Evaluating Item Quality in Large-Scale Assessments

Phase I Report of the Study of State Assessment Systems

SCALE
Stanford Center for Assessment, Learning, & Equity

Understanding Language

Stanford Graduate School of
EDUCATION

A project of

Understanding Language/ Stanford Center for Assessment, Learning, & Equity (UL /SCALE)

June 2016

Ruth Chung Wei
Director of Assessment Research & Development

Vinci Daro
Director of Mathematics Learning

Nicole Holthuis
Science Research & Development Associate

Kari Kokka
Mathematics Research & Development Associate

Daisy Martin
Director of History/Social Studies Learning

Nicole Renner
Director of English Language Arts Learning

Susan Schultz
Director of Science Learning

Jill Wertheim
Science Research & Development Associate

with the Stanford NGSS Assessment Project Team
(Jonathan Osborne, Raymond Pecheone, Helen Quinn, Paolo Martin)

Table of Contents

About UL / SCALE at Stanford University

Understanding Language/Stanford Center for Assessment, Learning, and Equity (UL/SCALE) is a recently merged research and practice center based at Stanford University that focuses on both language and performance assessment in K-16 settings. The mission of UL/SCALE is to support educators and policymakers in transforming systems to advance equity and learning for students – particularly for English Language Learners (ELLs) – by illuminating the symbiotic ways students learn language and academic content, and through the development and use of curriculum-embedded performance assessments. UL/SCALE provides technical expertise and support to schools, districts, and states that have committed to adopting performance-based assessment as part of a multiple-measures system of assessments to evaluate student learning and school effectiveness. UL/SCALE works with education agencies and practitioners to develop customized assessment, curriculum, and instructional materials, provide professional development for educators to address and conduct evaluations of the validity, reliability, and efficacy of the system. UL/SCALE has a deep commitment to equity and is a national leader in adapting curriculum, instruction, and assessment to address the language development/acquisition needs of English Language Learners and all students. Led by Stanford University Professors Kenji Hakuta and Ray Pecheone, UL/SCALE brings together state-of-the-art knowledge and tools regarding academic language and performance assessment, and applies a set of research-based guiding principles called "Learning Centered Design" to all design work. Find out more about UL/SCALE at: http://ell.stanford.edu and http://scale.stanford.edu.

Suggested citation: *Understanding Language/Stanford Center for Assessment, Learning, & Equity (2016, June). Evaluating Item Quality in Large-Scale Assessments, Phase I Report of the Study of State Assessment Systems. Stanford, CA: Author.*

Acknowledgements: *Special thanks to Gerald "G" Reyes and Claudia A. Long for editorial support, and to Pai-rou Chen for research assistance.*

Executive Summary

A new direction for large-scale assessment?

In the next few months and years, state legislators, education commissioners, and education testing directors (and their testing vendors) will be making critical decisions as they ramp up to the next testing session and beyond. We believe that this is an extremely important juncture in state assessment policy and practice because the kinds of large-scale assessments that states adopt next will have an important bearing on K-12 instruction and learning in the coming years and possibly for decades to come, especially in the core discipline areas of English language arts, mathematics, science, and history/social studies. The federal Every Student Succeeds Act (ESSA), authorized in December 2015, requires that state assessments ‘‘involve multiple up-to-date measures of student academic achievement, including measures that assess higher-order thinking skills and understanding.” The new law provides states with flexibility in selecting their own standards and achievement measures, and also permits states to adopt measures of student academic achievement that “may be partially delivered in the form of portfolios, projects, or extended performance tasks” (ESSA, 2015, S. 1177–24).

The federal Every Student Succeeds Act (ESSA), authorized in December 2015, requires that state assessments ‘‘involve multiple up-to-date measures of student academic achievement, including measures that assess higher-order thinking skills and understanding.”

These recent changes in federal policy regarding state assessments have created new opportunities to rethink the way state assessment and accountability systems are designed.

Why focus on large-scale assessment quality and item quality?

What often gets overlooked in the public discourse about large-scale testing is what the assessments actually measure and the quality of assessment items. Stakeholders often fail to differentiate between tests that are well designed and tests that are poorly designed. Instead, all state tests are treated with the same broad brushstrokes, with little reference to their content or quality, perhaps due to a lack of transparency and/or clear communication about what is actually being tested by these assessments.

In this paper, we present and provide an in-depth analysis of a selection of large-scale assessment items (mostly drawn from current assessment systems) to illuminate design features of high quality items across a range of assessment item formats. As our analysis shows, some assessment items do a better job than others at tapping into the key disciplinary knowledge, understandings, and skills that students are expected to learn in their academic studies.

As our analysis shows, some assessment items do a better job than others at tapping into the key disciplinary knowledge, understandings, and skills that students are expected to learn in their academic studies.

We hope to provide and advocate for greater transparency to the public about what state assessments measure, and what kinds of knowledge, understandings, and skills can be measured by different item formats.

A balance of well-designed assessment items of varying formats has a higher potential of tapping into the range of key disciplinary knowledge, understandings, and skills that are central to the disciplines than a test that consists of only one item format.

Through this paper, we hope to provide and advocate for greater transparency to the public about what state assessments measure, and what kinds of knowledge, understandings, and skills can be measured by different item formats. We also recommend a set of analytic tools and questions for interrogating the quality of test items that could be applied as new large-scale assessments are developed, field tested, and approved by state assessment agencies.

In sum, this paper aims to provide helpful information in relation to the following questions:

> *How do the item formats used in large-scale assessments (e.g., selected response, short constructed response, extended response/performance-based, technology-enhanced) differ from each other? What are the strengths and drawbacks of each type of item format for measuring important learning outcomes?*

Item Analysis

We conducted our analysis of a range of items from large-scale assessments or other sources by discipline. Four teams with expertise across four disciplines—English language arts (ELA), mathematics, science, and history/social studies—worked on analyzing selected items using parallel approaches.

