putting forth the ideas/consensus of the SAIC and sometimes it feels like the document is addressed to the SAIC and identifying things/questions the SAIC needs to address. For example "SAIC members will benefit from agreement on a common understanding of the terminology used to describe this emerging assessment." Shouldn't the document reflect our common understanding?
- It is suggested those questions be addressed and the document reflects a message from SAIC/CCSSO/WestEd to a larger audience (i.e., stakeholders etc. we will work with).

- We like the way Chapter 8 paragraph 2 on page 46 begins: “The SAIC envisions an assessment design that promotes accessibility using different item and text formats, technologies, designs, and accommodations in order to be as inclusive as possible.”
- We like the idea of the Key Questions and Key Takeaways throughout but Take Aways should answer the questions and the content in between needs to lead from one to the other.
- We believe a major purpose of the Framework is to lay the foundation for some type of shared/collaborative item development as Phase 2 of this effort. For this to be accomplished we believe Chapter 5 need to include the Item Specification Guidelines and the document to be developed outside of the Assessment Framework would be the Item Specifications.
- From Chapter 3, 2nd paragraph on page 16. Is says “The specific outputs of these targeted activities include science item specifications documentation, sample items, style guidelines, stimulus specifications, technology-enhanced item (TEI) specifications, functional HTML prototypes, performance task specifications, bias and sensitivity guidelines, and accessibility and accommodations guidelines.” To what extent will these be incorporated in the product(s) of our current effort?? This is important given that the “possible next step of this effort” would be some type of collaborative item development initiative
- After the preface and perhaps the introduction, Rather than saying the Assessment Framework “will” it should say It is recommended” or “In order to…, you need to …” For example the first paragraph under Achievement Level Descriptors on page 36.
- Overall I think the structure of the document is very well laid out and calls attention to several key issues that must be considered in developing a large scale NGSA.
- There are several cases where language throughout this document indicates that the large scale test and/or the NGSS themselves should be connected to curriculum and instruction, which doesn’t take into consideration the constraints LS tests in local control states are under. I’m not sure how to work wording here, but it would not be well received for a local control state that has relied on messaging that the standards are not curriculum to use a document like this that state the standards DO dictate curriculum and the test itself should be the measure of curriculum implementation. Specific instances are called out in the marked up document attached.
- Also of concern is the inadvertent negative portrayal of the PARCC consortium in Chapter 8 due to lack of information about their discussions and decisions around the Access and Accountability Features Manual. We would like to see this removed.
- This is a really good start to addressing the complex needs of developing an NGSA with the wide variety of constraints we all face. Thank you very much for starting this extremely difficult task.
- Overall, the framework provides clarity and is easy to follow. The use of supporting information from the BOTA report demonstrates the desire to utilize recommendations in the spirit of that report. Some of the key takeaways appear to be related to information obtained from previous state surveys, but there is no reference (in description or an Appendix identifier) provided.
- I would like to see the expected responses to the provided key questions. Are the intended answers provided in the text to each of these questions? Should the reader be directed to other components of the document for the needed information to answer?
- I like the structure of “Key Questions” and “Key Takeaways” – makes using document more efficient.
- Don’t document with dates—this Framework will be used for years to come and it will look dated.

- Spell out acronyms (at least at first in each chapter)
- Use complete names of documents (at least at first)
- Blue on blue was a little difficult to read.
- The Framework has an abundance of references to Smarter Balance and therefore a slant on many of the topics highlight projects that WestEd (NAEP) has done in the past. Very little mention of or reference to PARCC or PISA or TIMMS in the framework.
- The structure is straight forward and the level of specificity is adequate, however many of the terms may need to appear in a glossary and need to be used consistently e.g. bundle vs. cluster.
- There were so many acronyms and terms used that it made it somewhat hard to read through the entire document if you are not familiar with both NGSS/Framework and assessment language. (i.e. front matter, unpacking)
- Otherwise, we thought it was a good structure and quite clear on the intent.
- We feel that the framework gives us a general definition of terms, yet not so specific in details that it restrains further developments in the assessment process. We liked the sound research base, which is also on current understandings and the chapter breakdowns are at the correct level.
- In Chapter 2 (p.9), This chapter also seeks to reinforce the importance of considering the degree to which each states' adopted standards match the NGSS and influence the content specification process. (If the purpose of the SAIC is to develop guidance documents outlining a systematic, methodical, and research-based approach to the design and development of NGSS-aligned summative assessments; then this statement goes again the purpose and goal of this assessment framework.
- This statement could be taken out of context as states adopt the NGSS. While states have the ability to adapt or adjust the NGSS to suit the needs of their individual state, according to Achieve, you can't claim you've adopted NGSS (or be a member of the NGSS Network) if you have adapted or adjusted the performance expectations (i.e. language wise and content wise) and or the dimensions of the NGSS (i.e. eliminate all or part of a dimension).
- **Structure**
    - Preface might be re-named overview, Ch. 1 might be re-named background
    - "Key takeaway" should be re-named "Key Points" or "Summary of Chapter 1".
    - Also, the Key Questions should match the key summary or key points in a 1-to-1 manner if possible.
    - One possible idea, is to provide a more specific title to the "introduction" section at the beginning of each chapter. I felt more distracted by these sections and it might simply be due to the title and not the actual information there.
- **Level of Specificity**
    - Grade levels vs. grade bands vs. grade span, needs to be re-addressed and implications described to the users for consideration. If the practical assumption is that we will all have grade band assessments, then can we further define formative, interim and summative within this type of system in this document.
    - Please include a description in this document of how we're defining scientific phenomena.
    - The words "depth", "breadth", "complexity", and "sophistication" may need to be defined or aligned to the terminology of the ngss or the K12 framework in some way. I'm wondering because I was trying to determine if I was using these terms

correctly, and it seems that depth has a specific meaning in relation to CCSS aligned to a series of strategies. Are those example strategies needed here to help define these terms or perhaps this has already been done in one of the previous documents. How do we use these terms and do we use them differently for the SEPs, DCIs, CCCs and PEs?

- **Clarity:**
    - As a teacher, having only grade span assessments contributes to our problem in science education, especially with a lack of focus at the earlier grades. I wonder if these new assessments can be used to address the crux of this problem. For example, if we begin assessments at grade 3, are the assessments really going to support an increase in science education from grades k-2? Also, if the assessments are there to support accountability then the grades that fall between the grade level/grade span assessments are not accountable to addressing the standards to the right level of complexity—if we don't assess it, then we can't know. I would like some clarity to these issues in this document, if possible, that this group will need to take back to their state for consideration and decision-making. Perhaps this information could be framed with reference to equity in science education, and how assessments can support equity for underrepresented populations.
    - The Framework is well-organized, clear, and will be helpful as we proceed with the work. There are some sections that need to include more details, described in the following questions.
    - There are some acronyms that should be defined at their first use (e.g., PE on p.6, POE p. 39, NGSA p. 39), and some terms/concepts/process that could be summarized/defined) more thoroughly (backwards design, scaffolding).
    - We appreciate the amount of attention given teacher participation and other stakeholder input/review at key points in the test map, item spec, and item development processes. We also appreciated the way conclusions from the BOTA report are woven into the document to provide support for item clusters, measurement model, etc.

Question 4. For each chapter, there are questions presented in the "Key Questions for This Chapter" text box. Does the "Key Questions for This Chapter" text box accurately reflect the key questions from the chapter? If not, what key questions do you think should be addressed? Please provide a comment for each chapter.

Consider asking future reviewers to read the chapter and provide their key takeaways. I feel that those provided in the document are often either too general or too specific. Authentic responses from reviewers would allow for greater insight as to what information was gleaned from each chapter.

Chapter 1:

- Provided a good background. Key questions were an accurate reflection of the contents of the chapter.
- **Proposed additional question:** What are the elements/components of an effective assessment framework? (p. 5).
- Simplify the Case for the Assessment Framework by deleting all additional developments except 1) advances in cognitive research…, 2) growth in innovative assessment…, and 3) increased infusion of formerly excluded groups….• Approach to validation is not necessary as a key question in the introduction. Keep current first paragraph w/o first sentence and move the rest to Chapter 3.
- The questions seem appropriate for Ch. 1, however I'd address reliability in addition to validity in this section as well. (i.e., How will the Assessment Framework address validity and reliability?)
- Perhaps consider a question like "how will the Assessment Framework serve as a bridge between standards (to guide instruction or curriculum development) and assessment of those standards?" Maybe that question is embedded in the purpose question, but I like the focusing nature of thinking of the framework as bridging prior to reading.
- 1st question: Did not like "ultimate" purpose. There were multiple purposes, maybe use CCSSO language given to states when they signed up and talk about purpose for giving states examples of structures for large-scale assessment. Focus on large-scale, say this up front. Move these purposes to the very beginning of document. "The Case for an Assessment Framework" does not make a case at all. As written, it is a list of reports/documents.
- 2nd question: This needs to be better developed one short paragraph in not enough. Stephen Pruitt should not be named alone on page 7. Very awkward.
- 3rd question: Explanation for why validity is important is well said, however the section does not directly address how this Framework will address validity. Maybe pull out exactly where in the document each validity step is addressed.
- Validation discussion seems a little out of place here in the Overview. This may be better placed in Chapter 2.
- Yes, but there is an entire section on the case for the framework—this includes all of the research findings and advances that have come out in science learning and assessment technologies/recommendations. Then, the assessment process is rather weak for it to be one of the key questions—while people will be curious about this and need to know where it is located, it could possibly help to add in a few lines about timeline, or take it out as a separate question item and instead combine with the findings/advances section (if that is to be added).

- We suggest including the following question: What K-12 science education developments preceded the creation of the Assessment Framework?
- It would be helpful to have a question to assist that focuses on the tough design decisions states will need to address to develop an NGSS aligned assessment that is aligned to both the Framework for K-12 Science Education and the NGSS (as well as follows the guidelines in the BOTA Report in regard to an NGSS aligned assessment system).
- Should the name of the chapter be "background" instead of introduction, and add some additional subtitles in this chapter such as Introduction, purpose, Developing an assessment framework, advantages, challenges. The first two pages of chapter 1 were confusing due to arrangement of info. I kept revisiting them trying to clearly differentiate the advantages or "opportunities" and challenges –the area that was noted as a key takeaway. Wondering if key questions could be "what is the background for this assessment framework", What is the purpose? What is the process? What are the advantages/challenges?
- The section on "Case for Assessment framework" might be able to flow better if it was part of the introduction/background.
- Should the section in chapter 1 on validity actually belong with chapter 3 or another location? Also, should we emphasize that teaching and learning happens just as scientists and engineers work, and that reflected in the assessment vs. assessing standards the way students learn standards? Seems more objective but maybe not in regards to validity.
- Should one key question be: What does the assessment framework provide? Or What are the key elements does an assessment framework provide (with a description of the implications or issues for each of the bulleted items listed)? For example, I'm wondering if the new standards have been designed to address realistic breadth for the purpose of achieving more depth/complexity, why do we imply that it is impossible to assess each standard. It seems we have a responsibility to do just that either in some way.
- Key Questions are appropriate.

Chapter 2:

- The last question is not specifically addressed. In the takeaways on P. 14, the last sentence is incomplete.
- **Proposed additional question**: What are content frameworks? What purpose do they serve in the process of developing an assessment? (p. 9)
- **Proposed additional question**: What are the steps in the design and development process of statewide summative assessments? (p. 13)
- The question "What are the assessment boundaries for summative assessments?" is not clear.
- One of your Key questions states, "what are the assessment boundaries for summative assessments?" Other than literally defining assessment boundaries as things in the standards that "specify the limits to large scale assessment" the explanation for assessment boundaries and why they were included in NGSS (to clarify the extent to which students should be held accountable for certain levels of content information etc.) is lacking. May be important to explain that assessment boundaries were often focused on learning progressions that delineated one level of understanding from the next and thus created a "line" or boundary between say "middle school chemistry understanding" and the "next level."
- Third key question better worded as "How do evidence statements inform assessment development?

- First sentence of 2nd paragraph of introduction needs to be reworded – more direct and less circular.
- These questions seem to accurately capture the information in the chapter. Possible question: What is meant by content framework? (see #6, CH. 2 below)
- 1st question: well addressed
- 2nd question: We don't think this section was well developed. It doesn't really address boundaries per se. It talks more about the elements or structures of assessments. Suggest rewriting or renaming.
- 3rd question: not well addressed. Needs to be built upon. Don't use dates—document will sound dated. It would be good to give some examples of evidence statements. Need to be more explicit about what is meant by a claim and then the evidence.
- The Evidence Statements aid in developing not only assessments, but also clarify what all children need to know and be able to do in order to "own" the PE.
- Should add in "what are the key challenges that need to be addressed" and remove "what are the assessment boundaries for summative assessments?" And the last question could be modified to say "how do SAIC members anticipate addressing the challenges of an assessment system?"
- These questions accurately reflect the content of the chapter.
- A question around the clarification statements should be included. Additional guidance around the assessment boundaries (i.e. if you are wanting to measure whether a student exceeds then some of the items might need to include the content of the assessment boundaries). To me, the term "content" has a somewhat outdated connotation (e.g., lists of facts), which doesn't seem appropriate for the NGSS structure. Is there a better term – from a cognitive science perspective - that is inclusive of the interwoven SEPs, CCCs and DCIs? Perhaps "scientific understanding" is better because it is student-focused (admitting to different levels of sophistication), as opposed to "content" which suggests a disciplinary, all-or-nothing lens.
- First, I feel that the first section of chapter 2 might fit better in chapter 1, in defining the assessment boundaries in NGSS. Also, I dislike the title "introduction" at the beginning of chapter 2.
- Possible questions for consideration:
    - How do grade band progressions, evidence statements, and assessment boundaries inform interim and summative assessments?
    - What are the implications of grade band progressions, evidence statements and assessment boundaries on the development on as assessment framework?
- Key Questions are appropriate.

Chapter 3:

- Again the last question about what the states will report was not clearly answered. I believe this is because the survey results varied in state's responses.
- **Proposed additional question**: What are the components of the evidence-centered design? (p. 15)
- First sentence in Introduction should say the Assessment Framework's… not WestEd''s. We are putting forward a Framework of shared beliefs.
- The third question is important but not addressed.

- The Processes in Table 5 should each be a Key Question and we should work together to generate an answer/different possible scenarios to answer. This would be the "heart" of chapter 3 and the summary of those insights would be the Key Take Aways. (We believe that without the group taking steps such as this, there will be little value added as a result of this document – the Assessment Framework.
- These questions seem appropriate when considering a measurement model.
- Is there adequate information to support an answer to the last key question? page 15.
- Is the evidence-based approach the same as evidence-centered design? If so, please use consistent language. If not, please differentiate.
- 1st question: activities are addressed, but need an actual example of a claim and then the evidence that would support it; 2nd question: addressed well; 3rd question: Can this really be answered in the Framework? We don't think so. Rework or come up with new question.
- 1st 2 paragraphs on page 16—need to break down and explain the components; some of these come out of nowhere
- The reporting and methods of scoring for an NGSS-aligned summative system is not addressed in enough detail in the chapter.
- Add in a piece about where and how will the public be engaged to provide input and what resources are already available for item design or specifications? Also, we don't believe that the 3rd question was clearly answered.
- These questions accurately reflect the content of the chapter.
- An extension to the different types of data that could be reported out for different purposes should be included (i.e. state, district, school, parent, student, etc..) as well as what overarching claims can be made if a student is considered proficient on an NGSS aligned summative assessment (i.e. College and Career Readiness, AP/IB Potential, etc.)

There is little in the chapter on page 21 to support the Key Question about reporting.

Chapter 4:

- A good first look at what unpacking and clustering or bundling will do toward item development.
- I found one of the opening sentences of chapter four difficult to reconcile with the content of the chapter—which may be MY reading of the sentence. The second sentence says, "Using an evidence-based approach, each item type must be explicitly linked to particular combinations of dimensions within the context of a particular PE or bundle of PEs and must fit the proposed measurement model." From this statement I readied myself for a chapter explaining how item types (MC, TEI, Extended Response, etc.) would be matched with the most appropriate bundles of practices within PEs since it made sense that the practices would be the limiting factor on most item "types." However, the guiding questions directly after this section seemed more focused on the ASSESSMENT type/level. Again, however, the final paragraph states that the TYPE of assessment will be addressed in a later chapter and that this chapter is focused on item types and components.
- Upon reading the chapter, it appears that you are referring to item formats, architecture and organization and what they can do toward the goal of meeting the assessment expectations outlined by the NGSS. This includes the issues of content and context bundling, thus developing the need for complex item outlines (item bundling).
- Item types we have—what they can do—how we can mix them to do more….
- That's what I see you doing in this chapter.

- Overall a very “rich” chapter.
- Question 1 what types of items might be effectively included in… instead of should.
- A question should be added regarding the common language the group has agreed to use as a result of consulting together as a collaborative in the development of this framework. Top of page 23: Replace “SAIC will benefit from”... make it “SAIC has come to agreement on the common terminology which follows….”
- “ditto” for the challenges mentioned at the bottom of page 23. Key questions added a beginning of the chapter and the answers here!…
- These questions seem appropriate, but I think this chapter also answers the question “How do we identify item types that will elicit the types of evidence we need to make claims about student attainment of NGSS?”
- Why is item clustering essential to measuring the intent of the NGSS? (or similar question making an argument for the use of item clustering). Consider adding another key question related to the architecture of the item clusters.
- 1st: Yes, clear, but we do not necessarily agree with this approach. (see takeaway comments);
- 2nd: Yes very clear, good intention.
- The Sample Items will be helpful to determine how items address the 3-Ds of NGSS. Several of the “Eligible Item Types” are unacceptable e.g., MT (True/False or Yes/No); IC (Sentence Completion).
- “extant Science Items to Build”; Who will develop, review, and field test items?
- We feel that something should be added about the availability of already developed resources and/or allowing for the advancement of technology.
- We suggest including the following question: How will the Cross Cutting Concepts, Disciplinary Core Ideas and Science and Engineering Practices be incorporated in the item’s design?
- What types of items should be included is stated, but additional questions about what the items should be able to measure needs to be included. The different item types were modeled after Smarter Balanced, which currently excludes simulations or interactive computer tasks (NAEP), hands-on tasks (NAEP).
- Key Questions are appropriate.

Chapter 5:

- The guidelines will give vendors the specifications of what we desire the test items to look like. The key questions are answered.
- I really, really, really like the architecture prototype. The visual is very helpful.
- Key Question should be “What are the Item Specification Guidelines put fourth by the SAIC?”
- The Take Aways from Chapters 3 and 4 should become the key questions of Chapter 5. The answers developed in the Item Specifications Guideline in Chapter 5 result in the key Take Aways. The assessment Framework must provide Specification Guidelines. The accompanying document needs to be the ITEM Specifications. Once again, otherwise little value is added through this Framework and the collaborative will not be positioned to move into Phase 2 of its efforts.
- Add a key question: “What will it look like when the sample prototype is applied to an actual NGSA context?”
- The question seems appropriate I think. I feel like it’s confusing somehow, but I can’t pinpoint how. See # 5 and #6, Chapter 5 below.

- Needs more work. So far, so good.
- These questions accurately reflect the content of the chapter.
- In the item specifications document, there needs to be additional information and questions around how to develop a range of different items in regard to DOK. The PEs are listed as what a student should be able to do at the end of learning (summative), and the evidence statements provide what a student must do/show in order to be considered proficient for a given PE—additional guidance around how to develop 2-D or 3-D items to meet a varying range of DOK needs to be included.
- Key Questions are appropriate.

Chapter 6:

- The guidelines are in the form of questions for states to individually address in preparation for blueprint formation.
- This chapter seems good right now but may need additional questions and content as the remainder of the document is developed.
- The question is fine. Is it appropriate to include a question about ALD's as this is a second focus of the chapter.
- What processes should states consider in ALD development? What is the relationship between test blueprints and ALDs?
- Are guiding questions and guiding principles the same? If not, principles need to be better defined.
- Needs more specific questions related to the Guiding Questions to developing a blueprint, e.g. Reporting Categories – PE/PEs?
- These questions accurately reflect the content of the chapter.
- Additional guidance around how many item clusters should be present on the summative assessment would be extremely helpful. Additional guidance around whether all of the NGSS PEs should be measured on a summative assessment or a subset of PEs will be measured on formative and interim level (similar to what SBAC did). It might be useful to have a question focusing on the how the assessment framework will help inform the achievement level descriptors.
- Add a Key Question related to ALDs.

Chapter 7:

- Chapter 7 clearly defines the process of deciding what to test based on what knowledge and skills the student should have if he/she is in fact able to do the performance expectation, and then using that to tie the information with other PE's and determine the type of items that would go with the task.
- This chapter is constantly referring to test blueprint design when the overall discussion is about TEST DESIGN. This chapter discussion does not actually get into blueprint development, this is all about test design and specifications.
- Current questions are good. A third question could be added: "What might an item cluster look like when developed?" Give an example of a "developed item cluster.
- The questions seem appropriate and inclusive of the content in this chapter.

- A question that relates as to “how to unpack”.
- Second bullet—Why only address the first step in item cluster development? There is more to take away than step one.
- 1st: yes question appropriate and well-explained in chapter
- 2nd: should this be incorporated in chapter 4? Seems to overlap.
- **Development of Item Clusters**: no mention of AI scoring of constructed response items;
- **Criteria for Stimuli Development**: No mention of Hands-on investigation; preference would be authored Technical Passages vs “created” passages;
- **Item Reviews: Considerations**: must refer directly to alignment to NGSS;
- **Piloting and Feedback Loop**: Who will pilot and from whom will feedback be solicited?
- These questions accurately reflect the content of the chapter.
- Additional questions around how to bundle PEs and or examples of bundling should be included. What role do the model content framework that Achieve is developing have in regard to determining how to bundle PEs? A question that specifically addresses how the evidence statements are going to be used.
- Key Questions are appropriate.

Chapter 8:

- The chapter does deal with the issue of accessibility and the specifics that have to be at the forefront in development.
- Generally good to go. Switch order of questions.
- The questions seem appropriate, however a topic not covered here that needs to at least be raised in this chapter is the design of assessment forms and items for those students with the most significant cognitive disabilities. More below in question # 6 .This is a tough topic, I understand, however we do need to acknowledge it as states are required to deliver an alternate assessment as well. Will these items be appropriate for use on an alternate assessment as well? Whether the answer is yes or no, we need to acknowledge it…even if just to say this will be another individual state decision.
- Looks good.
- **Accessibility Considerations**: Have the recommendations from those who have participated in the development of PARCC been consulted?
- These questions accurately reflect the content of the chapter.
- Additional questions around how to develop assessments that meet the needs of SWDs and ELLs should be addressed. How do you develop comparable item types (as well as a fair and valid assessment) for blind students, when sighted students are given TEI, and simulation items?
- Key Questions are appropriate.

Question 5. For each chapter, there are bullets presented in the "Key Takeaways from This Chapter" text box. Does the "Key Takeaways from This Chapter" text box accurately reflect the key takeaways from the chapter? If not, what key takeaways do you think should be addressed? Please provide a comment for each chapter.

Chapter 1:

- great overview – good takeaways
- Should include the statement from page 4 that "the ultimate purpose of the SAIC is to support states in the development of a pool of high-quality summative assessment items."
- May not need validation take away.
- The current key takeaways are good, however I think this would be a good spot for identifying the purpose and/or constraints of a large-scale test. These ideas need to be expanded on in the text of Ch. 1 as well. The key takeaways here should identify the purpose of a large-scale test as well as appropriate uses of that information to avoid misunderstandings and limit the possibility of, say, a district picking up this document and using it to say that teachers will be evaluated on the state test since it aligns to curriculum even though we have no possible way of knowing what that curriculum is.
- Question about prioritizing standards (page 5). Consider unintended consequences of prioritizing standards.
- 1st bullet: Started to get there, especially on page 5, paragraph 3, however this needs to be more explicit. Talk directly about the challenges of assessing 3D standards: Item types that include 3D; Psychometrics that support 3D assessments; how to report out 3D results; how costs will play a role; 2nd bullet: Nice list, but how are these references used in the development of the Framework; would be nice to link where in the document these are being used. If they are not being used, they should be deleted. Side note: why are you using a reference for computer technology from 2001—this doesn't seem right. (top of pg. 7); 3rd bullet: addressed well (p. 5); 4th bullet: addressed well (pp. 7-8)
- Accurately Reflect Chapter.
- Yes
- The key takeaway "The NGSS, based on the K–12 Framework, pose a number of challenges for the development of assessments for the standard" could be a takeaway from each of these chapters. A more tangible takeaway could be the elements an effective assessment framework includes (Detailed description of the content (knowledge and skills) in a content domain; specification of the content that is eligible for assessment at each grade or grade span; information about how content will be assessed, including description of the characteristics of item types that will be used to measure the content in the domain; a blueprint for a test or item pool that meets the specified assessment objectives; and guidelines for administering and scoring the assessment.
- In the second bullet add input and feedback from the diverse stakeholders.
- key takeaway 1 refers to number of challenges for the development of the standards, and I'm wondering if it should equally describe opportunities and challenges, so the readers clearly understand the distinctions between the two areas through the bkgd info. on making the case vs. the assessment process.
- key takeway #3 refers to a "Bridge between", wondering if it should say "serves as
- guidelines for developers to create assessment items"
- Takeaways reflect the important content in the chapter.

Chapter 2:

- The variety of the state's responses should be discussed further.
- In the first statement, "content frameworks" needs to be clarified. the last statement if cut off.
- Third bullet in Take Away should be reordered….1 scope and purpose, 2 overall architecture, 3)curr and inst priorities, and 4)fiscal. Last Take Away should be focused on the possible impact of the Evidence Statements not on the lack of consensus (again – some of this is related to clarifying audience for the document—we don't think we are our own audience so something need to be said which will guide/prompt thinking of external audiences. The group may need to force itself to reach agreement on a few general statements and some questions that will need to be addressed "locally"
- Regarding Bullet #3- the state does not prioritize curriculum/instruction, nor do they intend to as this is a local decision. The state does not plan to even release guidance on MS/HS course mapping because of local control issues. I think it could be very problematic for us to use a document that states the priorities for assessment are based on the states' priorities for curriculum and instruction when those are not decided by the state.
- Final bulleted statement is incomplete. Unsure as to where the information from the last bullet came from as there appears to be discussion about this lack of sufficiency is provided in the text. Second bullet- Where does the statement in quotations originate? Should this include explaining phenomena in addition to addressing specific problems? I realize this is a quote from the Framework but the language now among Framework and NGSS writers seems to be explaining phenomena and solving problems.
- 1st bullet: ok, but on the vague side; 2nd bullet: ok; 3rd bullet: ok; 4th bullet: needs to be completed; need more info about the state survey responses in the text—what does this mean? What are the implications? There is no assessment boundary outcome.
- Although the Evidence Statements have been released, there needs to be more direction on possible use of Evidence Statements, such as developing rubrics for scoring PE or setting standards of proficiency.
- Ok, but the last takeaway was incomplete. We are also not sure if this takeaway is needed (but hard to know because we don't know what it is saying).
- The text box accurately reflects the key takeaways from the chapter.
- The key takeaway in regard to the evidence statements needs to be finished (currently is incomplete). Yes, although the term "priorities for assessment" in the 3rd bullet is somewhat ambiguous. The Gorin and Mislevy (2013) paper relies on "use cases" to illustrate and provide rationales for assessment design tradeoffs.
- In takeaway 3, I think we have to be careful about saying priorities at each grade, esp. if the assessment happens at each grade span. I might not say priorities for curriculum and instruction, as that comes from the standards.
- I don't really like any of these takeaway statements, but I'm not exactly sure how
- to fix them unless I align them to be the questions I suggested above.
- Takeaway #4 is cut-off from the bottom of the text box on p.14, and the whole
- statement is not visible.
- Consider adding a Key Takeaway related to Assessment Boundaries. The last Key Takeaway refers to the Evidence Statements but Evidence Statements are only briefly addressed in Step 4 under the heading of "Assessment Boundaries".

Chapter 3:

- Again the last question about what the states will report was not clearly answered. I believe this is because the survey results varied in state's responses.
- **Comment:** What are the **four levels** of reporting proposed by states? (Please see last bullet, Key Takeaways, p. 21).
- In the last statement, "four levels" need to be clarified.
- Key Take Aways need to include the agreements/scenarios that result from working through the processes of Evidence related to NGSS together.
- Last take away should say something like "it is recommended" or "SAIC recommends" that there should be four…
- The first bullet SHOULD be a key takeaway but is not really discussed throughout the chapter. More information should be given in this chapter about evidence-based design. See additional specifics in the attached mark-up. How will the final take-away regarding performance levels be addressed? The performance levels are critical to design and scoring of items and there will most likely have to be some kind of consensus or decision about this. I don't think the performance levels can be an individual state decision. A second question- how were the highest-priority categories of evidence decided? It might be useful to include that information in the text of chapter 3.
- Last bullet: where is this information provided in the chapter? It states 4 levels in the assessment framework. Should reader be referred to a certain part of the appendices for this information?
- Sentence above Key Takeaways is unclear. What is the purpose of this?
- 1st bullet: what are the high level components of an assessment? What is a claim-evidence pair? Need an example.
- 2nd bullet: Ok, but just a repeat of BOTA—should something be further interpreted from this?
- 3rd: Doesn't say how to use backwards design—need more info
- 4th: good job.
- 5th: Not addressed in text; needs to be added.
- The last bullet needs to be addressed in more depth in the chapter.
- The last listed takeaway doesn't seem to be in the chapter? Could backwards design be better defined in this chapter (3rd bullet)? It doesn't really appear to be described in the chapter as a whole and is confusing listed as a key takeaway.
- The text box accurately reflects the key takeaways from the chapter.
- There is insufficient content in the chapter to support the Key Takeaway described in the last bullet referring to four levels of performance and the focus of reporting. Evidence from the surveys in the appendices could be cited.

Chapter 4:

- A good first look at what unpacking and clustering or bundling will do toward item development.
- Changes like those below throughout the document will make it more "assertive" and help it to feel like shared/strong ideas are being expressed.
- Bullet 2, change "can provide" to just provide
- In Bullet 3, change "can help" to just "help"
- In Bullet 4, "may provide" to just "provide

- The key takeaways seem both appropriate and powerful. As a general note, while we are all in agreement and the research indicates online testing will be the most appropriate modality of assessment for delivering these items, we need to be mindful that a number of logistical and infrastructure issues persist in states and paper-based assessments may continue to be necessary for the next few years. Perhaps a slight rewording or disclaimer of the 4th bullet to include something like "every effort should be made to encourage online-deliver of these items, though some consideration must be given to adapt items for paper-based delivery for those states that do not yet have a technology infrastructure that allow them to move to a fully online delivery."
- 1st bullet: Yes, this is a take away. However, we had a few issues with the actual recommendations. First, we think that an individual item within a cluster should be able to address only 1 dimension, particularly if the item is only worth 1 point. Second, each item should not necessarily adversely affect another item within the cluster. It would be very helpful to have a psychometrician weigh in on this part. It doesn't make much sense to us. (p. 24). Third, stimuli should also be able to include images, diagrams, graphics, tables, and data. What is meant by experimentation, discussion, activity, and/or demonstration? Can you please spell this out for a large-scale assessment? Again, these don't seem to make much sense as written. The statement in the middle of pages 25 starting with "However, developers should propose creative thinking…" does not make much sense. If something cannot be done, then is cannot be done, and it should be acknowledge that is cannot be done! Of course, discussion around issues should always happen and the best possible solution should be found. Lastly, stand alone items should still be an option for part of a summative test to ensure that coverage of the standards is adequate from a psychometric point of view. 2nd bullet: good descriptions, define "main consortia" PARCC and SBAC? 3rd bullet: yes, this is one of the takeaways, what is meant by "task" here? 4th bullet: ok. 5th bullet: this was not addressed in a realistic way for state to use; WestEd and NAEP simulations are wonderful, but they are cost prohibitive for states and take a very long time to develop and field test
- The last bullet needs further development in the chapter. What are the negative and positive, intended and unintended consequences of interactive tasks?
- We don't really feel like the last takeaway was adequately addressed in this chapter. The 2nd bullet needs to define what the "main consortia" are.
- The last takeaway (Novel, interactive tasks bring unique challenges for schools, in terms of infrastructure and capacity, and for psychometricians, who must monitor technical quality, fairness, and consequences -- intended and unintended, positive and negative -- of testing with new item types) is a conclusion not necessarily unique to this chapter.
- For states that will offer both a paper/pencil and an online option, what are the key takeaways to determine whether the two tests are both valid and measuring the same construct?
- Key Takeaways are appropriate.

Chapter 5:

- The guidelines will give vendors the specifications of what we desire the test items to look like. The key questions are answered.
- The Key Take Aways form Chapters 3 and 4 should lead us to the Key Questions and Take Aways from the Item Specifications Guidelines Developed in this chapter possibly including information related to: science item specifications documentation, sample items, style

guidelines, stimulus specifications, technology-enhanced item (TEI) specifications, functional HTML prototypes, performance task specifications, bias and sensitivity guidelines, and accessibility and accommodations guidelines. • Revise take aways into statements about what the guideline "say" not "will describe/provide. • "In real life, applying the sample prototype to NGSA results in…."

- Bullet 1: The item spec guidelines will guide states and vendors in the development of items to be used in the development of an assessment…not the development of an assessment. Just a minor clarification, but I've found it's better to be as explicit as possible when talking about large-scale assessment. It's too easy for somebody that doesn't know the process and protocol for LS development to pick up a document such as this and misinterpret it.
- Needs more work. So far so good.
- Will each item provide scaffolding or each set of items provide the scaffolding?
- The text box accurately reflects the key takeaways from the chapter.
- Range of different items that will need to be developed to measure where a student is in regard to a learning progression need to addressed. Most likely will be covered in the item specifications document.
- Can you clarify in writing in this chapter why within a cluster each item must measure 2 dimensions?
- Key Takeaways are appropriate.

Chapter 6:

- The guidelines are in the form of questions for states to individually address in preparation for blueprint formation.
- First statement - Not clear what the blueprints are for – collaborative pilot testing? State assessments? ("Blueprints … will describe the overall design for the tests …" which tests?)
- A good set of Key Take Aways
- Remove "wills" in the first and third bullets and "should be" with "are" in the 4th bullet.
- Clarify bullet 1: "state-developed test blueprints will prescribe…" Bullet 2- remove language about large scale monitoring curriculum and instruction.
- 1st bullet: yes, it does address this; however it would be very helpful to have a couple of sample NGSS-aligned blueprints, reporting categories; NAEP tables were not very helpful because it is not NGSS. Also "time" in NAEP was confusing. What is meant by this? 2nd bullet: info from the BOTA report, but there should be more here. There not much to take away and there should be. Hoping for more guidance and concrete examples. 3rd bullet: good, ALDs are well-defined
- ALD will be a challenge!
- The text box accurately reflects the key takeaways from the chapter.
- Can you define or clarify the "bottom up approach" as described in the BOTA report in this section?
- Key Takeaways are appropriate.