Each team selected 6-8 items from a variety of large-scale testing programs, including state assessments, the College Board Advanced Placement Exams, the Program for International Student Assessment (PISA), the National Assessment for Educational Progress (NAEP), and other reputable large-scale assessments administered in the United States. The teams selected items that represented a variety of item formats and assessment targets within their discipline.

Each team also selected or developed criteria for classifying the cognitive complexity of assessment items. We chose to use taxonomies that have either been endorsed by scholars in each respective discipline as being useful ways of rating cognitive demand in the discipline, or, in the absence of any discipline-specific taxonomy, developed our own.

The tools used by the individual teams are listed in the table below.

English language arts	Karin Hess's Cognitive Rigor Matrices for Close Reading Across Content Areas and for Written and Oral Communication (2009, Updated 2013) (Adapted from Norman Webb's Depth of Knowledge framework and Bloom's Revised Taxonomy)
Mathematics	Norman Webb's Depth of Knowledge (2007) framework as adapted by Herman, Buschang, and La Torre Matrundola (2014)
Science	PISA Cognitive Demand Framework (2015) (Adapted from Norman Webb's 1997 Depth of Knowledge framework)
History/social studies	Daisy Martin (2016). Evaluating Cognitive Complexity in History Tool (new instrument)

In addition, each team conducted a qualitative analysis of each selected item, beginning with the following questions:

1) What disciplinary knowledge, concepts, and/or skills does the item assess?

2) What features of the item's design and scoring criteria support measurement of cognitively complex learning targets?

3) What other strengths and/or limitations does the item illustrate?

Overview of Findings

In selecting and analyzing features of high quality and cognitively complex assessment items, each team arrived at a set of analytical frameworks to describe some of the essential design features of large-scale assessment items that support more **valid assessment of central disciplinary understandings and skills.** By "valid", we mean that an assessment item actually measures what it is intended to measure, and that it is designed to measure the intended learning targets in a manner that more closely represents work in the discipline.

Below is an overview of some of the design features or considerations that the teams identified as being key to high quality and cognitively complex items. We present these features as both findings and recommendations. Although most of these features apply to all four of the disciplines, some of the design considerations play out differently across the disciplines, as the text below indicates. These differences are explained in more detail in each of Chapters 2-5 and are illustrated through narrative analysis of the assessment item examples.

Key Design Features of High Quality Assessment Items

Design Feature 1: Items focus on core disciplinary knowledge, concepts, and/or skills.

High quality items focus on student understanding and/or application of central disciplinary ideas and processes/practices. For example, ELA writing prompts aligned to the most recent conceptions of college-ready writing call upon students to use evidence from text(s) to develop an explanation, argument, or narrative rather than drawing solely on students' opinions or experiences. In mathematics, high quality items provide students with opportunities to demonstrate conceptual understanding of core disciplinary ideas, without unnecessary difficulty in deciphering what is expected. High quality science items emphasize the big ideas and themes in science, and only assess knowledge of details when those details are central to making sense of the big ideas. In history/social studies, if an item requires reading and writing, it requires historical reading and/or writing, and scoring criteria focus on disciplinary competencies rather than generic reading or writing mechanics.

Design Feature 2: Items integrate disciplinary knowledge, understandings, and/or skills.

High quality assessment items go beyond the measurement of decontextualized facts/knowledge, literal comprehension, or generic skills (such as literacy) to integrate two or more dimensions of disciplinary learning. In science assessment, this means integrating science knowledge, a cross-cutting concept (i.e., a big idea in science), and/or a science or engineering practice. In English language arts, knowledge of language and other ELA content should be integrated with core concepts in the study of literature, core cognitive competencies (e.g., finding, selecting, evaluating, and interpreting information), and/or metacognitive competencies (i.e., the awareness of and ability to use a variety of relevant strategies to understand texts). In history/social studies, high quality items measure knowledge within its relevant context, and knowledge is integrated with history/social studies practices such as analyzing the credibility of primary and secondary sources and/or applying the big ideas in history/ social studies (e.g., change and continuity, cause and effect).

Design Feature 3: The item prompt and materials (texts, other sources) are presented in a way that maximizes student access and engagement and reduces bias. This design feature is especially important in consideration of English Learners and students with language processing or language production challenges.

High quality items are worded as simply, concisely, and clearly as possible, using student-friendly language, without potential for different interpretations of what students are asked to do in the item. Other task demands that introduce construct-irrelevant difficulty, such as overly complex prompts, texts, or writing demands, are minimized. Across all disciplines, an engaging context that is meaningful and sensible to students from diverse socioeconomic, cultural, and language backgrounds is more likely to foster student persistence and support completion of an item. In mathematics assessments, high quality items are designed to provide students a variety of ways to enter into the task, e.g., through visual representations, technology enhancements, or response formats that provide enough content for students to reason with, even if they have not memorized a certain procedure or computation.

Design Feature 4: In constructed-response and extended-response items, the item is open-ended enough to allow for a variety of student responses.

Rather than expecting a single correct response, open-ended items allow a variety of ways to demonstrate the target understandings and skills, with scoring criteria focused on students' reasoning skills. This means that constructed-response and extended-response items in mathematics, for example, allow for multiple solution strategies and, when possible, allow for more than one correct answer. In English language arts and history/social studies items, the prompts are worded in a neutral way that does not bias students toward a particular response, and the materials/texts/sources that students must use to respond to the prompts represent a variety of viewpoints.

Design Feature 5: Items require students to work with source materials that are authentic to the discipline in a way that replicates the work of the discipline.

When possible, both the materials used ("stimuli") and the work that students must do in response to an item should replicate the real work of the discipline. For example, in history/social studies, this means an item might ask students to apply an analytic lens specific to the study of history (e.g., analyzing detailed information about a source) to interpret historical artifacts, texts, or other sources in order to answer a specific question about a historical event. In science, students might be asked to manipulate variables in a technology-enhanced simulation lab to test a set of hypotheses, or to work with a set of data to answer specific questions. High quality ELA items represent not only transferable literacy skills but also core ELA concepts and ways of thinking—e.g., that language has cultural, social, and personal power; that literature both reflects and plays a role in shaping culture; and that readers construct meaning from both text and context, including relationships among texts.