Chapter 7:

- Chapter 7 clearly defines the process of deciding what to test based on what knowledge and skills the student should have if he/she is in fact able to do the performance expectation, and

then using that to tie the information with other PE's and determine the type of items that would go with the task.
- Key Takeaways don't include all of the key points from this chapter (.e.g., review process, pilot testing)
- On page 45, the third "takeaway" states that PE's can/should be bundled as a teacher would in the process of investigating and explaining an overarching science phenomenon. I think a better way of putting that (in support of NGSS educational practice) would be to say something like: *"The BOTA report recommends bundling PEs to support rich item clusters, possibly grouping them as a teacher would in the process of investigating overarching science phenomena* ***TOWARD THE GOAL OF UNDERSTANDING CORE SCIENTIFIC CONCEPTS."***
- Add at least one more Take Away about the Item Cluster Development process (may need to add content to match it)
- Add a "When completed an Item Cluster "is" or "reflects" …
- These takeaways seem appropriate, but address the "curriculum and instruction" language per earlier comments about local control and current large scale testing constraints.
- Related to the last bullet about Stimuli – page 42 where criteria is described: "creative" and "appealing to students" seems vague, open to interpretation, possibly resulting in stimuli that don't meet desired goals.
- 1st bullet: is this "bottom-up" a recommendation? Don't states need to figure out the feasibility of this approach and then make decisions? 2nd bullet: ok. 3rd bullet: ok, BOTA reports talks about other item types as well. Shouldn't these be addressed? Why only cluster items? 4th bullet: yes, would be nice to have more info around stimuli. How long should they be? One issue to be aware of is putting too much info in a stimulus where not all the info is needed. Also, flipping back and forth between information given in the stimuli and the actual questions can be an issue for some students (accessibility issue). These issues should be addressed.
- The last bullet is vital to a valid assessment of NGSS.
- We feel that the 1st, 3rd, and 4th takeaway are very much related and almost redundant here. At least the 1st and 3rd could be combined as the 4th does add in the additional part about the stimulus, but the overall message from all 3 of these is that the assessment should support what instruction looks like in the classroom.
- We suggest a statement regarding the importance of unpacking performance expectations in order to matchup information, 3-dimensional progressions and assessment boundaries.
- Key Takeaways are appropriate.

Chapter 8:
- Should we address what accommodations or features will be allowable as an option for all students?
- Page 48: Applying Principles of Universal Design - Standards for Educational and Psychological Testing from AERA (2014) Part 1, Chapter 13 is more current than reference provided.
- More current to use ELs instead of ELLs
- How will language issues for ELs be addressed?
- Add a take away (may also need to add content) regarding "universal tools, designated supports, and accommodation as tools to open the assessment to broader "audiences".

- The key takeaways for this chapter also seem appropriate, however I would add in something about the use of universal design principles from the beginning of to ensure fairness for the broadest range of students possible.
- good. Liked inclusion of gifted and talented students.
- The last two bullets may warrant further discussion, e.g. below grade level reading level, use of scientific language (Disciplinary Literacy), etc.
- The text box accurately reflects the key takeaways from the chapter.
- More emphasis that universal design and accessibility need to be thought about at the beginning of the assessment and item design process (not be an afterthought).
- Key Takeaways are appropriate.

Question 6. For each chapter, is there information that should be included in the chapter that is not included? Please provide a comment for each chapter.

Chapter 1:

- **Comment:** The description of the process used to develop this assessment framework may need to be expanded to include an overview of the type of information gathered from states and stakeholders. Which states were involved? If more information regarding this topic is provided elsewhere, please provide a reference (p. 7).
- Page 8 discusses processes for "gathering input and feedback from diverse audiences at each stage of work, collecting evidence to verify that appropriate steps were undertaken to ensure that technical quality, fairness, and feasibility of the measures, documenting tradeoffs…" Is this going to be undertaken by the collaborative or each state or both?
- Should include the "value added" by this document and the impact of working together as a collaborative to develop it.
- There are some statements in this chapter that contradict the purpose of a large-scale test and do not acknowledge the constraints of LS tests accurately. I've added additional notes to the pdf, however there is 1 statement in particularly that is worrisome if additional clarity is not given (regarding priority content).
- Additionally, the next paragraph (Page 5, "Alignment of the assessment….)" contains inaccurate information and several of the misconceptions about "assessment" vs. "large-scale testing" that have led to the idea of teaching to a test or only teaching science in grades/subjects that are tested. Essentially- standards are not curriculum and are not intended to dictate curriculum, so a large-scale test cannot be aligned to a curriculum. Rather, the test is aligned to the standards. This is an important distinction that needs to be to clarified here since some states have relied on or are relying on the messaging that standards are not and do not dictate curriculum. Saying "the test" is aligned to curriculum directly contradicts this.
- Also, a large-scale test for a local-control state (nearly all of us) should not be used to assess the way a teacher teaches because we have no way of isolating specific teacher input on student attainment, particularly when the standards rely so heavily on learning progressions. It is unfair to use a LS test delivered at the end of a grade-band as a measure of teacher effectiveness. Granted, that won't stop some areas from trying, but indicating that the test IS a measure of teacher instruction is incorrect and undermines best assessment practices.

- Rather, a large-scale test is intended to measure the extent to which an implemented program aligns to the state's adopted standards for a given subject area and should make every attempt to do so in a curriculum neutral way.
- As written, Chapter one does not identify the purpose of a large-scale test or appropriate uses of that information. This is essential to avoid misunderstandings and limit the possibility of another entity picking up this document and misinterpreting the intent of a large scale test and the appropriate uses of those data. Jim Pellegrino would likely have a variety of resources to cite regarding this.
- Did not have a clear message for how the new standards will change current assessments. Need to specifically address needed changes and challenges for a 3d model.
- Last part of validation section should not focus on months, framework should project further.
- Although one of the Key Questions is about the processes followed to develop the Framework, there is very little detail in the section with the heading "Assessment Framework Development Process." Reference the appendices as some of the detail is included there.

Chapter 2:

- **Comment**: When will evidence statements for elementary and middle school be released? (p. 13). At what point will the discussion of these evidence statements be incorporated into this framework? Or, will they be outside the scope of this frameworks?
- **Comment**: Key Takeaways, last bullet, p. 14, states that "state survey responses indicate a lack of consensus" on the use of evidence statements as evidence of success. What support should this framework provide to states to facilitate the discussion around the relevancy of using evidence statements as evidence for success?
- Need to clarify what is meant by "content frameworks" on page 9. Clarify whether or not the collaborative will be developing claims (Step 3 in developing assessments (on page 13). Page 13 - Also clarify what is meant by "large-scale assessments" in 2nd paragraph.
- Although largely a rehash of NGSS, this seems fine.
- Just a thought to open for discussion- we've encountered some misconceptions about what is meant by "content" when referring to NGSS. Specifically, there seems to be a misconception that "content" refers solely to the DCI dimension and not the other 2 dimensions. I'd pose to the group- is this common in other states, and if so would it be worth defining content frameworks here as documents that identify the specifics of the SEPs, DCIs, and CCCs that items will be developed to align to?
- Would like to see some specific examples or a reference to where they can be found in the document for: NGSS-aligned items, NGSS-aligned tasks (need to define difference if you are using both terms here); NGSS-aligned blueprints; NGSS-aligned evidence statements
- The assessment boundaries are discussed, but the clarification statements might need to be also addressed in regard to their role in item cluster development. Additional guidance around the role the evidence statements should play in assessment development, as well as how to determine what evidence would be required to be considered at the different achievement levels (besides proficient). Grade band vs. grade level models (here in our state we have traditionally tested students in grades 5, 8 and high school—but the standards being addressed were at the 3-5 grade level, 6-8 grade level and all of the high school level standards).
- More discussion about the development/rationale for the Evidence statements is needed. The chapter does not provide a clear answer or much information related to the last Key Question

"How do SAIC members anticipate using the recently released Evidence Statements?" A reference to the appendices may be appropriate here also.

Chapter 3:

- Page 18 - "Each of these high-priority evidence categories is listed in Table 2, with a set of guiding questions that states might consider asking as they seek to collect these key pieces of evidence and examples of appropriate evidence for that category. The intent is to ensure that states have collected sufficient information to clearly communicate their measurement model or approach to educators, parents, policymakers, and other stakeholders." Will this be done together under the collaborative? Page 21 – would like more information about the following: "The Assessment Framework will ultimately include a summary of considerations, the result of a targeted review conducted by a panel of psychometricians who were convened to address the impacts on and constraints of an assessment system based on the proposed item cluster and alignment models. Additionally, this panel will provide input on the feasibility of the recommended reporting claims as supported by assessment blueprints. Characteristics of test design will be addressed, as informed by the measurement model, alignment expectations, reporting recommendations, and item cluster architecture."
- It may be worth mentioning either here or in Appendix A how the high-priority elements of evidence were selected. Additionally, some clarification or additional information regarding evidence-based design and backwards design.
- I might expand on the idea that summative assessment "content validity" is the enemy of coherence in science instruction. Alignment studies focused on breadth of coverage will tend to result in the "mile wide, inch deep" model of curriculum and instruction, in my opinion. This point relates to the earlier comment on "content" in Chapter 2.
- I'm a former science teacher and this section is very new to me. I need to further understand the implications of this chapter and what I will need to do with it, as this is not an area of expertise for me.
- A summary/description of the "backwards design" process as well as the College Board's Insight product mentioned in this chapter would be helpful. There is little in the chapter on page 21 to support the Key Question about reporting. That section should be expanded and have its own heading. Could also refer to the collaborative members' input on reporting from the surveys in Appendix C.

Chapter 4:

- Comment (pp. 23–24): Can an item cluster be aligned with more than one PE (a PE cluster)?
- How would the identification of interrelated PEs be facilitated? In other words, what criteria/procedures should/may be used to identify interrelated PEs that may be used to develop an item cluster?
- In cases where an item cluster is developed in alignment with more than one PE, how would the reporting of results reflect the testing on various PEs in an item cluster?
- Page 24 – the claim is made that "at least one item must be a constructed-response item" but there is no explanation or evidence to support this. Page 24 - Are the testing times needed for students to complete item clusters (15-45 minutes) realistic given a number of factors (i.e.,

overall time allowed for summative assessments, number of items (4-6) in clusters, student fatigue, etc.) We think this time should be reduced. Page 25 – we will need to be careful about bandwidth needed for certain stimuli (audio, video, animation, simulation) – need to be realistic about what schools can support (Smarter Balanced had to limit stimuli). Pages 25- 27 Should consider use of drawing tools in item types.

- Don't know if this comment belongs here, but on page 24 where you define each cluster/testlet as taking from 15-45 minutes, I think that may be an issue if you don't define the overall "length" maximum of a test. 15-45 minutes is a HUGE range. Perhaps the testlets themselves would need to be categorized in some way to ensure that you don't over define a test with "long" testlets vs "short" testlets etc. How will this be defined—by what psychometric or coding algorithm? Will you use field test data defining average assessment time, or "difficulty?" I don't know but this time disparity seems too large to ignore as just a minor differential between item clusters.
- Actual examples are critical in this chapter. Either one example or better yet several examples showing different approaches to this process abd the corresponding results.
- Again, just a disclaimer about how we support the move to online testing, and I think we are all moving in that direction, but ultimately we have to acknowledge that not everyone is there yet, nor will they be for at least a few years, and it may be appropriate to discuss the known issues for delivering and scoring these item types for paper delivery- if they even can be.
- Would like to see more about feasibility—perhaps this should be one of the guiding questions. Need to give more information about cluster development: what are the score point ranges to aim for? Do the item types need to be consistent from cluster to cluster? If they are different, how is this psychometrically justified? How many PEs/standards can/should be assessed within a cluster?
- Could use a list/table of definitions for the new terms (item clusters) developed in context of the NGSS aligned assessment for easy reference. Cognitive Demand could be identified with the various item types. Some of the types mentioned have such low cognitive demand that they should not be considered.
- I would suggest expanding on the scaffolding discussion to show how summative NGSS assessments could be computer adaptive (and target the student's zone of proximal development). The advantages for measurement precision weren't mentioned explicitly, but a mini-pool of machine-scored items developed for each cluster would be more likely to accurately measure student understanding than a fixed item cluster (in addition to increasing student engagement).
- I really like p. 30 on Stealth assessments. I want to explore more about the partnerships among game design, instructional design, and assessment design.
- Interdependence of items as opposed to interrelatedness of items should be carefully defined. We will need input form the technical review team on the interdependence of items. In our state's scenario sets, our guidance has been that performance on an item in a scenario set should not impact performance on other items. Smarter Balanced math tests do have some "dependent" items now. There should be a compromise position between "no dependency" among items and "all items interdependent" that is defined. A definition for scaffolding as well as examples of scaffolding in different item types/examples of scaffolding tying items to each other would be helpful. Figure 2 – it's not clear where scaffolding fits into this diagram. Is the scaffolding the "appropriate assistance"?

Chapter 5:

- Guidelines could include a selection of options rather than of questions.
- Comment: Is SAIC going to develop (p. 32):
- Item specifications that states can then customize to address their assessment needs? Or
- Item specification guidelines that states will use to develop their own item specifications? Or
- Both?
- Comment: How will the PEs used to develop item clusters be selected/identified? Using DCI, SEPs, or CCC? (p. 33).
- Page 31 – should have descriptions of the various sections in the bulleted list for the Item Specifications Guidelines..
- The Guidelines need to be here so that the Item Specifications can be the "additional document"
- This was confusing to me. So is it fair to say that the SAIC will develop Item Spec guideline guidelines for states? So each state will take the SAIC guidelines and use those to develop their own individual guidelines? And if that's the case, is the SAIC document only going to include those 4 sections ID'd in chapter 5? And then it would be up to member states to include or at minimum reference those 4 sections in their own individual state item specs? I think this is the intent, however I'd like to see this made a little clearer in this document if possible. I know it's a short chapter, but I found it confusing and had to really think about what we've discussed at our meetings to piece this information together. Any state that decides to join in Phase 2 may not have that same background information, so this intent may not be clear to them. Additionally, as mentioned before, any entity that views this as a public document could misread this section entirely and erroneously come to the conclusion that the SAIC was the sole source of item spec guidelines as well. That would make it extremely difficult to counter claims that this is NOT a consortium and states are maintaining a healthy amount of autonomy.
- Not sure how the prototype layout is associated with the material in this chapter.
- Needs more development.
- Needs more development of the item specification guidelines
- It sort of feels awkward as a standalone chapter, maybe add in a timeline including who will develop the item specs.
- Before the 4/24 meeting, we had a question regarding how the sample prototype could look if we had to administer a static assessment. Our question was addressed during the 4/24 call and we would suggest including the prototype information from that call in this framework.
- This section will be addressed more in depth in the item specifications documents, but some general guidelines to help guide that if items are developed by varying states and contractors, that there won't be so much variability that the items developed aren't aligned with NGSS and or don't measure the intended construct. The item specification guidelines were described only generally, so it wasn't clear how much variability might result when items are developed by multiple states and contractors. This will have implications for the coherence of the summative NGSS assessment.
- Include "development of" stimuli in addition to selection of stimuli.

Chapter 6:

- Cost and time have to be mentioned somewhere – is this the place?
- **Comment:** Is SAIC going to develop blueprints that states will customize to meet their needs? (p. 34).
- **Comment:** In addition to the guiding questions, this needs to provide guiding steps for developing an NGSS-aligned blueprint (p. 35).
- **Comment:** What would be the rationale for using NAEP-like structures (as reflected in Tables 4–6) (pp. 37–38).
- I think it would be useful if the item specifications guidelines include information regarding the ways in which items perform (generally), the amount of time it takes to "attack" each item type, cost/benefit analysis for item types, scoring issues with various item types, feasibility issues and specific psychometric concerns with various item types. This is not a negativity approach, but an approach with transparency so that states can make the best decisions possible given their specific restrictions (budgetary, length, time etc.). I do feel that all item types would be clustered within testlets—but within that framework, what are the best combinations of item types with regard to time, money, psychometrics and student tolerance?
- Achievement level descriptors section, first sentence should not say will, rather is should use language like "it is recommended", "in order to", ""you will need to". The Assessment Framework needs to provide direction at least to some extend and not just document considerations that need to be accounted for.
- I think some clarification needs to be provided in this chapter that it will be up to individual states to develop specific blueprints and ALD's. Again, those of us involved in the collaborative now have that understanding, but if it's not stated clearly, then this could be misconstrued by any entity not involved in these discussions to refute the idea that states are maintaining autonomy here.
- Also a question- if each state will develop their ALD's or PLD's on their own (not a bad idea), has there been any consideration to how this would affect interoperability of items in phase 2?
- Description of the relationship between the ALDs and the test blueprints. As presented in the chapter, they appear to be discreet activities.
- Add Revised Blooms Cog skills and DOK—would be nice to have science defined for both of these, maybe Karin Hess's work? Add concrete blueprint and reporting category examples.
- The purposes of the summative assessment should be one of the main considerations of blueprint developers, but they aren't clearly described in this document. It sounds like student-level data on the three dimensions at a fairly small grain size is desired, but the case for this level of detail in a drop-from-the-sky assessment isn't made explicit. Use cases and sample reports should accompany any blueprints. Assessment purpose and use (Herman, 2010) needs to be described.
- I'm not sure where it belongs, but nowhere in the document is "college and career readiness" addressed. With that focus in the Common Core and the main consortia tests, can it be addressed in this document?

Chapter 7:

- examples
- **Comment:** SAIC mentions that "the goals of the NGSS are outlined as PEs that must be translated into **content frameworks** to enable the specification of items, tasks, and test blueprints (First bullet, Key Takeaways, p. 14). **We need clarification on the following:** What are content frameworks? At what stage are the content frameworks developed or needed in the assessment development process? How are these developed? Or Are these state standards frameworks developed by individual states?
- Page 39: last paragraph discusses "unpacking the standards" – not clear if this means the evidence statements and if they will be sufficient (although it sort of states this).
- Page 40:
- Section on Unpacking the Performance Expectations repeats information from previous page.
- Need more information on Evidence Statements and how they will be used.
- Page 41: How, who and when will the "teams of classroom teachers as well as content experts and assessment experts" work on bundling PEs?
- Page 43: It would be helpful to have a more detailed item review protocol for the collaborative and/or states to follow.
- Page 44: Need more information on pilot testing (who, when, which students, etc.)
- Examples are lacking.
- Item review process need to be tighten up, perhaps looking at the Smarter Balanced process and criteria.
- In terms of Item Specifications in the review process, it should say meet the Item Specifications developed by the collaborative.
- A general note on ensuring that we aren't trying to state that the standards and/or test dictate curriculum. I think we can make minor adjustments to language to make this distinction more clear (marked in PDF).
- A general note and/or clarifying question: It appears that PE bundling will be determined by the individual states. Is this correct? And if so, how does that affect the share-ability of items developed to a particular bundle of PE's? I guess the real question is- will states need to dictate how PE's will be bundled for LS purposes (again, having an effect on curriculum and instruction), or is there research that suggest students would be able to apply their learning regardless of the bundle and/or phenomena chosen on an assessment? And does that have implications for assessment item validity and reliability?
- More guidance could be provided as to the process of "unpacking" the PEs.
- Key takeaways lack information about development, review and piloting that is shared in the chapter. Consider highlighting pieces of this work as it relates to item cluster development
- In Bundling on page 41 – does this imply that bundling will be decided prior to developing the assessment? Will this bundling be shared, and implications or consequences if it is?
- In the development of item clusters on page 41, I felt that the phrase "certain elements" needs clarification.
- Can this chapter be included in chapter 4? Seems redundant a little redundant. Maybe put closer to chapter 4? More information about how to break down and cluster PEs/standards would be helpful—a couple of concrete examples would go a long way here. There was some concern over creating a document that will be "breaking down" the standards and that this document could become the default standards document. Can this concern be addressed?