Development of carefully constructed items that more closely reflect central disciplinary learning targets not only supports more valid and high quality assessment, but also represents disciplinary learning outcomes in ways that can support more productive teaching and learning.

Design Feature 6: The use of technology-enhanced items is purposeful—i.e., the technology elevates the cognitive complexity of the item or makes the item more accessible.

Given that the required use of unfamiliar technology in assessment situations adds potential barriers for students with less experience with technology and may introduce construct-irrelevant error, the use of technology must add value to an assessment. In some technology-enhanced items that we encountered, we found that they simply replicated the demands of selected-response items without enhancing the cognitive complexity of what is being measured or supporting greater access to students. If technology-enhanced items are used in large-scale assessment, they must (1) offer a productive scaffold for student reasoning without reducing the cognitive complexity of the item, or (2) elevate the cognitive complexity of the item. Technology-enhanced items (in the form of simulations, video, and interactive platforms) present especially high potential for more authentic assessment of science and engineering practices.

Conclusion

High quality assessments measure a full range of disciplinary knowledge, understandings, and processes/practices embodied in new state and national standards. This means that a range of item formats will need to be included in large-scale assessments, including selected-response, constructed-response, extended-response/performance-based, and technology-enhanced items. While some selected-response items are able to measure disciplinary knowledge and skills, they are limited in their ability to measure higher-order thinking skills in authentic and valid ways. Well-designed constructed-response items and performance tasks are able to probe much more deeply into students' reasoning and their ability to draw on their knowledge and skills as they are needed to investigate questions and solve problems. With a new set of assessment and accountability possibilities supported by ESSA, state assessment agencies now have greater latitude to include a greater variety of item formats, including curriculum-embedded, hands-on performance tasks in the classroom, as well as innovative technologies that allow for simulated investigations when hand-on resources are not available. Development of carefully constructed items that more closely reflect central disciplinary learning targets not only supports more valid and high quality assessment, but also represents disciplinary learning outcomes in ways that can support more productive teaching and learning.

CHAPTER ONE

Introduction

by Ruth Chung Wei, Ph.D.

A new direction for large-scale assessment?

We believe that this is an extremely important juncture in state assessment policy and practice because the kinds of assessments that states adopt next will have an important bearing on K-12 instruction and learning in the coming years and possibly for decades to come...

In 2010, when the federal government awarded the two Common Core testing consortia, the Partnership for Assessment of Readiness for College and Careers (PARCC) and Smarter Balanced Assessment Consortium (SBAC), $170M and $160M grants respectively through the Race To The Top program, there was great optimism about the power of state collaboration and the economies of scale that were anticipated from cost-sharing the development of large-scale assessments. Since then, however, enthusiasm about the Common Core assessments has waned for a variety of reasons. Some states have dropped out of the PARCC and Smarter Balanced consortia on the heels of unfavorable public responses to the consortia's assessments, while other states have completely abandoned use of the Common Core State Standards due to political pressures. As of October 2015, it appeared that PARCC will have only six states plus Washington, D.C. planning to administer their assessments in the 2015-16 academic year, while Smarter Balanced still has 15 states/territories planning to administer the full assessment in 2015-16.[1] In other words, a clear majority of states are planning to use other assessments.[2]

A majority of states are now working on their own as they transition to new standards, new assessments, and new accountability systems. State education agencies considering new assessments are in a pivotal moment as they get their testing programs up to speed within the next few years. Some states will fall back on their previous testing vendors (e.g., ETS, Pearson, Measured Progress, American Institutes for Research, the College Board) and other states will look to smaller, lesser known testing vendors, most of which will have shorter track records with developing assessment items aligned to the latest state and national standards.

In the next few months and years, state legislators, education commissioners, and education testing directors (and their testing vendors) will be making critical decisions as they ramp up to the next testing session and beyond. We believe that this is an extremely important juncture in state assessment policy and practice because the kinds of assessments that states adopt next will have an important bearing on K-12 instruction and learning in the coming years and possibly for decades to come.

Why should we care about large-scale assessment quality?

The 2014-15 academic year proved to be a year of significant upheaval for educational

[1] J.R. Woods (October 2015). State Summative Assessments, 2015-16 school year. Denver, CO: Education Commission of the States. Retrieved on November 23, 2015 from: http://www.ecs.org/ec-content/uploads/12141.pdf

[2] For state-by-state testing plans for the 2015-16, updated in July 2015, see the infographic at http://www.edweek.org/ew/section/multimedia/map-the-national-k-12-testing-landscape.html?intc=highsearch

assessments across many states, as parents, teachers, and other stakeholders have challenged the dominance of high-stakes testing as a key metric for district, school, and teacher accountability policies. Many of the complaints about testing revolved around the amount of time spent on testing, the amount of instructional time spent on "test prep," the cost of building a technology infrastructure to administer computerized tests, the use of high-stakes tests to evaluate teachers, and the lack of transparency about the content of the tests. Reports from the media indicate that parents are worried that the high-stakes tests are colonizing instructional time and stressing out their children. Stakeholders from both ends of the political spectrum have joined forces to support a parent opt-out movement to protest the high-stakes tests.

...state tests are treated with the same broad brushstrokes, with little reference to their content or quality, perhaps due to a lack of transparency and clear communication about what is actually tested in these assessments.

But what often gets overlooked in the public discourse about large-scale testing is what the assessments actually measure and the quality of assessment items. Other than a few items that have been scrutinized repeatedly in the media (e.g., Pearson's talking pineapple debacle), the media, parents, stakeholders, and even the defenders of high-stakes testing too frequently fail to differentiate between tests that are well designed and tests that are poorly designed. Instead, all state tests are treated with the same broad brushstrokes, with little reference to their content or quality, perhaps due to a lack of transparency and clear communication about what is actually tested in these assessments.