- No, but we feel that we should have been reading about these items earlier when we talked about bundling/item clusters. Could you reorder the chapters so that this chapter is moved closer to the bundling/item cluster chapter?
- We suggest more detail regarding the relationship between evidence statements and item cluster development. Sample item(s) could be useful to illustrate how the process outlined in this chapter could come together.
- We appreciated the descriptions of the development processes, particularly the inclusion of current educators.

Chapter 8:

- Accommodation possibilities and accessibility possibilities could be mentioned.
- This is a very strong and clear chapter. Something regarding universal tools, designated supports, and accommodations (see SB) may be helpful.
- A topic not covered here that needs to at least be raised in this chapter is the design of assessment forms and items for those students with the most significant cognitive disabilities. This is a tough topic, I understand, however we do need to acknowledge it as states are required to deliver an alternate assessment as well. Will these items be appropriate for use on an alternate assessment as well? Whether the answer is yes or no, we need to acknowledge it…even if just to say this will be another individual state decision.
- Additionally, there is some inaccurate information regarding the PARCC consortium's Access and Accountability manual as well as their approach to accommodations for SWD's and ELL's, as well as accessibility features for all students. I think this misinterpretation and misunderstanding of the PARCC manual has given the impression of an apparent bias against the PARCC consortium and ends up placing PARCC in a negative light which is inappropriate. Specific examples are called out in the marked up document, but for clarification purposes here the A&A Manual calls out Accessibility features and accommodations, both for ELL's and SWD's, in the manual. Appendix A translates those features and accommodations for students taking the paper-based form knowing that the goal is to move entirely towards online testing. Page 44 in the PARCC manual begins the section on features PARCC deems appropriate for ELL students, including extended time, word-to-word dictionaries, and translated directions (same as SB). PARCC also allows for the complete translation of the math portion into Spanish as long as this is allowed by individual state school code.
- Good.
- Not really, but we feel that accessibility for all kids wasn't really described (except in the key takeaway)—instead, by referencing the subgroups listed in NCLB it makes it sound like we are only addressing for these groups, not all students.
- No additional information needed at this time.
- The connectedness to culturally responsive pedagogy needs to be further developed as we continue to address the diverse needs and accessibility of students. Did this section address closed captioning options? I know this was a point of question for PARCC and SBAC

Question 7. Was enough attention given to Evidence Statements?

- It's hard to address them when they are not out!
- More detail is needed regarding evidence statements.
- No. It is not clear how the evidence statements will be used for the development of assessment items. Will they be part of the item specifications?
- I don't think there was enough information shared here. The document says that evidence statements can serve as supporting material in the design of curriculum and assessment (from front matter for the ES) and goes into some bit of detail about what that means. However—I know that Jim Pelligrino and others used (all but INVENTED) evidence statements in their work with Advanced Placement in order to outline the Evidence Centered Design basis for item development. It would be exceptionally helpful if you could incorporate some information for how they did that HERE, so that testing companies/states and others could use the information to work with item writers in the development of NGSS items. Page 40 provides a summary of what Evidence Statements ARE, I think this is the kind of document where it is important to illustrate what Evidence Statements are meant to DO, how they are to be used, and exactly what that process entails.
- With evidence statements, as well as the other big concepts/ideas interspersed in the Assessment Framework, illustrative examples—what the result looks like when it is applied/used would be very helpful. In general, we need to have examples applying these to assessment situations, particularly summative assessment situations. Once again keeping in mind the next steps each of us will be taking and a possible Phase 2
- I'm sure others have asked, but will we be able to incorporate MS and ES evidence statements as well? I believe Stephen mentioned there is a draft ready and they are close to being finalized.
- Consider sharing more about what the ES are and are not. It is also important to emphasis how the ES were developed - using specific structures for each SEPs at HS (and future MS/ELEM). ES are already being misused by classroom teachers as curriculum guides and lesson plans providing as much clarity as possible in how they will used in the development of item clusters and assessment in general would minimize misconceptions. Guidance on how to "bundle" using the evidence statements would be helpful as well. Examples! Page 40—Consider providing more detail/examples related to the last sentence "allow for multiple methods and context of assessment and for assessment of multiple PEs". No, I don't think so, because the last sentence on page 40 left me wanting further explanation. The statement sounds like a tall order for the ES to accomplish but it's hard to know without examples.
- Would be good to give example of one or two and break it down with claim-evidence lingo.
- No, further development on how they can be used in should be included.
- We guess so…these appeared to only be mentioned as a resource to draw upon and then later in chapter 6 and 7 as the highest level of proficiency defined. Did you intend for there to be more attention given to them? I would say that it is fine the way that it is if the only intention was to make sure that people are aware of the resource and that the level of proficiency has already been defined for each PE.
- What did not come across in the chapter 7 paragraph is that evidence statements can serve as the building blocks for the state-administered assessment. By including some of the evidence statement criteria, especially what qualifies as observable components of student performance,

the importance of the connection between evidence statements and item cluster development could be clearer.
- No. The Evidence Statements only discuss what "data" needs to be collected in order to determine if a student is proficient for a given PE and doesn't assist in determining where a student might be (learning progression wise) in regard to mastering a given PE. The role of how the evidence statements could be used in regard to developing the achievement level descriptors (ALDs) for an NGSS aligned assessment would also be helpful. I don't know what more would be added here at this time, as we haven't viewed all of the evidence statements. However, this section may need to be revisited with more emphasis on the connections to the Evidence Statements.
- I'm not sure yet. I don't think I know enough about the implications of the evidence statements for each grade level.
- No — see previous comment for chapter 2. A discussion/specific examples of how the overlap between specific parts of evidence statements from different PE's can facilitate the bundling of PE's for a cluster would be helpful

Question 8. What additional information needs to be provided in the Assessment Framework to accurately explain the item cluster design?

- I think an example would be good. It would be lengthy but would be easier to wrap your head around. Many people haven't worked with the NGSS and would not understand what even unpacking is.
- It would be helpful to have more information about cluster stimuli (page 42). The criteria should have more detail and it would help to have examples - perhaps non-examples as well as good examples. Should also include information about copyrighted information used for stimuli.
- At what point do we stop assuming that psychometricians will just do their job (magically) and offer them suggestions on how we expect them to do it? Page 30 begins to hint at the fact that psychometrics will play an important role in assessment, but it doesn't BEGIN to offer suggestions for how this can or should be considered. I think it should—because when states are navigating RFP's with testing companies, many will want to talk about the use of Item Response Theory, or Computer Adaptive Testing models (based on IRT) because that is the "in thing" right now—but with these testlet type assessment items IRT won't work—and CAT testing will be difficult. IRT is based on the assumption that each item on a test is individual—without one affecting the outcome of another. But testlet exams, by nature, have items that DEFINITELY affect one another—therefore IRT models will not work—and generic CAT algorithms (usually based off of IRT models) will not work either. There needs to be a psychometric discussion in this framework to allow states and testing companies to understand the technical requirements of a test like this.
- Mark Wilson and others have research into testlet models—their research would be valuable assets in a section such as this.
- A very good question .We are not sure of the answer but know it must be answered before significant progress will be possible.

- Guidance around the identification of phenomena, or maybe just resources to indicate context about the importance of phenomena for developers and writers unfamiliar with NGSS. Whether states will need to identify their own bundles of PE's or not to deliver a valid and reliable NGSA using shared items.
- It seems as though more detail should be provided about how, or things to think, when bundling PEs. Examples would be helpful of course!
- See comments within chapter comments.
- Does not appear to clearly state that the SEP or CCC in the PE can only be the ones used in creating assessment items for the item cluster. See comments in other sections.
- Nothing at this time
- Adding the further clarification from the webinar/question session on April 24th needs to be undated here. The explanation provided by Steve and Kevin on the call gave us a deeper understanding of the variety of levels and methods item writers can get at the core concepts. The need for multiple item types to get at the depth and breadth of understanding for these standards was further highlighted in the call. Updating Figure 3 of Chapter 5 (p.32) is a good example of what we are describing.
- Additional guidance around the item clarification statements, assessment boundaries, and the role the evidence statements should play in the item cluster design. Examples of 2-D and 3-D items that build up an item cluster.
- Explain why each item is 2d and each cluster is 3d.
- Explain why the items types are listed.
- Examples of item clusters.

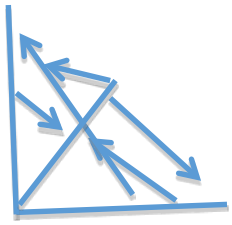
Question 9. What (if any) additional guiding questions for blueprints need to be included?

- We need to truly address time per test and the number of PE's tested.
- Need greater clarity on what tests/assessments the blueprints will be used for (e.g., pilot tests, classroom summative assessments, large-scale state assessments?).
- Perhaps this could be "brainstormed" during the face to face meeting.
- More information about how performance assessments can be incorporated into state assessment systems would be helpful. There are a lot of reports coming out about these. It would be good to have a section addressing these. Also, BOTA stresses the importance of a moving cautiously, in a deliberate manner. This cannot be stressed enough. Would like to see it more clearly stated in the Framework. Talk about a transition timeline and the importance of this—overlapping standards may need to assessed at first. Need to take into account high school graduation requirements, etc. These are critical considerations that need to be taken into account when developing/contracting for a new assessment.
- We do not have any other questions at the moment, but are particularly interested in both integrated science blueprints as well as end-of course blueprints to meet both potential needs. Just wanted it noted that content limits should be kept in the background for item specifications for later development activities.
- How many item clusters (range) should be on an NGSS aligned summative assessment?
- How many total items (range) should be on an NGSS aligned summative assessment?

- What role does a computer adaptive test (CAT) play in regard to the blueprints?
- What role does a combination of a CAT and a fixed form performance task have (similar to SBAC)?
- Matrix sampling models (examples of)
- Additional Item Types:
- Simulation
- Hands-On Tasks
- Interactive Computer Tasks
- Performance Tasks
- The technology is there to produce an innovative assessment such as through the use of simulations and gaming, the question is to what degree is it valued and how do we need to prioritize it? In the classroom we are encouraged to make learning fun. How are we making assessments fun through games and simulation? I would like to see more guidance in this area, if possible or reasonable.

Question 10. Others

- I don't mean to harp on psychometrics, but this came up in conversation in our group discussions more than once, and I realized it was a real issue. Our psychometrician truly DID NOT UNDERSTAND exactly what was meant by 3-D assessment. In psychometrics 3-D measurement in testing means something different than what we are necessarily talking about between and among items. So, I can't draw this in word very well, but imagine a 3-D graph. Within that graph an item or piece of information may or may NOT have an impact on other pieces of information within the measurement.
-



(these connections are not guaranteed)

- What are the overlaps between information? IS there overlap or do students have understandings independent from one another? This has always been an issue with the idea of "double coding" items—is it true that students truly must understand BOTH bits of information to answer the question or is it possible for one piece to be available to the student without the other piece thus creating two pieces of information in one item—not 2-D, but two 1-D within one question. (I don't know if that makes ANY sense in writing—but there it is). With all of our "3-D" language, I'm not sure that content experts are always on the same wavelength as the psychometricians they are speaking with.
- You were getting into all of this in your webinar—but I think it is important to address this in the framework to some degree so as to ensure that states and vendors are on common ground with common language when they begin to discuss these issues.

SAIC Assessment Framework Feedback from Workbook

Combined States' Responses (Post 5/14/15 Meeting)

Question 1. Who should be the primary audience of the Assessment Framework? Who should be considered as secondary audiences?

- Primary:
    - Those who are involved at the state level, State DOE.
    - next generation science assessment developers including state department of education (SDE) staff and testing contractors.
    - Test/Item developers, Science Assessment Coordinators
    - To develop state assessments to measure student performance on the new science standards.
    - State assessment personnel including content specialists, testing companies under contract to state partners
    - Assessment builders
    - State-level assessment personnel
    - Vendors
    - State Assessment Office, Science Coordinators, STEM Director, State and District Office of Curriculum & Instruction, Vendors
    - State Assessment Staff
    - Test Developers/Vendors
- Secondary:
    - Teachers, curriculum developers and others in the field.
    - Educators involved in assessment development, SDE management overseeing assessment issues. To a lesser extent others audiences such as educators, the public, etc. who may be interested in "big picture" decisions about assessments such as how results will be reported.
    - Assessment developers: SEA, Vendors
    - Internal assessment supervisors and possibly test vendors
    - Science Curriculum & Instruction Coordinators
    - There are a variety of audiences with an interest in the assessment of the NGSS and the use of information gained through the assessment process. These include: the Board of Education, educators, parents, and community members to name a few. The use of guiding questions and summaries of "essential elements" from the framework would helpful for those who may not have a need for the more technical information.
    - To influence curriculum instruction.
    - District Test Coordinators
    - Administrators
    - Teachers
    - Professional Development
    - Higher Education (could be primary)
    - District Curriculum Leads

    - Accountability personnel responsible for reporting assessment data.
    - Vendor staff, Professional organizations, USED staff
    - Teachers
    - Science Teaching and Learning Staff
    - Technical Advisory Committee members
    - Regional Science Coordinators
    - Teachers and other educational professionals
    - Professional development providers
    - States level personnel and vendors that are directly involved in the development of a large scale test require an in-depth understanding of the issues surrounding 3-dimensional assessment of learning. It may also be beneficial to create primers for secondary audiences- such as legislators/state boards, teachers, and parents.
    - States also need "dumbed down" versions to communicate with stakeholders, Legislative General Assemblies and the public in general.
    - State Department of Education staff
    - Contractors/Vendors of Science Assessments
    - US Department of Education staff (NAEP)
- Not a functional resource for parents and students as this has been created
- State Education Agency Staff (science content and assessment specialists, psychometricians, accountability staff)
- Test Vendor(s)
- Item development writers, reviewers (content, accessibility, and bias/sensitivity)

Question 2. How do states anticipate using the Assessment Framework?

- Release as a component of RFP
- As a foundation for item construction
- Vendor use for maintaining quality control
- Science specialists in learning services
- State school board
- State Department of Education leadership
- We should still strive to maintain transparency but develop key pull out pieces of the framework for communication purposes with superintendents, districts, etc.
- As possible general guidance for the development of assessments aligned to the NGSS. It will likely strongly influence NGSS assessment development in our state, although we may modify it. We will share with our management and psychometricians as well as testing contractor and our state science assessment committee to inform them of the challenges we face and direction we are headed and to get input and feedback into NGSS assessment plans. While the primary purpose is to guide large-scale state summative assessments, the framework can also serve as guidance for interim assessments.
- Internal information and guidance as well as professional development tool. I'd like to use it to teach district science supervisors and test coordinators and then elicit input from them especially as we work to develop our own RFP.

- The Framework will lay the foundation for our state's development of NGSS-aligned science assessments. Initially, the Framework and Item Specifications will serve as a tool to guide conversations with the various stakeholder groups that will be involved mapping out the implementation and assessment of the NGSS. It is also anticipated that the documents will provide specific direction to assessment developers in our state or as part of cross-state effort.
- We plan to use the assessment framework as a guiding document and quality control document for our state specific blueprints and item specs.
- Affect changes to Curriculum Instruction
- Development of State Assessments
- Communication to stakeholders, Legislative General Assemblies and the public in general.
- As a guide to help inform our future assessment development, including what we may want to include in a new assessment RFP/RFR.
- The Framework will guide development of summative assessments aligned to NGSS and must be incorporated into assessment development, including guidance in RFP for NGSS assessments.
- Within an RFP, to guide development with operational vendors;
- As a Quality Check document;
- To share with curriculum instruction at SEA level;
- Perhaps share an abridged version with State Board of Education to help them understand the key points (3-page primer);
- To guide item development as our state transitions to new standards;
- To assist with test design;
- RFP building
- Assessment quality control
- PD/Curriculum development
- Creation of support materials
- Guidance to vendors to create primary and secondary assessment materials
- This document along with the item specifications guidelines (with the NGSS and the Framework for K-12 Science Education) will be the guiding documents to assist in developing test specifications and blueprints. These documents will assist with the development of which NGSS will be assessed at each grade level (i.e. 3-5, 6-8 and High school), how many item clusters will be present on each of the different grade band tests, as well as the overall structure of an item cluster.
- Use it to develop an RFP to select a vendor to develop an assessment
- Use it as a quality control document through the item development and test construction process
- Share with Office of Curriculum & Instruction to inform their work
- Create an abridged version to use with admin/boards/districts including key issues and considerations
- To guide the decision making process and provide support/rationale for decisions including: development of test maps/blueprints, assessment system design, assessment development cycle design, committees, reviews, validity evidence.
- The Assessment Framework from the SAIC will be used as the foundational document and then customized to fit our state's needs and recommendations from our stakeholders and advisory committees

Question 3. What test design decisions that states will need to address should be captured in the Assessment Framework?

- Field test design
- Matrix sampling (as suggested by BOTA report) matrixing standards not students.
- Special populations
- Cost considerations
- Test form and design working with clusters
- Accommodations that would be acceptable/available
- Technology Specs--Is it going to be online and to what extent?
- Standard format for writing items compatible to test delivery engine. (Conform to existing standard of item writing)
- Assessment challenges and implications. Possible solutions.
- Overall design of NGSS assessments (e.g., blueprints, time for tests, matrix sampling)
- Translating item clusters into actual test forms.
- Scoring considerations.
- Reporting.
- Pilot testing (e.g., which grades, sampling issues, time of year, etc.)
- What will be the item cluster/item formats—will they remain static or are they "pliable" between clusters?
- How many items and/or clusters will be necessary for a valid measurement of student proficiency? And at what level are they "proficient?" See #11 for more on this.
- Paper Pencil versus computerized (I think you've addressed this already though)
- Grade cluster vs. grade level testing
- This is the $1,000,000 question (perhaps literally) as we work through possible designs for both formative and summative assessments of students, schools, systems. On big question is the role of "internal" assessments and "external" assessments for both monitoring and informing improvement. One thing that should be kept in mind, even if not directly addressed are the recommendations in the BOTA report (those we accept as well as those we opt not to).
- In our state, adoption of NGSS may open the door to changing assessment structures, for example, moving the elementary test from 4th to 5th grade and moving towards grade-banded assessments rather than single grade assessments.
- These types of decisions could be examined in the Framework, especially with respect to how they relate to the K–12 Framework. Essentially, what does the K–12 Framework for Science Education tell us about teaching and learning and how does assessment fit into that landscape?
- There needs to be some discussion about the "claims" that can be made with the information gathered from this assessment.
- For example, "Claim: The student is college and career ready for biology" after having only passed 1 life science item cluster that addresses at most 3 PE's. It would be helpful to include information about the types of claims that can be supported by data from this type of assessment.
- Whether on-line or paper and pencil or some combination of both.
- Agreeing on accommodations to be supported (see question 14).
- How long in testing years will it take to fully implement the Assessment Framework?
- A strategy for modifying the assessment model as
- Pilot/field test design issues need to be addressed

- Matrix sampling needs to be addressed—how many do you need?
- What makes a valid test, especially if you are assessing standards over a 3 year period?
- Clusters—how do they come together to form a test?
- What coverage of the standards is necessary to make the test valid?
- Special populations take into consideration with the design.
- Cost considerations (especially for the development)
- List of differences to consider between a state and local assessments
- The Framework needs to guide development of assessment blueprint (inclusion of UDL, ELL, etc.). Clear guidance for developing PE clusters (bundles) in order to translate into development of items that are aligned with and accurately measure the NGSS.
- Addressing field test /pilot issues
- Assessment framework guidance
- How to address NGSS to provide adequate coverage of content standards
- How many items do we need
- Future considerations of TE-items
- Matrix sampling
- How will clusters work in the test design process – and still be valid?
- Coverage of standards
- Special populations
- How do we incorporate UDL with NGSS clusters?
- Design and cost considerations
- Pilot/field test issues
- Matrix sampling issue
- Grades vs. grade bands –what is the balance?
- Access and equity issues for underrepresented and special needs groups
- Budget and the framework's ability to address the myriad of issues budgets create
- Facilitate conversations that would forward conversations that could take place with policy makers to build assessments that align to our state's standards.
- How many item clusters should be on the elementary, middle and high school grade banded assessments?
- Matrix sampling vs. Computer Adaptive Testing (Will the Feds. Approve of Matrix sampling for the ESEA Science testing requirement)?
- What types of stimuli will be used?
- How many different types of stimuli will be used?
- Braille, closed captioning, translation options (How to develop an equivalent Braille version of a test that might contain simulations and machine scored graphic response (grid) items?
- How to afford the development of a new science assessment?
- What our test blueprint will actually look like? (examples would be helpful)
- Hand scoring, vs. machine scored, vs. artificial intelligence……
- Score reports, score reporting categories?
- Claims/statements about student performance being measured by the assessment
- Standards being assessed
- Process used to cluster items
- Guidance around developing ALDs
- Guidance around score reporting
- Address field test design

- Address matrix sampling
- What would make test valid (over multiple years)
- Coverage of test to make it valid
- Address specific populations
- Address cost considerations and ideas to minimize cost
- Test form/test design & how much of it is covered
- PEs to be assessed, e.g., how to identify standards for a 10th grade high school test, 11th grade test; 8th grade test?
- Guidance on test maps/blueprints
- Assessment system design
- Assessment development cycle sequence and components, committees, reviews, etc.
- Field test design issues
- Matrix sampling recommendations

Question 4. How should grade levels, grade bands, and grade spans be addressed in the Assessment Framework?

- Grade levels, bands suggested by the framework should be flexible and should define the benefits and risks of testing at different grade levels/bands, but should be a decision of the states.
- The assessment framework should be flexible on this issue since various states will treat this issue differently. In our state, we will likely have grade-by-grade standards for K-8 with more flexibility in high school. The grades that items should be targeted for and pilot/field tested with should be addressed.
- Given that the standards are written at grade levels at the elementary grades and then are written at grade bands in middle and high school, I think it is important to address them as they are written. State writing teams can always combine content from grade levels to create a band of tested information if they want (elem. Level) but already banded information will need to be BROKEN OUT INTO GRADES through POLICY in order to be assessed at grade levels for middle or high school. This needs to be spelled out and addressed (made very clear) for state officials, LEAs and vendors prior to making random grade level placement choices for testing purposes.
- This should follow NGSS.
- For elementary, grade levels should lead the discussion, keeping in mind that “testing structure” varies from one state to another, with some testing grades 3, 4 and 5, some only grade 4, and some only grade 5. There should be at least some general discussion of what these scenarios might mean in terms of NGSS assessment.
- In middle school, the DCIs should lead, with the discussion focused on the 6–8 grade band. However, the Framework should probably acknowledge that there may be both curricula and assessment differences depending upon bundling and the grade level content arrangements. The same types of scenarios mentioned for 3-5 apply to 6-8.
- At the high school level, the Framework will need to include a more detailed look at the various assessment possibilities given “common curricula models” and course taking

patterns/requirements. Integrated science, course related, cumulative at say grade 11, etc., are all possibilities.

- This will ultimately be a state decision and many states have legislation guiding these decisions. A brief discussion of the pros, cons, and considerations for each of these would be very beneficial.
- Different models should be developed to support not just states administering state science assessments at grades 5 and 8, the end of the grade range, but other possible combinations of grades. States might choose to test at every grade or at grades 4 and 7.
- Ideally, it would be great to have different examples—grade bands and grade level expectation examples.
- This may have to be fairly flexible due to changes that may occur to federal and state policies. In order to be of accommodate as many partners states as possible, grade bands would most likely best guide assessment development, e.g., K-2, 3-5, 6-8, 9-12.
- Must be flexible – and describe benefits and risks of each design as states will make their own decisions;
- What statements do states want to say about science scores?
- What will ESEA specify for state requirements, especially as they relate to matrix design?
- Peer Review considerations
- Create pros and cons regarding the use of each; use the framework to strike a balance of what combination of grade levels, grade bands, and grade spans should be used to address the most pros across the board and negate the cons.
- Flexible- In our state we have assessed using grade bands (3-5, 6-8 and High School) and will most likely continue to do so. What might change is when (11th grade vs. 12th grade) the high school assessment is given and counted for participation in regard to ESEA.
- I would hate to see the high school assessment only address an EOC Biology, because then ability for students to show how the crosscutting concepts are conveyed across the different disciplines of science is eliminated. (Especially in regards to Earth/Space Science).
- Pros/Cons of different options would be beneficial.
- Pros/cons depending on each scenario
- statements of achievement/growth of student learning in Science
- What are states doing to meet ESEA-NCLB
- Also, advise on matrix design
- Propose possible models/recommendations for determining the PEs to be assessed, including risk/benefit for considering each.
- Elementary – since most states assess only a single grade, 4th or 5th, should the PEs eligible for assessment include that grade only or should PEs from a previous grade also be eligible?
- Middle – if tested in 8th grade, are PEs from the entire grade band eligible?
- High school – recommendations for determining the eligible PEs depending on the grade when the test is to be administered.

Question 5. How should the balance between the breadth and depth of the performance expectations (PEs) be addressed in the test design?

- This question is probably best answered after you answer who are you testing and how has that test taken shape. Also keep in mind that the standards themselves are written to increase the breadth and depth of science knowledge, so the test should be designed to measure that way as well.
- The breadth will be addressed with the matrix. The DOK can be used to address depth which would be determined by the task - we are probably going to get 4's especially with performance tasks.
- The question relates to multiple levels in a comprehensive NGSS assessment system. Local, classroom assessments can address greater breadth & depth of NGSS (over time) while external, monitoring assessments should assess "bigger picture" or "enduring understandings" of NGSS. The Framework should address how summative assessments fit into the bigger system.
- Regarding breadth and depth of summative assessments, two or three options should be presented to address this issue with advantages and disadvantages of each. Use of matrix sampling should be addressed as well.
- For TEST development limit the depth and breadth to the content attributed in the DCI, CCC and SEP already "unpacked" in the colored boxes of the standards documents. To add more depth or breadth would add a layer unfair to the assessed party by asking for content not made explicit in the outline of "required" material. Feel free to teach more—but THIS will be fair game for assessment.
- This MAY be a concern for PE's where arguments are "well you can't understand this unless you've done that." In my state—that's all well and good, but the test will ask what's listed in the black, white, orange, green and blue of the documents—nothing extra—because at the end of the day, what is in that document is what we said would be assessed when we adopted the standards. It's considered "fair play" to have what will be tested in writing and not assumed.
- There is the theory of the K–12 Framework/NGSS and the practice of how states are implementing assessments. The K–12 Framework/NGSS would suggest that the standards already present a balance of breadth and depth, which should be reflected in the assessment. However, states who have adopted (or are working toward adoption) may not have an assessment structure that supports this. For example, in HI, we currently test 4th grade (4th grade standards only), 8th grade (8th grade standards only), and HS Biology (EOC, Bio standards only). Somehow, the Framework will have to acknowledge that there is an ideal way to address NGSS and then the realities that states will face.
- A discussion of the tradeoffs between breadths vs. depth would be helpful here. This is ultimately another area where states are going to have to make decisions and the more information available the better.
- First, we probably need to confirm our working definition of 'depth' and 'breadth'—I would propose the following: depth=complexity (reasoning/thinking); breadth='big picture' or entire context.
- I believe that a matrix sample or matrix/common hybrid address the breadth issue—as best as is possible. We prefer not to prioritize or otherwise eliminate any of the PEs from the pool of 'potential' for assessment purposes.
- In terms of depth, we owe it to practitioners to embody the INTENT of the PEs (the intersection of the 3 dimensions) as the driver for depth or complexity. This requires a focus always on at

least a practice or a cross-cutting concept with each DCI component of the PE per task or item. Ideally, however, congruency to the intent of the PE requires that for every 'bundle' there must be one component/task/item that fully gets at the 3D-ness of the PE or PEs within the bundle. If a student can't ultimately address all 3 dimensions, we can't say they have 'met' the standard.

- Depends on how states want to report information out. Needs to be flexible. Again, give different examples: clusters of PEs; content, practices, crosscutting concepts.
- Other questions need to be considered before this question can be sufficiently answered:
- Who will participate in the assessments (grade levels, ELL, IEP, etc) ?
- How will the assessments be administered?
- What needs to be reported (for what purposes)?
- What do we want to report?
- How do we want to report: student level with matrix sampling?
- Must be flexible to be used for different purposes and different states;
- The important question is to figure out why we are doing this in order to address the breadth and depth needs of the assessment. How does this design address what would be needed at the student level, school level and district level while addressing federal accountability? There are different needs for reporting here and it would be great that everyone could look at this and get what they need from the test reports without being buried in information not relevant to them.
- Flexible. It might be helpful to show examples of different bundling approaches that could be used, as well as show an assessment system model (similar to SBAC) that shows how certain PEs would/could be assessed in a rotation fashion (matrix sampling).
- The PEs state what a student should be able to do/perform at a proficient level and so additional guidance around how to use the PEs for items, ALDs and score reports would be extremely helpful (what about developing items that exceed the PEs and or are below the proficient level (concerned about a floor and/or ceiling effect).
- What have we answered, who, what, why
- Ask more questions before depth/breadth and Student level vs. matrix sampling
- What are the stable points across states, ex. Range of 3 variables
- 1st: what do we want to report and how are we going to use it?
- Provide content sampling plan recommendations
- Will need to consider what information will be reported
- Consider Model Course maps from Appendix K of the NGSS, course sequences, end-of-course/through-course vs comprehensive; summative vs. classroom/interim

Question 6. In general, how do states anticipate using the NGSS Evidence Statements?

- The NGSS evidence statements were designed to give indicators of what a student should be able to do to be considered PROFICIENT (not above, not below proficiency)
- The evidence statements can give insight into the bundling of PEs. They can be used to develop tasks
- Unpacked standard version.
- They can be used to find bundles for assessment.
- The evidence statements should not be used as a curriculum.
- To provide more specific guidance in the development of assessment items. The evidence statements provide greater detail as to what is expected of students in the performance expectations, although some evidence statements need further elaboration. Evidence statements can also be used to assist in the bundling of PEs since they can provide guidance as to the overlap or connections between PEs.
- We will use the NGSS Evidence statements in a process similar to Understanding By Design. To really work around and through what you need to do with kids in the classroom, you have to understand what you need/want them to know and be able to do when you are done. The Evidence Statements and PE's provide a scaffold for understanding exactly that—what should kids know and be able to do once you've taught them science?
- Now—how do you teach THAT? How do you get to THAT? What would teaching and or learning THAT look like?
- From an assessment perspective the question is—how do you get them to show you THAT?
- The evidence statements indicate what students would have to do to illustrate their knowledge/ability. Now—what prompt and/or question would elicit that sort of evidence?
- These may form the basis for proficiency levels. They will also serve as guidance, along with the Item Specification and other resources, to assessment developers.
- The evidence statements are indicators of student proficiency of a particular PE, so we would use them as the indicators of Proficient.
- We plan to use those as 'success criteria' that defines what proficiency looks like—essentially, the 'proficient' cell of a rubric or scoring guide. We'll also have teachers practice using them to provide/inform FEEDBACK on student work—as well as a means to gauge the quality of tasks/learning experiences/items…in simple terms, adding the words "Did students (text of the ES)?" will enable teachers to look at student work for specific markers of quality and then give focused feedback on that element of success; and adding the words "Will students be able to (text of the ES)?" as a 'screen' for looking at lessons, units, tasks, etc. as a means to considering if those experiences even give students an opportunity to demonstrate attainment of the standard as intended.
- As a resource for developing item specifications.
- Not sure how else they might be used…they are very new
- The Evidence Statements will guide the development of items, possible Rubrics (may be item specific or generic), and proficiency levels for student responses on assessments.
- As a resource for developing item specifications;
- Similar to claims and target reporting
- As a resource for item development
- To address the question of what do we want students to know, show
- As proficiency statements, to help guide Performance Level Descriptors

- We plan to use them as a resource to align our state assessment with classroom assessment. We made use the high school evidence statements while building our End-of-Course and anticipate doing the same upon the release of the K-8 evidence statements.
- Assist with the development of different item types and to make sure that the items being developed will measure the appropriate rigor of each PE or bundle of PEs.
- To assist with the development of item rubrics.
- Achievement Level Descriptors
- Define what a student can do to demonstrate proficiency level
- For use in formative and should be used to help coming-up with cluster
- How do you prevent misuse?
- From the assessment perspective, the Evidence Statements will be embedded in the item specifications (as already shown in the Item Specs Guidelines draft) and will also guide the expectations for evidence of "proficiency" in student work.

Question 7. What role should learning progressions play in the assessment design?

- As the learning progressions are envisioned, and not tested, we should be cautious about them. They have not been proven as the standards have not been tested across the grades.
- The learning progressions should be used in a more formative than summative way.
- The learning progression does flow but is dependent on where the student is.
- Summative assessment use of learning progressions may come after test has some results.
- The learning progressions were used as the foundation of the development of NGSS and are therefore already built into the standards. The learning progression may be helpful in the development of specific scoring rubrics for some items and in the development of the Achievement Level Descriptors.
- Learning progressions should be used in scoring guides to diagnostically assess student understanding and provide those assessing with information that can be used to improve or change instruction/curriculum/practices or other issues identified by the data received through assessment. See below for details:
- Duschl et. al. posit that "well-tested ideas about learning progressions could provide much needed guidance for both the design of instructional sequences and large scale and classroom based assessments" (p. 22). But how will learning progressions change the face of assessment? And to what end? To recognize the anticipated effects of learning progressions on assessment it is important to first recognize the difference between current practices and proposed practices.
- Current practices in large scale, as well as most classroom, assessments are most often based on a unidimensional item response theory (Torre, 2009). Under this model, students are asked to respond to items that have a distinctly "right" or "wrong" answer with the underlying assumption being that "students with higher proficiencies [will] have higher probabilities of answering the item correctly" (Torre, 2009, p. 164). But using learning progressions to build assessments merges cognitive and psychometric theories that could facilitate a more diagnostic approach to assessing what and how students understand a given construct (Torre, 2009; Alonzo & Steedle, 2007).