We should pay attention to the quality and content of large-scale assessments, however, because we know that state testing is not going away. The federal Every Student Succeeds Act (ESSA), authorized in December 2015, continues to require that states administer annual testing in English language arts and mathematics in Grades 3-8 and once in high school, and, in science, once at each of three grade spans (3-5, 6-9, and 10-12).

We also know that state tests have the power to drive curriculum and instruction within schools, especially when there are significant consequences attached to the outcomes of those tests. Research shows that when results from state assessments are used for high stakes, such as high school exit requirements, teacher evaluation, or school and district accountability, those assessments drive much of what teachers teach and how they teach their students (Arirasian, 1987; Shephard & Dougherty, 1991; Smith, 1991; Amrein & Berliner, 2002; Madaus, 1988; Madaus & Clarke, 2001; Jacob, 2005; Abrams, Pedullah, & Madaus, 2003; Rentner, et al., 2006; Diamond, 2007; Vogler, 2007; Griffith & Scharmann, 2008; Plank & Condliffe, 2013).

Other studies have found that when large-scale assessments are well designed to measure more cognitively complex learning goals, such as higher-order thinking, and do not focus only on basic skills and factual recall, this has a positive impact on teachers' instruction (Yeh, 2005), resulting in teachers using more open-response questions, creative/critical thinking questions, problem-solving activities, writing assignments, and inquiry/investigation activities in their classrooms (Vogler & Virtue, 2002). Research on the inclusion of performance assessments in large-scale assessment programs of the 1990s also provides evidence that richer, more complex forms of assessment support richer curriculum and instruction (Koretz, Stecher, Klein, & McCaffrey, 1994; Matthews, 1995; Stecher & Mitchell, 1995; Wolf, Borko, Elliott, & McIver, 2000; Chung & Baker, 2003).

The Every Student Succeeds Act requires that state assessments ‘‘involve multiple up-to-date measures of student academic achievement, including measures that assess higher-order thinking skills and understanding.” Although ESSA includes a number of stipulations and restrictions regarding assessment, such as the aforementioned required schedule for state testing in ELA, math, and science, alignment of assessments to “challenging state academic standards,” required tracking of data for subgroups, and limited permission for parent opt-outs (95% student participation is required), many states welcome the greater flexibility ESSA allows in selecting their own standards and achievement measures. ESSA even specifies that measures of student academic achievement “may be partially delivered in the form of portfolios, projects, or extended performance tasks” (ESSA, 2015, S.1177–24).

Some items do a better job than others at tapping into the key disciplinary knowledge, understandings, and skills that students are expected to learn in their study of English language arts, mathematics, science, and history/social studies.

In addition, ESSA authorizes that up to seven states may be granted the opportunity to pilot innovative assessment systems that may include “competency-based assessments, instructionally embedded assessments, interim assessments, cumulative year-end assessments, or performance-based assessments that combine into an annual summative determination for a student, which may be administered through computer adaptive assessments” (ESSA, 2015, S.1177-84).

Altogether, these recent changes in federal policy regarding state assessments have created new opportunities to rethink the way state assessment and accountability systems are designed and delivered.

Why focus on item quality? Not all large-scale tests are created equal, nor are the items within a test. In this paper, we aim to provide the public, test developers, and policymakers a closer look at some large-scale test items and what they actually measure. Some items do a better job than others at tapping into the key disciplinary knowledge, understandings, and skills that students are expected to learn in their study of English language arts, mathematics, science, and history/social studies.

Some scholars, including those here at UL/SCALE, argue that performance tasks have the greatest potential for assessing students’ complex understandings, but they acknowledge that there are practical concerns with administering performance tasks in large-scale assessment settings. Part of the public outcry about too much testing time in 2014-15 was due in part to the inclusion of performance tasks in the PARCC and Smarter Balanced assessment systems, tasks that require a longer administration time over multiple sessions. When designed with care, however, even machine-scored items can tap into thinking skills central to the disciplines. A balance of well-designed assessment items of varying formats has a higher potential of tapping into the range of key disciplinary knowledge, understandings, and skills that are central to the disciplines than a test that consists of only one item format.

The purpose of this paper is to present examples of high-quality, large-scale assessment items of a variety of formats (e.g., selected response, short constructed response, extended-response/performance tasks, technology-enhanced items) that are likely to measure the key disciplinary knowledge, understandings, and skills that are central to the four disciplines that are typically part of state testing programs—that is, English language arts, mathematics, science, and history/social studies. In defining what is important in each discipline, we refer to the Common Core State Standards (2010), the Next Generation Science Standards (2013), or the College, Career, and Civic Life Framework for Social Studies State Standards (2013), but we do not depend on them given that many states do not use these standards/framework. Instead, our definitions of key disciplinary knowledge, understandings, and skills are based on research and consensus within each disciplinary field. Through this paper, we aim to illuminate design features of high quality items across a range of assessment item formats and recommend a set of analytic tools and questions for interrogating the quality of test items that could be applied as new large-scale assessments are developed, field tested, and approved by state assessment agencies. By illuminating the content and demands of different types of test items, we also hope to provide and advocate for greater transparency to the public about what state assessments measure, and what kinds of knowledge, understandings, and skills can be measured by different item formats.

A balance of well-designed assessment items of varying formats has a higher potential of tapping into the range of key disciplinary knowledge, understandings, and skills that are central to the disciplines than a test that consists of only one item format.

This paper on item quality is part of a larger study on the state of state assessment systems being undertaken by Understanding Language/Stanford Center for Assessment, Learning, & Equity (UL/SCALE).

The entire research series, which includes three components, aims to answer the following questions:

1. How do the item formats used in high stakes assessments (e.g., selected response; short constructed response; extended response/performance tasks; technology-enhanced tasks) differ from each other, and what are the strengths and drawbacks of each type of item format for measuring important learning outcomes?

2. What types of item formats are used across state assessments, and in what proportion?

3. What will it take for a state's assessment system to measure a broader range of content knowledge, college and career readiness skills, and other valued learning outcomes?