- Alonzo and Steedle (2007) propose two viable methods for cognitive diagnostic assessments using learning progressions. First, the authors suggest using open ended items with diagnostic rubrics tied to learning progressions. Whether asked orally or in writing, student responses can be recorded and evaluated using a diagnostic progression rubric. Such a rubric is described more completely in the work of Jeffrey Steedle and Richard Shavelson (2009) who explain the use of double digit rubrics where the first digit provides a "relatively broad description of student understanding (macro level) and the second denotes a particular instantation of the first (micro level)" (Steedle & Shavelson, 2009, p. 704-705).
- Using the open ended response format, Alonzo and Steedle found that response consistency was intrinsically tied to and influenced by the context of the question. The authors explain that the low 60% consistency rating illustrates a significant problem for large scale standardized testing using these item types. But used in a classroom, "a teacher could profitably use information about his or her students' consistency to examine issues of transfer, helping students see the same underlying principles applying in a variety of situations" (2007, p. 417).
- In addition to the open ended response items, multiple studies have investigated the viability of ordered multiple choice (OMC) items tied to learning progressions (Torre, 2009; Alonzo & Steedle, 2007; and Briggs et. al., 2006). OMC items are built by pairing a question stem with at least four response options, all of which correspond to defined levels within the tested learning progression (Alonzo & Steedle, 2007). Generally speaking all of the response choices are "correct" within the parameters of a defined learning progression and the option chosen by the student is illustrative of his/her level of understanding. When given a variety of items, it was found that students answering OMC items illustrated their placement on a learning progression with equal, if not slightly better, accuracy as those students assessed using open ended or interview techniques. In fact, OMC's proved to not only be easier to score, but "they provide[d] a more unambiguous and slightly more precise indicator of student thinking" (Alonzo & Steedle, 2007, p 416). Again, however, the authors warn that large scale assessment would not be as well served using these diagnostic item types. Though the OMC's seem to accurately reflect student thinking "there is a danger that we will over identify students as holding the "A" level conceptions . . .[and while]. . . a classroom teacher could easily [check] in with his or her student to find out" what the students' complete thoughts are, "this luxury is not available in standardized testing situations and no "easy fix" to this problem has been identified" (Alonzo & Steedle, 2007, p. 417).
- Understanding the types and successes of assessments tied to learning progressions is exceptionally important not just to knowing how to use progressions in the classroom, but for determining how to measure a progression for validity. The National Research Council claims that a well-constructed learning progression can not only outline the most common path toward concept development, it can identify likely roadblocks and misconceptions along the way (Duschl et. al., 2009). Both OMC's and open ended items using double digit rubrics have been shown to be viable tools for measuring the alignment of learning progressions with student performance outcomes. The information gained could be used judiciously to inform instruction, curriculum development, assessment strategies and standards reform (Duschl, Schweingruber & Shouse eds., 2007; Wilson, 2009; Corcoran et. al., Summary, 2009; Songer et. al., 2009 & Alonzo & Gearhart, 2006) assuming, of course, that the learning progression in question is properly researched, developed and validated.
- The role will probably depend on how the state has adopted NGSS and the assessment structure for the state. The Framework will have to look at learning progressions, especially as

they relate to bundles. However, the Framework will also have to acknowledge that different course sequences will mean different learning progressions.

- It is important to ensure the utilization of learning progressions. It has been useful, for example, to discuss learning progressions in the context of including science at all grade levels, not just "tested" grades.
- Learning progressions could inform bundling as well as the selection of phenomena or context for bundled item sets. I would think they would also help inform item/task specific scoring guides.
- May provide information for Achievement Level Descriptors
- Be cautious about using them, not a ton of research on them
- Learning Progressions need to be used to guide the development of valid assessments of "what students really know and are able to do." A caution approach needs to be taken when gaps are detected: Are these gaps in learning or instruction?
- Integral part of NGSS;
- Caution: acknowledge them and the fact that they're not fully proven
- Used for instructional purposes perhaps, not for summative purposes
- ALDs/PLDs: to categorize student scores
- Provide information with achievement level descriptors (used cautiously) so the assessment framework is usable across the different states. Learning progression use can determine our ability to align our assessment with district curriculum. This alignment would address formal and informal educational platforms.
- Could possibly be used in the Achievement Level Descriptors (By the end of grade 5, by the end of grade 8 and by the end of grade 12).
- If a computer adaptive test design is used, then the items that are present in the bank could consist of items from each of the different aspects of the progression and the student score report might be able to display where a student is on the given learning progression (achievement level).
- Development of formative assessments
- Development of descriptors of student performance for each claim by grade level
- Assist in making determinations during achievement level setting.
- Perhaps progressions could be considered the development of a set of items for an item cluster to be used in an adaptive test. Do we include any out of grade items in a cluster?

Question 8. How should the Assessment Framework address automated scoring of constructed-response items?

- I think that different forms of scoring should be considered and should include:
- machine scoring, human scorer, and teacher scored pieces that result from classroom embedded pieces.
- If we are to assess a hands on task, we would need to include teacher scored pieces.
- Because of cost and time, computer scoring would be best. Is the technology where it needs to be with this?
- We have to make sure the constructed responses are very specific to the practices. The items have to be balanced and appropriate.
- It would be helpful to include information about how automated scoring might be used and what the latest research shows about its effectiveness (perhaps in an appendix). It would also be helpful to provide information about which types of items automated scoring could be used for.
- With the honesty that I have so far seen in the document(s). Automated scoring of CR items is not well established and is often not reliable when compared with human scoring results under most circumstances. Natural language items are most often scored using key words and do not often take into consideration the contextual use of the key words, thus introducing incorrect scoring when compared with human scorers. Offering the information that exists with all of the pros and cons (cost and accuracy included) is what should be included in the framework document.
- Decisions get made by states and districts. Information is all that can be offered by the framework.
- There are a variety of positions associated with this issue and, as new machine scoring algorithms are developed, there may be new tools that are more valid and reliable.
- Bottom line: The Framework should address/discuss the role of constructed response items and not focus on the process of scoring such items at this point in time.
- It would be beneficial to include information concerning the pros and cons of AI scoring as states are dealing with a variety of constraints/circumstances.
- At this time automated technology is not sufficiently advanced to consider for a summative assessment. The approach is to assume human readers for the first couple of years and continue to investigate and conduct trials in future years. To get teacher buy in, one should consider offering scoring with a human reader and a second read using automated scoring.
- Give different ways to score CRs –AI, teacher scoring, distributive scoring, classroom-embedded, etc.
- Research into how, where, and when AI scoring has been used for Science assessments. Maryland has been successfully using AI scoring for the grades 5 and 8 MSA science assessments for several years.
- Fiscal constraints with hand-scoring
- Proceed with research backing reliability of AI scoring
- Look at the different ways and the different types of items, and then the best scoring recommendation can be made based on the given research.
- Flexible- give the pros/cons of AI scoring
- Provide recommendations on constructed response items that could be machine scored, scored by artificial intelligence, and responses that should be human scored.
- Provide recommendations based on the research and links to current research on automated scoring.

Question 9. How can item specifications and item cluster alignments be prevented from becoming default curriculum in states?

- You might think to keep these specifications secret, but why, tests should test the standards and the standards should be taught. The test design should vary the types of clusters every year. We should strive to have a great bank of items from which to draw.
- We need to have a communication plan and let teachers know that there is more than one way to teach and bundle a PE
- If the item specifications and item cluster alignments are public, then it is very hard to avoid them becoming the default curriculum although proper communication on this issue can help. In our state, we have not released item and test specifications for this reason. They are only used internally for assessment development purposes.
- In many ways this is the job of professional development staff and administration. However, in looking over the specifications document you've provided so far, the specifications are not prescriptive and would be difficult to form into a "curriculum." A specific item spec and alignment would not be released to the public until the item was released, which would not be until the item had already either been used or rejected. And from what I'm seeing in the document so far, different items would have different specifications dependent upon the PE or bundle of PE's that were used to develop the stimulus—therefore, nailing down the "curriculum" from a set of released items would be unwise.
- On the other hand, knowing how item specifications are created, how PE's are bundled and how item types might be linked with certain practices and or combinations of standards dimensions would only serve to help (in my opinion) a teacher or curriculum developer understand the standards more deeply and gain an appreciation of what sorts of LESSONS and phenomena would help a student gain a broad ranging understanding of the science that would prepare them to answer any set of items/clusters dealing with the NGSS that might be thrown at them at any given time.
- In some ways, there is no way to prevent this. On the other hand, if the bundles on which items are based are conceptually sound (which we would hope they would be), these could actually be useful to states/teachers. Since NGSS PEs are so rich, the bundles could become of basis of some very effective instructional units/learning opportunities.
- Our state has had on-line adaptive science assessments for a number of years. If the adaptive nature of the assessments is maintained across DCIs and PEs there are exciting ways for the assessment to provide students with new contexts in which to apply their knowledge. This will be an exciting assessment challenge and one, if successful, which would help teachers to realize that "teaching the bundle" is not sufficient.
- This is a challenge as often phenomena become curriculum, particularly since that is how we are looking at bundling PE's together.
- We do not have a good answer based on past experience. The saying "Preaching to the Choir" comes to mind. We say don't do and schools do it anyway. When an assessment is part of high-stakes accountability, people naturally are going to do everything possible for their students to perform well on the assessment.
- Should make them public. Will not become default standards because document is complex, long. Most educators will still rely on the standards.
- Identifying the purpose of item clusters and develop various sets of clusters, do not rely on only one set of item cluster alignments to guide development of assessments.
- Bundling may vary by state (PE);

- Need to explain that states make comparable assessments from year to year, not identical assessments and items from year to year;
- Clusters should be based on good instruction;
- Be careful not to use favorite classroom scenarios
- Consider a modular type of arrangement of the clusters so that schools that use an integrated approach are not hindered if the assessment ends up looking like an end of course assessment;
- Be able to offer assessment at different points in year.
- Bundling could play an important role in this piece. The information will always show the range of information that will be assessment knowing there are a myriad of combinations that do not make it worth it to just focus on the content. This way, teachers will teach to the breadth and depth of the standards and not try to game the assessment.
- Matrix Sampling
- PE cluster
- Computer Adaptive Testing (CAT)
- Multi science discipline based item clusters
- Multi grade level PEs bundled together
- The item specifications and item clusters will be a valuable resource in the redesign of science curriculum aligned to the NGSS, but states should follow the recommended steps in developing a curriculum which is to first look at student needs and then select learning experiences and order those experiences to meet the needs of students and objectives of what they should learn for the school year.
- A content sampling plan that varies the item clusters/PE bundles assessed each year.
- A continually evolving set of item cluster alignments as we learn more through the item development and field testing processes
- Rich classroom examples using a variety of PE bundles

Question 10. What are some alignment expectations for items (i.e., what are key questions that items need to be evaluated on)?

- DOK
- Standards
- Range of alignment
- Perhaps there will be more of a primary, secondary standard alignment.
- Alignment should be specified for NGSS (PEs, CCC, SEP, DCI, aspects of the evidence statements) and Depth of Knowledge. Alignment may also depend on whether claims and targets will be used in an evidence-based design approach.
- Obviously the DCIs, CCCs, and SEPs must all be there. The key question for the Framework to address is how items can be brought together in a fair, valid, and reliable way to measure these three “independently “as well as an “interdependent” whole.
- Do the questions/tasks reflect the cognitive complexity required by the standards?
- Do the questions/tasks assess the depth and breadth of the standards at each grade level/grade span?
- What conclusions may be made about student understanding?

- How does the question/task align to the evidence statements?
- Are the items in a cluster interdependent or can they stand alone?
- Do they meet NGSS Standards?
- Each state will need to vet items to meet State Standards.
- Will the time to administer an item negatively impact overall time needed for the assessment.
- Will students have received instructed in the area being assessed
- Would like to see psychometricians weigh heavily in on this.
- (Note: Dual alignment: what makes it dually aligned? There is a lot of gray—it is not black and white.)
- Key Questions:
    - alignment to the NGSS
    - dimensions assessed (no one dimensional items)
    - authenticity of data used in stimulus
- Examine the DOK of the standard and determine what could make an item dually aligned;
- Look into dual alignment to more than one science PE;
- Also consider aligning with the CCSS math and ELA pieces that fit.
- Items need to be 100% aligned to the performance expectations and should be connected to the item cluster guidelines.
- Which science discipline (s)? Possible multiple tags/alignment
- 2 dimensional vs. 3 dimensional
- Item Type (CR, SR, TEI, PT, SIM, etc…)
- Appropriate level of rigor
- Are there similar items that if a student answers them, they will be able to answer other items (i.e. Frenemy items)
- Key Vocabulary (which vocabulary needs to be defined in the stimuli?)
- Reading Lexile of stimuli and Item cluster
- Is the item cluster aligned to a given PE or set of PEs?
- Which dimension (SEP, CCC and DCI or combination of) is being measured?—tricky!
- (Even with a 2 point or 3 point item, it might be hard to determine which dimension or combination of dimensions a student is proficient in).
- DOK of item vs. DOK on cluster
- Is the item aligned to:
- The intended item specification(s)?
- The DOK of the PE?
- The intended Evidence Statement(s)?
- The intended DCI/SEP/CCC?
- The appropriate learning progression
- Bias-ELL, Sped
- Appropriate item type(s), TEI only when item is actually enhanced by TEI
- Dimensionality

Question 11. What are some guiding questions for blueprint design?

- How do we address the engineering standards - what are those tasks going to look like?
- Dimensionality - time testing - number of standards tested - types of items - readability - user-friendly
- Special attention needs to be focused on graduation requirements - what does an EOC look like through NGSS eyes? Matrix testing could come under public scrutiny.
- Score reporting guidelines
- Device usage/platform
- Distribution and value of items
- Partial credit considerations
- What are the claims and/or targets (reporting categories)?
- How much information do we want to get from the assessments?
- What is the overall time of the test?
- Will the assessment(s) include fixed forms or be adaptive? Is a mixed design possible/desirable?
- Will matrix sampling be used? What matrix sampling designs are most effective?
- I thought that the questions listed on page 34 of the Framework were pretty comprehensive. However are always the basics like:
- At what level are we going to report an item cluster? Is each item getting reported separately and then also as a cluster? Is the PE reported separately or as a bundle if the stimulus is bundled? If each item is reported separately for its level of content (dimensional learning), is the cluster then reported differently and separately on its own as a PE level measure?
- How many ITEMS are needed to have a large enough sampling of student work to validate a score? How many item CLUSTERS are needed to have a large enough sampling to validate a PE level score?
- How are items delivered (fixed form, Computer adaptive (I REALLY DON'T RECOMMEND THIS FOR A TESTLET FORMAT), modified adaptive—this is a newer format that does not use randomization algorithms but basically creates a myriad of scrambled forms based off of a base skeleton
- The response to #10 is a starting point but in addition, the blueprints will need to address
- What will be measured – the grain size and extent with each.
- Is there a need to define "possible bundles for assessment purposes?
- The appropriate use of various item types.
- Balance between DCIs, CCCs, and SEP both in terms of "weight" and proportion of items
- Assessment limits including language and mathematics limits if appropriate. (Will this need definition across a bundle as well as DCI/CCC/SEP?)
- Various possibilities for reporting performance…..
- Performance levels
- Reporting categories
- Any distinguishing between student, class, school, system level reporting
- It may useful to include a checklist of tasks that an individual state would need to accomplish (e.g., PE Bundles, etc).
- Blueprint must reflect what is housed in the Assessment Framework.
- Since the framework is so extensive, the blueprint needs to define what shall be measured at the state level and what is expected of schools to instruct and measure at the school level.
- The blueprint may/will differ by state, therefore the blueprint becomes a suggested model.

- The blueprint must specify grade level expectations.
- What is the purpose of the assessment?
- What will be reported out on and/or what will the reporting categories be?
- What type of questions will be used?
- Questions:
    - What issues need to be considered at the HS level for blueprints in states using the assessments as a graduation requirement?
    - Are the items truly assessing the three dimensions of NGSS?
    - What is the purpose of the assessment (what is the assessment measuring): student progress, teacher effectiveness, instructional shifts, etc.?
    - What are we going to measure with timing constraints?
    - Is this assessment going to be an End of Course assessment and/or a summative assessment?
    - Will the high school assessment be of a different test type than elementary and middle school assessments?
    - What are we measuring?
    - How do we communicate this so that the standards and not blueprint is taught?
- The resulting matrix that we could make available to the field could play an important role in this piece. The information will always show the range of information that will be assessment knowing there are a myriad of combinations that do not make it worth it to just focus on the content. This way, teachers will teach to the breadth and depth of the standards and not try to game the assessment.
- How many item clusters?
- How much time should students be expected to spend (range) on a given item cluster?
- Are there any of the NGSS PEs that shouldn't be assessed on a summative ESEA assessment?
- How do we avoid the PEs from being unpacked?
- Score Reporting categories? (different models)
- What is the purpose of blueprint design?
- What are we going to measure? Breadth vs. time (measure subset of what is defined)
- H.S. problems –how do we measure at H.S. e.g. diff. states have diff. graduation requirements
- Geared towards end of course vs. summative
- Different design questions for different levels
- Integrated summative science (elementary and middle)
- Blueprint design considerations for various reporting strand configurations
- Weighting considerations DCI (LS, MS, ESS, ETS) vs SEP vs CCC
- Test length: Number of questions/clusters
- Prioritization of PEs
- Role (if any) of standalone items
- Design considerations for various grade level tests
- Content sampling plan considerations
- Separate or Embedded Performance task(s)

Question 12. How should college and career readiness be addressed in assessment development and in the Assessment Framework?

- To state in a test that if student performs at a proficient level that he/she is college and career ready can include many factors, and some might think we should be cautious. We want to be encouraging students to pursue a career in the sciences.
- As rigorous as these standards are, we should be comfortable stating that with a proficient result, our students would be college and career ready. Perhaps stating that a proficient result on this assessment is "yet another indicator of college and career readiness" would satisfy all.
- This is already addressed in the NGSS. College and Career readiness may be utilized in the development of the Achievement Level Descriptors.
- Unless we have a desire for a high school assessment to be considered for college credit or for college acceptance of science "readiness" without question at a certain score—then I'm not sure what I would put here. It would be a generic—STEM is an important factor in any child's readiness for college and/or career ….etc etc etc….
- In theory, the NGSS are college and career ready standards so an NGSS-aligned assessment that students would do well on would mean that they are prepared to move on beyond high school.
- I think we grapple with this on a bigger scale, since some of what we place value on in K–12 is not valued in higher ed and vice versa. There is a practical gap between K–12 and higher education in science education. However, the Framework should provide a definition of what college and career readiness looks like according to the K–12 Framework and NGSS. It should also be honest enough to say that an NGSS-based assessment is not the same as, for example, an AP exam or a college placement test. A student who excels in high school science aligned to NGSS will still have room to learn and grow in college, but will have a solid foundation.
- We should be very purposeful in stating claims that may be made associated with the assessment. More research is needed on this topic.
- The expectation today at the state and federal level is for state assessments to measure student performance so that a determination is made on the student's college and career readiness. The student results are then rolled up to provide a school, district and state status.
- Give some examples of what a CCR achievement could be. Maybe it is a larger part of the assessment system—not just the state summative part.
- Further research needs to be done as to what College and Career Ready actually means. This warrants further discussion.
- Need continued research in this area
- Much like College and Career Readiness (C&C) is embedded in our current standards, C&C should be embedded in the framework. How this is done will take some research considering there are different interpretations of C&C. The way C&C is addressed should be general enough so all states can use this at the state and district level.
- More research needs to be completed (debunking myths) of what metrics are needed in order to determine whether someone is considered college and career ready in science (college credit bearing courses). I am very concerned that a one-time snap shot of a student can determine which level of college science courses a student should be able to enroll in at either a community/trade school or 4-year university.
- Level 3 &4 – on track for being prepared, not sure how that applies for NGSS esp. with matrix sampling

- Appendix D – all standards, all students
- various levels of college level entry chemistry
- 4 yr. vs. community college
- Basic, proficient to entrance level chem.
- What does the research say on CCR for Science?
- CCR should be clearly defined – what does college and career ready mean in terms of science/NGSS?
- Any evidence (existing or to be collected) that the standards lead to college/career readiness should be included.

Question 13. How should the Assessment Framework address alternate assessments and alternative assessment forms?

- Maintaining a portfolio is what we recommend as it shows growth. There will be a need for an alternative set of standards that are based on the NGSS.
- First, clarify what is meant by this. Alternate assessment may be mentioned and that further work needs to be done, but this seems like a separate project.
- You know, I'm not 100% on this. In our state we have an actual separate test for our severely limited students. The Alt assessment is, in some ways MORE task based than the regular because students usually have to respond using cards or by pointing or interacting with materials. We have standards extensions that are used for these students as well—standards that stretch content (not practices or cross cutting concepts) to some of the most basic constituent pieces. I'm not sure where our department is going with that in terms of NGSS at this point. I think that the special education assessment group at SCASS is VERY active and would be a good resource to communicate for what should be or would be appropriately included in this section.
- The K–12 Framework places heavy emphasis on science for ALL students. This should be reflected in the assessment system. The Assessment Framework will need to acknowledge the need for assessments that meet the needs for all students (ELLs/cognitively challenged/physically challenged etc.) and should recommend potential ways of meeting the needs of these students as well as cite examples of assessments/models/research-based tools that have effectively done this.
- The Framework is not the place for an extensive discussion on Alternate Assessment but it cannot be ignored.
- A discussion considering pros and cons of different possibilities would be beneficial. Experts in this area should be consulted concerning the development of alternate or alternative assessments.
- The Assessment Framework needs to address alternate assessments and alternate assessment forms. The approach the Kentucky Department of Education is using is to prioritize a set of progressions from the NGSS to roll out to teachers with students participating in Kentucky's Alternate Assessment. Assessment items are written to the prioritized standards using a familiar assessment format identified as Attainment Tasks. Attainment tasks are picture based performance events that require students to complete a task, working step by step as directed by the teacher. The items in the task are multiple choice questions made up of 3

distractors and 1 correct response. Attainment tasks use activities that is based on an authentic task (e.g., similar to a task that might occur in real life) with a focus on the three dimensions of the NGSS; emphasizing application and analysis. The items are structured around classroom application and student participation.
- Give examples or suggestions for how to address these. For example, perhaps states should create a guide to help instruct how students who are below grade level should be assessed.
- Provide guidance for aligned items and examples of how alt assessments could be developed to meet various needs.
- Describe that it will be planned for but not spelled out in Framework.
- Key Piece: What are states doing now that can be carried out in the NGSS assessments?
- Text to speech and speech to text issues have to be addressed. Access and equity issues for underrepresented and special needs groups is an integral part of any state assessment but there needs to be a flexibility so that states can address issues unique to their state.
- General guidelines around how to develop alternate assessments and or alternative assessment forms (Braille, Translations, paper/pencil vs. online) should be both flexible, but give enough information to help guide SEAs to starting thinking about how to develop these two types of assessments that are linked or aligned to NGSS.
- The Assessment Framework should not address alternate assessments, but should address alternative assessment forms such as braille and other languages
- Any current recommendations, supporting research, or possible existing models should be included. Don't try to cover everything in this round of the document.

Question 14. To what degree should the Assessment Framework address accommodations and allowable assessment features?

- I think looking at the other consortia accommodations and features would be good but we have to see these through a science specific focus.
- As much as possible we have to provide accessibility for those populations.
- Offering different form with fewer interaction types, "drag and drop" etc. Choose from pick 1, pick 2, pick all that apply, or write an essay.
- These issues should be addressed and the most up-to-date thinking and research in the field should be referenced. In our state, we will likely follow the same guidance as that used for Smarter Balanced.
- The NGSS are standards for ALL students. This includes students that run the gambit from mild learning disabilities to gifted LD to severely limited or exceptionally gifted. I believe that if our standards are for all students, then our assessments are for ALL students, and that means that we need to address accommodations and allowable features to the extent that any of us would in a generic RFP. Beyond that, each state can add their individual requirements/expectations as desired.
- The K–12 Framework places heavy emphasis on science for ALL students. This should be reflected in the assessment system. See #13.
- Each state / groups of states collaboratively will need to determine the specifics of how this is addressed but a general discussion of accommodations and supports with citations should suffice.

- The framework should include a discussion on these topics but may need to be a more general conversation until specific items/tasks/clusters are developed.
- The Assessment Framework needs to address assessment designs for on-line and paper and pencil. The document should state which item types are appropriate for each assessment model. Since this activity is for item development only, comparability between on-line and paper and pencil assessments is not a consideration.
- Now to the specific question of accommodations. With 12-13% of our state's students receiving an accommodation, it is critical that the document provide guidance on accommodated materials like Braille, audio, large print, assistive technology etc. Certain item types may not work in online, paper and pencil or both. The document should spell out what is viable and what is not.
- Suggested accommodation and features should be included, maybe as a list.
- The Framework needs to include guidance such as the parameters in other major assessments (PARCC, Smarter Balance, NAEP, etc) and the recommendations from NBS.
- The Framework already covers some of this, may need to add specifics regarding item types that are specific to science.
- Must also address translations to other languages.
- NAGB and the Consortia have done some work around this issue. Language issues have to be addressed re: multiple language accessibility without being prescriptive. These accommodations and features need to be general enough so that states can address issues unique to their state.
- General guidelines (similar to what SBAC and PARCC have developed) are helpful. For us, being able to streamline our science assessment to have similar accessibility supports language and functionality not only keeps the assessment systems consistent for students (prevents confusion), but also helps prevent additional confusion to the field in regard to what an accommodation, designated support or universal tool is (i.e. SBAC terminology).
- There should be a strong focus on accessibility features in the assessment framework. A list should be included in the framework.
- Both major consortia have done significant work in this area. We should utilize and refer to the work that has been done. If there are recommendations unique to science those should be included.

Question 15. How should the Assessment Framework address support for English language learners?

- Read to - certified administrators would be a problem for our schools
- Have the test computer presented in language needed.
- This brings up so many questions:
- Which languages? Instructions only? All questions?
- Read aloud should be in English only.
- When do we test a new ELL student?
- Is this really a framework issue?
- The latest research and thinking should be referenced. In our state, we will likely follow the same guidance as that used for Smarter Balanced.

- The K–12 Framework places heavy emphasis on science for ALL students. This should be reflected in the assessment system.
- This needs to be addressed, probably more extensively, but in a way similar to #13 and #14. Smarter Balanced, the CCSS, and WIDA all have information which could contribute here.
- BTW: Somewhere in the Framework we should clearly state the mathematical and overall language level, and vocabulary expectations
- The framework should include a discussion on this topic but provide a general conversation of possible options rather than a prescription.
- Guidance is needed, especially in the area of how soon should an English language learner be assessed in the new Assessment Framework. The document needs to address and may site references to research in this area to do's and don'ts.
- Support ELLs. At a minimum, it is important to include to show that it is something states need to take into consideration when developing a test.
- The Framework must address concerns regarding an assessment accessible for ELL, such as compute translated assessments, "read to" procedures, and which languages are needed.
- Use guidance from English Learner community;
- Each state will implement consistent practices already in place for other state assessments;
- Possibilities: Computer translated, reading the test to students in their native language, whatever accommodations and features are made need to be general enough so that states can address issues unique to their state.
- General guidelines- similar to SBAC. It would be helpful to have guidance around cognitive load (NGSS) and ELLs. It would be ideal of course to have similar translations and pop up glossaries be available for students on an NGSS aligned assessment that are currently available on the SBAC math assessments. General guidelines around Text to Speech, Translations, etc… are very helpful.
- Should include language supports for ELLs.
- Discuss Universal design considerations
- Discuss ELL testing options that might be preferable for science tests —stacked language (SBAC Math), glossed words, translations

Item Specifications Guidelines Feedback

Please provide feedback, questions, or affirmations related to the item specifications guidelines documentation.

- I appreciate that the document is soundly based on research, and anticipate that we will continue to keep that research at the forefront. The thoughts and ideas as a result of this collaboration are reflected in the document.
- How will possible claims and targets be addressed in the item specifications?
- Pg 9. – Use more recent references to Universal Test Design elements (current citation is from 2002).
- More extensive information about the potential use of automated scoring should be included (e.g., what types of items it could be used for).
- Item Specification Tables:
    - Available items types should be more specific (i.e., specify the types of TEI items are most appropriate for a given PE or bundle of PEs).
    - Allowable stimulus materials should be more specific (e.g., instead of just "graph" or "table", specify the types of graphs or tables that could be used, specify types of simulations or animations that could be used. This section is currently much too generic.
    - Sample item stems with each PE (or bundle) would be very helpful, especially to provide guidance as to how questions can be worded to elicit evidence of student understandings of the practices and crosscutting concepts (combined with DCIs).
    - Pg. 16 shows an example of how a SEP is combined with a CCC, but without a DCI – this doesn't seem appropriate or even possible in the context of NGSS.
    - Clarify what it means to have a DCI "in common" between two PEs (e.g., PS1 or PS1.A?). Must bundled PEs have DCI "in common" or can they just be related?
- Questions:
    - What aspects of the item specifications will be developed during Phase 1?
    - Need greater clarity as to what aspects of the item specifications will items be written to (e.g., the evidence statements?, all aspects including PEs, CCCs, SEPs, DCIs? This is especially unclear when PEs are bundled.
    - Will an item specification table be created for each Performance Expectation and then additional ones for the bundles? Having both would be the very helpful.
    - Will multiple versions of PE bundling be developed?
    - What is the time estimate for an item cluster? Seems it would be different for pilot/field tests and the live test forms (i.e., more items would need to be pilot/field tested than would be used on the live tests).
- An excellent start. It will be exciting to see the next rendition.
- Consider the possibility of having at least one pair of "contrasting" examples where the DCI, CCC, and SEPs at a particular grade level are bundled and assessed in two different arrangements.
- Need more psychometric information in here. Would be helpful to know where psychometric team agreed, and where there was disagreement.
- Maybe put something in there about alignment to Common Core. For example, graphing is big in NGSS, but only bar graphs and pictographs are expected in grade 5. It might be important to make sure states are within their "limits" and not ask for line graph interpretations at grade 5 and below.

- Cognitive levels are addressed well in the front matter, but then not included in the prototypes/models. Was this intentional?
- Chapter 2: It may be helpful to include information about "below grade level" achievement (i.e., Alt). Our state develops a "resource guide" for SPED educators to help them understand how these students can show progress.
- Maybe have groups of teachers (at appropriate grade levels) determine the "vocabulary requiring definition" words. Contractors and a few state dept. of ed staff may not always know what should be included and what should not be included.
- p. 34, 2nd bullet, this was in the Framework as well and needs clarification; it doesn't make sense because you will need to field test more items than you will need in a cluster, so they cannot all be inter-related; also there are psychometric considerations that should be spelled out if you want to make items that are inter-related.
- p. 38, 1st paragraph under Item Cluster Stimuli: this seems to conflict with earlier language about putting pertinent information next to the item instead of within the stimulus. This paragraph should be clarified.
- Really liked Appendix B, nice summary
- Editorial: Page 6, Under Cog Complex: The NGSS places (not place)
- There may be issues with developing questions if the evidence statements that focus only on a single dimension are used to guide item development.
- We appreciate this platform for feedback since this has provided us feedback and guidance in our work moving forward.
- Style Guide Specifications need to be included (similar to SBAC and PARCC)- so that the same font, graphics guidelines, stimuli guidelines, units (ft. vs. ft), etc… needs to be included (especially for Phase II to be successful).
- The prototype items are in development currently, but sample student work with sample rubrics would also be incredibly useful also.
- General guidelines around the psychometrics for 2-dimensional and 3-dimensional aligned items- how to determine which dimension the student is proficient in if they don't get all of the item correct (i.e. student scores 1 point out of a 2 point item)?
- Basic Terminology, p. 4:
    - Add definitions for "dimension," "domain," "item alignment", "item cluster specification," "item cluster alignment", and "target".
    - Alphabetize the list of terms.

      DOK, p. 8:
    - Can a multi-part item have an overall DOK>2 rather than requiring "each" part to have a DOK>2? Fewer constraints may lead to more flexibility in scaffolding within a cluster allowing for a richer set of items.

      Universal Design/Vocabulary and Language, p. 9:
    - Would like to see a stronger recommendation for keeping the vocabulary at a grade-appropriate level. The PEs were written for an adult audience. The use of a scientific term in a PE does not seem to be an appropriate criterion alone for using that term in an item.
- Chapter Three: Item Cluster Specifications, p. 12:
    - Is an item cluster specification the same thing as an item cluster alignment?

- Chapter Three: Item Cluster Specifications, p. 13:
    - Change number of items in the "Number of Items in IC" to a range,
    - How do the number of items needed for a CAT factor into the number of items or range of items listed in the item specification?
- Chapter Three: Item Cluster Specifications, p. 14-17:
    - Please provide guidance on how the Evidence of Learning statements were sorted among the various tables. How to decide which belong on DCI and SEP, on DCI and CCC, etc…? If states are going to be making those decisions, we may want to consider ways to be consistent. Word documents of the Evidence Statements and the NGSS would be really helpful!
- Chapter Three: Item Cluster Specifications, p. 20:
    - The placement of the 3-D CR item in a cluster should be flexible—it doesn't have to be the last item or the first item
- Chapter Three: Item Cluster Specifications, p. 22:
    - The assessment boundary in the table for 5-PS1-2 is incorrect, it's a repeat of the Clarification just above it.
- Appendix A, p. 34:
    - In the second bullet, "Each item is inextricably linked to the stimulus and to the other items…the cluster of items must be constructed in such a way that individual performance on each item is adversely affected if an item is responded to without the context of the other items in the cluster." This language raised concerns for our NTAC. Examples of concerns: a cluster of items that are interrelated risks becoming essentially a single item; if an item or two in an interdependent set is unsuccessful in field testing/piloting risks losing the entire cluster. It would be helpful if these issues could be addressed in the document.
    - p. 44, The multiple choice example on this page indicates "While the MC item itself does not directly align to an SEP, the combination of the simulation and the follow-up question asked does…" This statement seems to imply that an item aligning to 2 DCIs or 2 SEPs is necessarily a 2D item. Is this the intent? Do you mean to imply that we should not be critical of an individual item that seems to fall short of being at least 2D or DOK>2 if that item is needed for the scaffolding of the cluster?
    - Once good prototypes exist, consider replacing the NAEP examples with examples that are aligned to NGSS and illustrate appropriate dimensionality and clustering.

APPENDIX E: Annotated Resources for the Development of Assessments for the NGSS

A growing number of papers and reports on various issues related to Next Generation Science Standards (NGSS) assessment have established a need for a central online repository for these documents. As a result, in January 2015, the Center on Standards & Assessment Implementation (CSAI) made available just such a collection. The CSAI collection can be accessed at http://csai-online.org/collection/1565. Its overarching mission is "to support districts and states as they endeavor to design and implement NGSS-aligned assessments—from a single assessment component to a full end-to-end assessment system—to be consistent with the vision and goals of the NGSS."

In this appendix, a subset of the CSAI collection is summarized and briefly annotated. The purpose is to give a thumbnail sketch of the contents of the collection; it assumes that those seeking more information will directly access the primary-source documents through the CSAI site. It is also assumed that those using the resources housed on this site will use their professional discretion when utilizing the information included—not all resources were created with the specific purpose of informing NGSS-aligned assessment, and, thus, these resources should be used with the appropriate lens. The collection is divided into four categories, as detailed below.

Assessment of the NGSS: An Overview

This section includes some of the seminal NGSS documents, including the NGSS themselves.

A Framework for K-12 Science Education: Practices, Crosscutting Concepts, and Core Ideas
http://csai-online.org/resources/framework-k-12-science-education-practices-crosscutting-concepts-and-core-ideas

Developed by a committee of nationally and internationally known experts in the fields of science and science education, this document establishes a solid, evidence-based foundation on which the NGSS were formed.

Next Generation Science Standards
http://csai-online.org/resource/158

This website is easily navigated and comprehensive in information related to the development process for, explanation of, and access to the NGSS. This resource is relevant, timely, and complete in its presentation of the NGSS. Additional links to supplemental resources are updated regularly.

Developing Assessments for the Next Generation Science Standards (BOTA report)
http://csai-online.org/resources/developing-assessments-next-generation-science-standards

This document lays out much of the underlying considerations of NGSS assessment and was referred to frequently in the development of the SAIC Assessment Framework.

The Assessment Continuum

Reports and other documents on the full spectrum of NGSS assessment—formative, interim, and summative—are included in this section. The reports focusing strictly on summative assessment are as follows:

What Can Be Learned From Current Large-Scale Assessment Programs to Inform Assessment of the Next Generation Science Standards?
http://csai-online.org/resources/what-can-be-learned-current-large-scale-assessment-programs-inform-assessment-next

This paper provides an in-depth examination of current large-scale assessment programs, with an emphasis on supporting effective assessment of science and engineering practices (SEPs). The analysis of multiple item and task samples and the presentation of specific alignment data make this a valuable source of information for assessment developers, policymakers, and educators.

NGSS Evidence Statements for High School
http://csai-online.org/resources/ngss-evidence-statements-high-school

The high school evidence statements describe how students can use the SEPs, crosscutting concepts, and disciplinary core ideas (three dimensions) to demonstrate proficiency for the performance expectations. They are intended to identify clear, measurable components that can be used when designing assessments for the NGSS.
The evidence statements for grades K–8 were released in June 2015 and are available at http://www.nextgenscience.org/k-5-evidence-statement and http://www.nextgenscience.org/middle-school-evidence-statements.