This paper addresses the first question by closely examining the measurement characteristics of a range of item formats in each of the four discipline areas. In this component of the study, we provide a set of item quality analytics that state assessment developers and others can use to evaluate the measurement quality of items that are included in their state assessments.

The final report will be available later in 2016.

Item Selection and Analysis – Our Methodologies

Each of the four disciplines represented in our study (i.e., English language arts, mathematics, science, and history/social studies) used basically the same methodology for selecting items for analysis, although there were some differences. These differences are described in each individual content chapter. An overall view of the methodologies used in our item review is presented below.

Item Selection

In our review of item quality, a team from each of the discipline areas drew on sample items from released tests from a variety of large-scale testing programs. As part of our review of all fifty states' assessments' systems, we created a bank of released tests from the most recent year of administration that was available. We call this bank of released tests the "50 State Assessment Collection." However, we did not limit our search for sample items to state assessments. We also searched for released items from the College Board Advanced Placement Exams, the Program for International Student Assessment (PISA), the National Assessment for Educational Progress (NAEP) and other reputable large-scale assessments administered in the United States. These items include selected-response (multiple-choice) items, short constructed-response items, extended-response items/performance tasks, and some innovative technology-enhanced items. It is important to note that the selected items do not represent all of the "best" high-quality items and item formats that exist in large-scale assessments; rather, they represent a limited sample of high-quality items that could be identified given our access to released items, our limited time, and our ability to gain copyright permissions for this publication. These sample items are not "perfect" along every possible dimension of quality, but they do illustrate different approaches to measuring disciplinary knowledge, understandings, and skills in ways that are often underutilized in large-scale state assessments. In some cases, the sample items were selected because they show a particular strength as a counterpoint to a common problem in large-scale assessment item design.

Item Analysis

For each of the four discipline areas (English language arts, math, science, history/social studies), we selected or developed criteria for classifying the cognitive demand and disciplinary rigor of the items. Because "cognitive demand" and "rigor" are defined in different ways across the disciplines, we chose to select taxonomies that have either been endorsed by scholars in each respective discipline as being useful ways of rating cognitive demand in the discipline, or, in the absence of any discipline-specific taxonomies, developed our own. The English language arts, mathematics, and science teams each selected existing taxonomies as lenses to quantify the analysis of items that had been identified as high quality items. In the area of history/social studies, however, there is no such existing taxonomy, so our Director of History/Social Studies Learning opted to develop a discipline-specific taxonomy based on research and consensus in the field about cognitive complexity in history/social studies. Each of these quantitative tools used to assess the cognitive demand of items is described briefly below (and more fully in the subject-specific chapters that follow). This section also describes the method we used to analyze the quality of the selected items, focusing on three important dimensions.

Cognitive Complexity Analysis Tools

English language arts	Karin Hess's Cognitive Rigor Matrices for Close Reading Across Content Areas and for Written and Oral Communication (2009, Updated 2013) (Adapted from Norman Webb's Depth of Knowledge framework and Bloom's Revised Taxonomy)
Mathematics	Norman Webb's Depth of Knowledge (2007) framework as adapted by Herman, Buschang, and La Torre Matrundola (2014)
Science	PISA Cognitive Demand Framework (2015) (Adapted from Norman Webb's 1997 Depth of Knowledge framework)
History/social studies	Daisy Martin (2015). Evaluating Cognitive Complexity in History Items Tool (new instrument)

With one exception (history/social studies), we see that three of the four cognitive complexity measures are adapted from Norman Webb's Depth of Knowledge (DOK) framework. DOK has become popularized through the dissemination of the Common Core State Standards as a way of differentiating the cognitive demand of assignments. At the same time, we also note that every discipline has a specific way of adapting the DOK framework to the particular demands of the discipline. These adaptations of the DOK framework acknowledge the different kinds of knowledge that are valued across disciplines.

Qualitative Analysis Dimensions

In addition to the cognitive complexity tools for assessing the cognitive demand of items, we analyzed the items using a qualitative lens that includes at least these three common dimensions as important criteria for item quality:

1. Are key **disciplinary knowledge/concepts/skills** assessed? The common perception among educators and the general public is that most items on state assessments have emphasized the measurement of decontextualized knowledge (facts), and excluded the assessment of more complex conceptual understanding and skills. This perception may account for complaints about "breadth versus depth" in assessments, a criticism teachers often voice about state tests, particularly at the middle and secondary levels.

2. Does the assessment item **integrate disciplinary knowledge/concepts/skills** or are these dimensions assessed independently of each other? In general, we find that the more an assessment item integrates these three dimensions of disciplinary demands (factual knowledge, concepts, and skills), the more cognitively complex the item is and

the more authentic to the discipline. This does not mean that every item on a test should integrate all three kinds of disciplinary demands, but that more cognitively complex items are more likely to measure at least two or more dimensions of these disciplinary demands. A summative assessment that includes items representing a wide range of disciplinary demands would provide more valid and comprehensive evidence of student learning and performance in the disciplines than an assessment that includes items that each measure one dimension alone.

3. What is the **item format** and how does the response format impact the cognitive demand of the item? Selected-response items are not always focused on decontextualized facts, but they are limited in their ability to elicit key disciplinary demands, such as writing an evidence-based argument or evaluating someone else's argument. Short constructed-response and extended-response items (including performance tasks) have more potential for assessing the higher-order thinking skills that are valued in the disciplines (e.g., making use of source materials in one's writing, designing a way to investigate a science-related phenomenon, evaluating someone else's claims).

These three dimensions were applied across the disciplines. There were also some discipline-specific dimensions used in the item analysis. For example, in English language arts and history/social studies, a unique criterion for item quality is the authenticity and complexity of the texts, materials, and sources that students interact with when responding to an assessment item or a series of assessment items (i.e., a "thematic block"). Here, by authenticity, we mean that students are asked to interact with the types of texts or sources central to study of English language arts and history/social studies. The English language arts item analysis also examined whether an item assessed metacognitive strategies in reading/literacy as well as authentic uses of literacy.