System Design and Implementation

The conceptual and logistical challenges in developing NGSS assessments are the focus of this section. Among the broad selection of resources included here are an analysis of high-level designs for a system of assessments to support the NGSS; a Venn diagram showing the relationships and convergences among the Common Core State Standards (CCSS) practices for mathematics and English language arts and the NGSS SEPs; and several sample science items and tasks that were developed prior to the creation of the NGSS and that illustrate some of the aspects of assessing student learning of the NGSS.

Assessment System Design Options for the Next Generation Science Standards (NGSS): Reflections on Some Possible Design Approaches
http://csai-online.org/resources/assessment-system-design-options-next-generation-science-standards-ngss-reflections-some

This paper explores how a system of assessments can be designed to allow students to demonstrate their proficiency with the NGSS performance expectations, providing three possible high-level designs of the system and one less-optimal design for contrast. The author incorporates a discussion of the National Research Council's (NRC's) assessment triangle and the application of evidence-centered design principles into her analyses.

Science Assessments: Innovations in the Next Generation of State Assessments
http://csai-online.org/resources/science-assessments-innovations-next-generation-state-assessments

This site contains a video recording of a November 2013 webinar hosted by the Alliance for Excellent Education. Panelists discuss activities in states that have adopted the NGSS; outline options for assessment; and describe efforts underway in Kentucky. The webinar featured the following panelists: Nancy A. Doorey, Director of Programs, K–12 Center (ETS); Karen Kidwell, Director, Office of Program Standards, Kentucky Department of Education; Stephen L. Pruitt, PhD, Senior Vice President, Achieve, Inc.; Robert Rothman, Senior Fellow, Alliance for Excellent Education; and Kathleen Scalise, PhD, Associate Professor of Education, University of Oregon.

Relationships and Convergences Among the Mathematics, Science, and ELA Practices
http://csai-online.org/resources/relationships-and-convergences-among-mathematics-science-and-ela-practices

Developed by the Stanford University Understanding Language initiative, this Venn diagram is a tool that effectively illustrates the overlaps and groupings of student practices (and student capacities) among four sets of seminal documents: the CCSS in English language arts and literacy in history/social studies, science, and technical subjects; the CCSS in mathematics; A Framework for K-12 Science Education: Practices, Crosscutting Concepts, and Core Ideas; and the Framework for English Language Proficiency Development (ELPD) Standards corresponding to the CCSS and the NGSS.

Council of State Science Supervisors: Large-Scale Assessment Discussions
http://csai-online.org/resources/council-state-science-supervisors-large-scale-assessment-discussions

The Council of State Science Supervisors Science Education Assessment page offers examples of states' approaches to NGSS assessment implementation and provides insights into the process from states' perspectives. Both PowerPoint slides and recorded webinars are provided for presentations, which allows users to benefit from the discussions that ensued during the original presentations.

Sample Science Items and Tasks

Within this subsection are several papers providing examples of extant science assessment tasks that both follow a principled and research-based approach to assessment and demonstrate some of the characteristics of assessing student learning that are called for in the NGSS.

How Can Simulations Be Components of Balanced State Science Assessment Systems?
http://csai-online.org/resources/how-can-simulations-be-components-balanced-state-science-assessment-systems

This 2011 policy brief offers guidance and recommendations on how simulation-based science assessments can be included in the next generation of state science assessment systems. The authors report on research conducted on the simulation-based science assessment module SimScientists. The brief does not address the NGSS directly, as it was published before the release of the standards, but it is applicable in that it develops assessment targets that reflect research on model-based learning based on the NRC's Framework for K-12 Science Education.

Presentation of National Assessment of Educational Progress Sample Science Tasks
http://csai-online.org/resources/presentation-national-assessment-educational-progress-naep-sample-science-tasks

This site contains slides from a presentation by Peggy Carr of the National Center for Education Statistics, including samples of innovative item types and innovative evidence types that demonstrate how hard-to-measure constructs, similar to what are called for in the NGSS, were addressed in the 2009 NAEP Science Assessment.

Presentation of a Program for International Student Assessment (PISA) Science 2015 Task Variant
http://csai-online.org/resources/presentation-program-international-student-assessment-pisa-science-2015-task-variant

These slides, from a presentation given by Janet Koster van Groos and Eric Steinhauer of ETS, include examples of a Program for International Student Assessment (PISA) Science 2015 Task Variant, which illustrate how scientific literacy is assessed by requiring the use of both scientific competencies and content knowledge, similar to the multidimensional science learning called for in the NGSS.

Presentation of Advanced Placement Science Task
http://csai-online.org/resources/presentation-advanced-placement-science-task

These slides are from a presentation on the redesign of the Advanced Placement Physics course and exam, given by Karne Lionberger of College Board. They illustrate how the new curriculum emphasizes the pairing of science content with practices, which is akin to the multidimensional science learning called for in the NGSS.

Challenges, Implications, and Opportunities

This section includes several reports on the tradeoffs and challenges that should be considered when designing a system of NGSS-aligned assessments, including task design and scoring, psychometric modeling, and practical and logistical considerations. The included reports focus not only on summative assessments, but also on issues related to formative assessments and program (curriculum intervention) assessments.

Designing NGSS Assessments to Evaluate the Efficacy of Curriculum Interventions
http://csai-online.org/resources/designing-ngss-assessments-evaluate-efficacy-curriculum-interventions

In this paper, the authors present an approach to designing assessments for the intended purpose of evaluating curriculum interventions for NGSS implementation in schools. The use of evidence-centered design is discussed and illustrated in an example from an urban middle school in which assessment is used to evaluate science curriculum materials.

Proficiency in Science: Assessment Challenges and Opportunities
http://csai-online.org/resources/proficiency-science-assessment-challenges-and-opportunities

This paper provides a concise summary of the new perspective on science proficiency that is presented in the NRC's Framework for K-12 Science Education and the NGSS, and provides a high-level overview of its implications for assessment. Although the paper does not provide the level of detail provided in other resources released more recently, it does condense the important facts and challenges into a helpful summary.

Inherent Measurement Challenges in the Next Generation Science Standards for Both Formative and Summative Assessment
http://csai-online.org/resources/inherent-measurement-challenges-next-generation-science-standards-both-formative-and

The authors discuss the implications of the complexity of the NGSS for assessment, particularly for summative assessment for individual high-stakes accountability reporting. They provide important insights and guidance on how to address these issues, as well as several strategies and methods for mitigating the complexities. Note that additional resources have been released since the date of this resource's publication, including NGSS evidence statements that further support assessment of the standards; thus, a discussion of how these resources should be considered in assessment design is not included in the paper.

APPENDIX F: Examples from SimScientists and NAEP

Examples of simulation-based modules

Figures F-1 through F-3 are examples of simulation-based modules from SimScientists that interweave information about ecosystems with active student practices, such as developing and using models, planning and conducting investigations, and interpreting patterns in data sets.

Students can play, pause, and scrub the animation shown in Figure F-1 to make careful observations of the feeding relationships in the ecosystem, which are the interactions between system components.

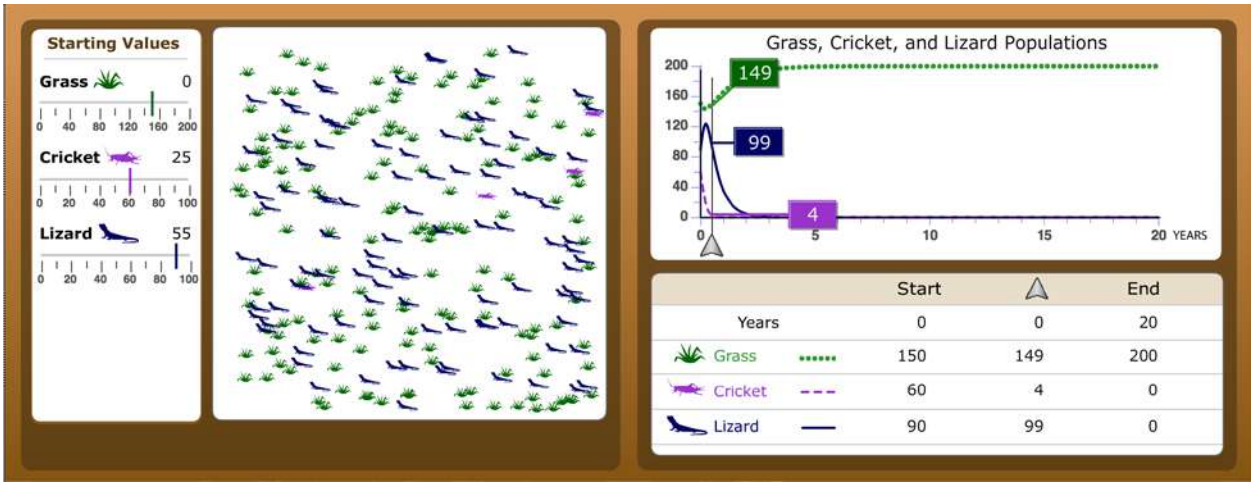
Figure F-1. Students observe a simulated ecosystem.

In the model shown in Figure F-2, the arrow represents the flow of matter and energy from the cricket to the lizard.

Figure F-2. Students use their observations to develop a model that represents the ecosystem interactions.

In the model shown in Figure F-3, students can set the starting values for the organisms, run trials, and collect data on the population changes, which are the emergent phenomena. Students are expected to interpret these data and to construct explanations of the population changes, based on their mental models of the system: population changes are the result of interactions (feeding relationships and the flow of matter and energy) among system components (the organisms and abiotic factors in the ecosystem).

Figure F-3. Students manipulate a simulated ecosystem.



Examples of new task formats

Figures F-4 through F-6 are examples of new task formats developed for the National Assessment of Educational Progress (NAEP).

In the task shown in Figure F-4, students can set the starting values for the mass and temperature of each substance, run trials, and collect data on the temperature changes.

Figure F-4. Students conduct an investigation of energy transfer.

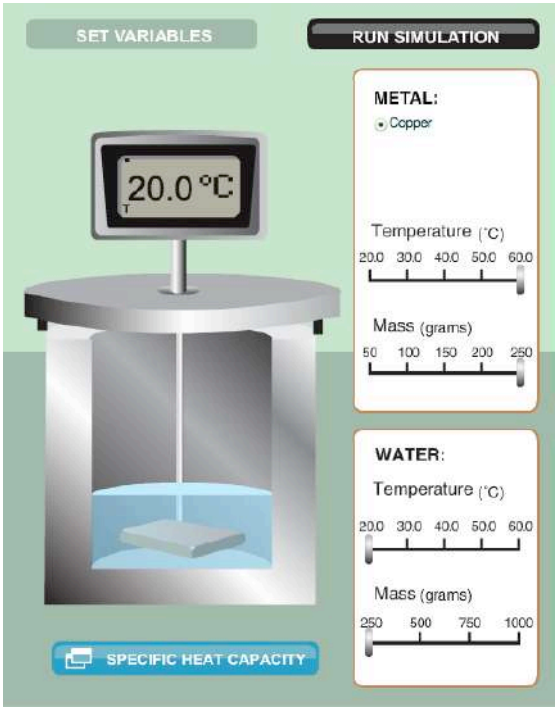
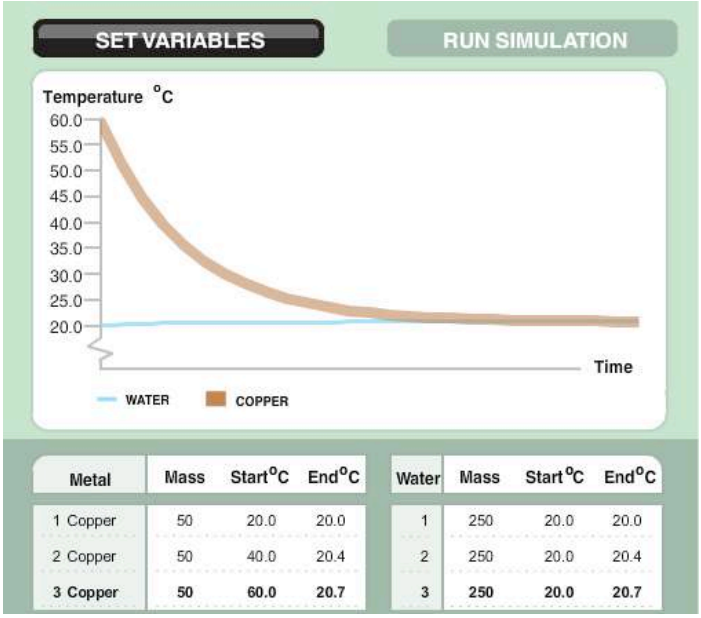


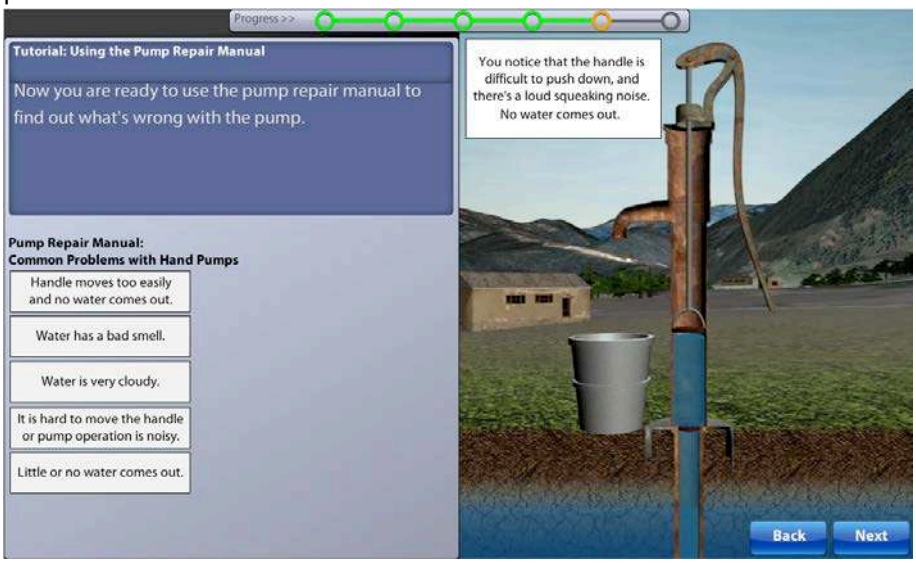
Figure F-5. Students use their data to explain changes in temperature based on their mental models of the system of interacting molecules and principles of energy transfer.



Metal	Mass	Start °C	End °C
1 Copper	50	20.0	20.0
2 Copper	50	40.0	20.4
3 Copper	50	60.0	20.7

Water	Mass	Start °C	End °C
1	250	20.0	20.0
2	250	20.0	20.4
3	250	20.0	20.7

Figure F-6. Students troubleshoot a broken pump by systematically investigating possible sources of the problem.



APPENDIX G: Documentation and Timeline of Key Activities

Documentation of Decisions

Development of a research-supported, technically sound, state-endorsed item pool aligned to the NGSS will entail detailed documentation of options considered, tradeoffs weighed, and decisions reached during all phases of work. These types of information promote trust with stakeholders, as steps taken are transparent and evidence is on hand to defend decisions related to the next-generation assessment.

Guiding questions to consider during the documentation process include the following:

- How and by whom will decisions about assessable PEs, measurement models, and item types be documented?
- Who will review and provide final approval for the specifications and blueprints? How will subsequent changes be recorded?
- How and by whom will all steps in the item development process be documented?
- How will decisions based on reviews, interviews, pilot testing, and field testing be documented, recommendations be implemented, and impact be monitored over time?

It is important to note that the members of the SAIC represent a diverse group of states and entities. Some are members of one of the two major Race to the Top assessment consortia (the Smarter Balanced Assessment Consortium [Smarter Balanced] and the Partnership for Assessment of Readiness for College and Careers [PARCC]); others do not participate in either consortium. Some members adopted the NGSS by name; others adopted the full NGSS but with a name change; and still others opted for a partial adoption. Some members plan for a full computer-based administration of NGSS assessments, while others plan to use a mix of computer-based delivery and paper-and-pencil delivery. Finally, some members fully embrace the recommendations in the BOTA report, while other members support a more state-mediated approach to the transition to NGSS.

Table G-1 below outlines the major activities and processes for the development of the Assessment Framework.

Table G-6. Timeline for Key Activities and Deliverables for the Assessment Framework

Activity Date	***Type of Activity***	***Purpose of Activity***
Bi-weekly	SAIC Collaborative Check-in	Project status update
Bi-weekly	CCSSO Conference Call	Project status update
February 13, 2015	Kickoff with member states/entities - WebEx	Orientation
February 13, 2015	Online survey	Assessment Framework State Survey #1 administered

Table G-7. (continued)

Activity Date	*Type of Activity*	*Purpose of Activity*
February 23, 2015	In-person meeting with all collaborative members, Austin, TX	Assessment Framework outline/overview
February 23, 2015	Online survey	Assessment Framework State Survey #2 administered
April 3, 2015	WebEx with the Collaborative	Project Update
April 20, 2015	Draft Framework made available	
April 20, 2015	Review of Assessment Framework by Psychometric Review Committee	Individual reviews by CCSSO and content experts
April 24, 2015	WebEx with the Collaborative	Reporting options
May 14, 2015	In-person Collaborative Meeting, Chicago, IL	CCSSO and collaborative review of draft Assessment Framework and draft item specifications
May 15, 2015	In person Prototype Sub-Committee meeting, Chicago, IL	Discussion of draft prototypes
June 12, 2015	WebEx with Technical Team	Content Expert and Technical team discussion of Assessment Framework
June 12, 2015	WebEx with Psychometric Review Committee	Discuss feedback and recommendations
June 22, 2015	Prototype Draft Preview Materials sent to Prototype Sub-Committee	Grade 5 prototypes sent for preview and feedback
June 24, 2015	Prototype Draft Preview Materials sent to Prototype Sub-Committee	High School prototypes sent for preview and feedback
July 1, 2015	WebEx with Prototype Sub-Committee	Discussion of draft prototypes
August 7, 2015	Member review of final Assessment Framework	Individual reviews by members
September 11, 2015	Delivery of final Assessment Framework to CCSSO	Final delivery made by WestEd
	Delivery of final Prototypes to CCSSO	Final delivery made by WestEd