In mathematics, item analysis also included paying special attention to design features of items that 1) strengthen item validity, such as clarity and accuracy of the language used; 2) strengthen accessibility, such as an engaging, relevant context, and 3) provide scaffolding within the item, such as opportunities for multiple points of entry or multiple solution strategies.

Across all of the disciplinary item analyses, the unique design features of technology-enhanced items were also carefully reviewed to assess the value added by technology in terms of whether it provided scaffolds for students interacting with the item, enhanced the authenticity of the item (e.g., a simulation), or elevated the cognitive demand of the item.

Limitations

In presenting selected items as promising examples of large-scale assessment items that measure a broad range of assessment targets, including the more cognitively complex ways of thinking, reading, and communicating in each discipline, we again take special care to emphasize that these items do not represent the "best" or "perfect" assessment items. Rather, these items are a limited sample of high-quality items that allow us to see in a concrete way how different item design features are able to elicit evidence of a variety of disciplinary knowledge, concepts, and skills. There are likely to be additional innovative assessment types that we were not able to access because they are proprietary or have not been released.

In addition, although almost all of the items we selected have been piloted or used with students, we know that a key way to evaluate an item's quality and what it measures is to conduct cognitive labs or think-alouds with students as they complete the item. We did not have the resources to conduct these cognitive labs. But we do suggest that test-developers build in time and resources to conduct such analyses on innovative items to evaluate the validity, accessibility, and fairness of new item types and formats.

Lastly, we know that it is the collection of items within an assessment on which an assessment's validity and reliability rests, and we acknowledge that we have conducted our analysis on items that have been de-contextualized from the collection of items in which they were administered. However, our analysis is not meant to focus on lifting up particular items as exemplars – but rather to explain and illuminate, through the use of selected items, the cognitive complexity and qualitative criteria (e.g., design features) of quality items.

The next four chapters present an analysis of a selection of large-scale assessment items in each disciplinary field: 1) English language arts, 2) mathematics, 3) science, and 4) history/social studies. These were completed by UL/SCALE's disciplinary experts in curriculum, instruction, and assessment. The authors of each chapter are noted at the beginning of each chapter.

References

Abrams, L. M., Pedulla, J. J., & Madaus, G. F. (2003). Views from the classroom: Teachers' opinions of statewide testing programs, *Theory Into Practice*, *42*(1), 18-29. http://dx.doi.org/10.1207/s15430421tip4201_4

Airasian, P. W. (1987). State mandated testing and educational reform: Context and consequences. *American Journal of Education*, *95*(3), 393-412.

Amrein, A. L., & Berliner, D. C. (2002). *An analysis of some unintended and negative consequences of high-stakes testing.* Tempe, AZ: Educational Policy Studies Laboratory, Arizona State University.

Chung, G. K. W. K., & Baker, E. L. (2003). T*he impact of a simulation and problem-based learning design project on student learning and teamwork skills.* National Center for Research on Evaluation, Standards, and Student Testing (CRESST), CSE Technical Report 599.

Diamond, J. B. (2007). Where the rubber meets the road: Rethinking the connection between high-stakes testing policy and classroom instruction. *Sociology of Education, 80*(4), 285-313. doi: 10.1177/003804070708000401

Every Student Succeeds Act (2015, December 10). Retrieved from: https://www.gpo.gov/fdsys/pkg/BILLS-114s1177enr/pdf/BILLS-114s1177enr.pdf

Griffith, G., & Scharmann, L. (2008). Initial impacts of No Child Left Behind on elementary science education. *Journal of Elementary Science Education*, *20*(3), 35-48.

Jacob, B.A. (2005). Accountability, incentives and behavior: the impact of high-stakes testing in the Chicago Public Schools. *Journal of Public Economics*, *89*, 761–796.

Koretz, D., Stecher, B., Klein, S., & McCaffrey, D. (1994). The Vermont Portfolio Assessment Program: Findings and implications. *Educational Measurement: Issues and Practice*, *13*(3), 5-16.

Madaus, G. F. (1988). The influence of testing on the curriculum. In L. N. Tanner (Ed.), *Critical Issues in Curriculum: 87th Yearbook of the National Society for the Study of Education, Part I.* (pp. 83-121). Chicago, IL: University of Chicago Press.

Madaus, G. F., & Clarke, M. (2001). The adverse impact of testing on minority students: Evidence from one hundred years of test data. In G. Orfield & M.L. Kornhaber (Eds.), *Raising Standards or Raising Barriers? Inequality and High-Stakes Testing in Public Education* (pp. 85-106). New York, NY: Century Foundation Press.

Matthews, B. (1995). *The implementation of performance assessment in Kentucky classrooms.* Louisville, KY: University of Louisville.

Plank, S. B. & Condliffe, B. F. (2013). Pressures of the season: An examination of classroom quality and high-stakes accountability. *American Educational Research Journal*, *50*(5), 1152–1182. doi:10.3102/0002831213500691

Rentner, D. S., Scott, C., Kober, N., Chudowsky, N., Chudowsky, V., Joftus, S., & Zabala, D. (2006, March 28). *From the capital to the classroom: Year 4 of the No Child Left Behind Act.* Washington, DC: Center for Education Policy. Retrieved September 22, 2015 from http://www.cep-dc.org.

Shepard, L.A., & Dougherty, K. (1991, April). *Effects of high-stakes testing on instruction.* Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Smith, M. L. (1991). Put to the test: The effects of external testing on teachers. *Educational Researcher*, 20(5), 8-11. doi: 10.3102/0013189X020005008

Stecher, B. M., & Mitchell, K. J. (1995). *Portfolio-driven reform: Vermont teachers' understanding of mathematical problem solving and related changes in classroom practice.* CSE Technical Report 400. Los Angeles, CA: University of California Center for Research on Evaluation, Standards, and Student Testing.

Vogler, K. E. (2002). The impact of high-stakes, state-mandated student performance assessment on teachers' instructional practices. *Education*, *123*(1), 39-55.

Vogler, K. E., & Virtue, D. (2007). "Just the facts, ma'am": Teaching social studies in the era of standards and high-stakes testing. *The Social Studies, 98*(2), 54-58. http://dx.doi.org/10.3200/TSSS.98.2.54-58

Wolf, S., Borko, H., Elliott, R. L., & McGiver, M. (2000). ''That dog won't hunt!'': Exemplary school change efforts within the Kentucky reform. *American Educational Research Journal*, *37*(2), 349-393. doi: 10.3102/00028312037002349

Yeh, S.S. (2005). Limiting the unintended consequences of high-stakes testing. *Education Policy Analysis Archives*, *13*(43). Retrieved September 22, 2015 from http://epaa.asu.edu/ojs/article/view/148. doi: http://dx.doi.org/10.14507/epaa.v13n43.2005

CHAPTER TWO

Evaluating Item Quality in English Language Arts Assessments

by Nicole Renner

Introduction

Teachers and scholars of English language arts (ELA) consistently discuss what constitutes appropriate curriculum, pedagogy, and assessment in the discipline. Controversies and debates around topics like direct grammar instruction, formalist criticism vs. reader response theory, and the role and importance of literature, including shifting conceptions of the literary canon and definitions of the word “text,” have dominated much of the discipline’s discourse for decades.

The discipline has also consistently held certain core values, including some common conceptions of cognitive rigor and complexity. The 1996 NCTE/IRA Standards for the English Language Arts are widely accepted and have repeatedly been affirmed as representing central values of the discipline. The vision these standards present of ELA learning is simultaneously rigorous and humanist. They emphasize the cognitive complexity of encountering the world through text, demanding that students, for example, “apply knowledge of language structure, language conventions . . . media techniques, figurative language, and genre to create, critique, and discuss print and nonprint texts” and that students “conduct research… by generating ideas and questions, and by posing problems. They gather, evaluate, and synthesize data from a variety of sources…to communicate their discoveries in ways that suit their purpose and audience” (NCTE/IRA, 1996). The standards also recognize that ELA is fundamentally a humanities subject, specifying that students should “read a wide range of literature from many periods in many genres to build an understanding of the many dimensions (e.g., philosophical, ethical, aesthetic) of human experience.” Despite widespread agreement on these core values of the discipline, many large-scale English and literacy assessments, textbooks, and curricular materials have failed to provide opportunities for students to demonstrate the complex skills and dispositions that reflect these values.

Despite widespread agreement on these core values of the discipline, many large-scale English and literacy assessments, textbooks, and curricular materials have failed to provide opportunities for students to demonstrate the complex skills and dispositions that reflect these values.

Instead, assessments have tended to focus on low-level reading comprehension, speculative rather than evidence-based interpretation of text, and identification of literary elements in decontextualized passages of text (Applebee & Langer, 2013; Applebee, 1993; Lynch & Evans, 1963; Mihalakis & Petrosky, 2015).

Elements of the NCTE/IRA standards, however, have recently become visible in the Common Core State Standards (CCSS) and other new state standards that increasingly reflect the notion that cognitive complexity and rigor in ELA exist in the intersection among complex text, close reading, critical perspectives, and evidence-based communication—all embedded in authentic academic, civic, and personal purposes. This recent resurgence of interest in deeper learning and cognitive complexity has provided a fresh opportunity to focus large-scale assessment on cognitively demanding tasks that “require students to read, re-read, and analyze complex texts to develop oral and written explanations,

The ELA discipline is poised to realign ELA assessments and classroom practice with the core disciplinary values and best practices reflected in the NCTE/IRA standards, eliciting evidence of students’ complex thinking about texts and the worlds they represent.

interpretations, and arguments” (Mihalakis & Petrosky, 2015). Particularly with the passage of the Every Student Succeeds Act (ESSA) in 2015, the ELA discipline is poised to realign ELA assessments and classroom practice with the core disciplinary values and best practices reflected in the NCTE/IRA standards, eliciting evidence of students’ complex thinking about texts and the worlds they represent.

Large-scale test developers will need well-defined specifications and examples to develop assessments that fully embrace this shift in approach toward assessment. However, test developers need not start completely from scratch; they can build upon existing resources and assessments that, to varying degrees, assess student performance of important skills and understandings within the discipline of ELA at high levels of cognitive complexity. To aid developers, we conducted a review of large-scale ELA assessment items, looking for items that illustrate a set of core features that support cognitively complex assessment within English language arts. In the sections below, we present the methods we used to conduct this review, followed by four noteworthy item features that emerged from the review and sample items that illustrate each of these features.

Methods

This section first describes the data collection methods used to select English language arts items that we consider to be promising (i.e., they assess important skills and understandings at high levels of cognitive complexity), and then it describes the criteria, frameworks, and procedures we used to analyze each of the selected items.

Item Selection

The goal of our assessment review was to select items of high cognitive complexity that measured authentic disciplinary ways of thinking, reading, and writing, regardless of item type or content focus. To find exemplars of cognitively complex items among those publicly available from large-scale ELA assessments, we convened a team of experts to review a wide range of state and consortium assessments, as well as national and international assessments (e.g., Advanced Placement, PISA). This team—which consisted of two ELA content experts, with support from a team of seven interdisciplinary assessment experts—used an item selection process that considered the text(s) referenced in the item, the content or substance of the item, the item’s evaluative criteria (e.g., scoring rubric) when available, and the item’s format and design features.

Item Type, Format, and Grade Level. The team’s final selection of sample items represents a range of types, formats, grade levels, and content foci. Altogether, we chose six items to present as sample items.

Table 1

Summary of item types selected for analysis by grade

Grade Level	Selected-Response	Short Constructed- Response	Performance Tasks
4		1	
8	1		1
9-12	1		2

The items we selected skew toward eighth grade and high school because we found more examples of high-quality, performance-based items at those grade levels. However, the qualities of these items can be found in elementary-level items as well. In some cases, we describe in the analysis that accompanies each item how these upper grade level items could be adapted for lower grade levels while retaining their high cognitive rigor.

While part of our goal in this analysis was to highlight innovative item formats that increase the measurement potential of standardized assessments, most of the ELA items we selected are quite traditional in format. We only selected technology-enhanced items if the use of technology actually elevated the cognitive rigor of the item. We found many examples of technology-enhanced items that accomplished nothing that a traditional selected- or constructed-response item could not already do. In other cases, technology was used to give students the opportunity to read and gather information from “dynamic texts” (OECD, 2013) such as simulated websites, which made the items more authentic to how students typically gather information in real life. These uses of technology, however, still did not significantly affect item quality or cognitive rigor.

Content Focus. The six items we selected for analysis represent three content foci: reading informational text, reading literary text, and writing. Because open-ended questions about text require writing, there was some overlap between “reading items” and “writing items.” We distinguished among the three content foci by determining what the item was intended to measure as indicated by the prompt, evaluative criteria, and any available metadata.

Table 2

Number of items selected by content focus

Reading Informational Text	Reading Literary Text	Writing
2	2	2

We chose not to select any items that assessed grammar, conventions, usage, or vocabulary because we believe there is limited potential for high cognitive complexity in such items. Furthermore, we know that grammar instruction is most effective in the context of writing (Calkins, 1980; DiStefano and Killion, 1984; Harris, 1962), and that discrete “school grammar” instruction usually does not improve the quality of student writing (Hillocks, 1986; Hillocks & Smith, 1991). Accordingly, typical grammar assessment items that ask students to correct errors, select the “best” version of sentences or paragraphs, etc. do not authentically measure students’ ability to use standard written English grammar and conventions in their natural context—writing. Instead, they measure a student’s editing knowledge/skills.

Design Features. Finally, some of the items we selected represent design features that contribute to cognitive complexity or to increased measurement potential. One such design feature we will highlight in the item analyses below is the use of coherent “item blocks” around one or more texts, wherein the items build upon each other toward a culminating performance task. This strategy supports student engagement and achievement on high cognitive rigor tasks because it creates a form of scaffolding, helping students access progressively deeper layers of textual interpretation as they progress through the item set. This provides a more coherent assessment experience for students and yields potentially greater evidence of students’ depth of understanding of complex texts because of the opportunity it provides for sustained engagement with the text.

Item Analysis: Theoretical Frameworks

We used several theoretical frameworks to analyze each sample ELA item. To evaluate the cognitive complexity of each selected item, we applied Hess’s Cognitive Rigor Matrices for Close Reading Across Content Areas and for Written and Oral Communication. We applied the Cognitive Rigor Matrices, which are neutral with regard to content domain within English Language Arts, to each of the items selected for analysis. For items with a content focus on reading literary text, we also applied George Hillocks and Larry Ludlow’s 1984 Taxonomy of Skills in Reading and Interpreting Fiction. These frameworks are discussed below and can be found in Appendix A and Appendix B, respectively.

Cognitive Rigor Matrices. The Cognitive Rigor Matrices are popular as tools for evaluating the cognitive complexity of assessments and assignments. They are used as a framework for developing and evaluating assessment items by various educational and testing organizations, including the Smarter Balanced Assessment Consortium (SBAC), the New York City Department of Education, Chicago Public Schools, and the Teachers College Reading & Writing Project. They are also used by numerous states, including Arizona, California, Colorado, Georgia, Iowa, New Hampshire, Ohio, Utah, and Vermont for a range of purposes, including evaluating state assessments and as professional tools for teachers to design rigorous classroom assignments.

As shown in Appendix A, the Cognitive Rigor Matrices map Bloom’s Revised Taxonomy (2001) and Norman Webb’s Depth of Knowledge (DOK) levels (2002) along two axes to provide a multidimensional look at the cognitive complexity of items. The vertical axis, representing Webb’s DOK levels, characterizes an item’s procedural complexity and the depth of content understanding required to successfully complete the item, while the horizontal axis, representing Bloom’s Revised Taxonomy, characterizes the type of cognitive processes the item calls upon. The matrices illustrate each intersection of Bloom’s Revised Taxonomy levels and Webb’s DOK levels with examples of discipline-specific tasks.

It is extremely important to note that Bloom's Revised Taxonomy and Webb's Depth of Knowledge levels are conceptually distinct and focus on different features of an assessment item or an assignment. As Hess, Carlock, Jones, and Walkup explain in their article about cognitive rigor,

> *Although related through their natural ties to the complexity of thought, Bloom's Taxonomy and Webb's depth-of-knowledge differ in scope and application. Bloom's Taxonomy categorizes the cognitive skills required of the brain to perform a task, describing the 'type of thinking processes' necessary to answer a question. Depth of knowledge, on the other hand, relates more closely to the depth of content understanding and the scope of a learning activity, which manifests in the skills required to complete the task from inception to finale (e.g., planning, researching, drawing conclusions) (2009, p. 3).*

We assigned each item a single Depth of Knowledge level because evaluating the DOK of an item requires a holistic examination of the demands of the entire item; accordingly, an item cannot simultaneously represent multiple DOK levels. However, more complex items frequently call upon multiple types of thinking, which is why some items in our analysis represent multiple levels of Bloom's Revised Taxonomy.

Hillocks and Ludlow's Taxonomy of Skills in Reading and Interpreting Fiction

Hillocks and Ludlow's Taxonomy of Skills in Reading and Interpreting Fiction describes seven distinct and key skills that are authentic to professionals in the discipline, such as "reading experts, teachers of literature, and literary critics" (1984, p. 8). The skills are divided into two categories: literal level of comprehension and inferential level of comprehension. Unlike Bloom's taxonomy, this schema is hierarchical, whereby the skills are arranged in a progression of increasing difficulty and complexity. Skills within the literal level are 1) Basic Stated Information, 2) Key Detail, and 3) Stated Relationship. Skills within the inferential level are 4) Simple Implied Relationship, 5) Complex Implied Relationship, 6) Author's Generalization, and 7) Structural Generalization. See Appendix B for further explanation and examples of these seven skills. Applying this taxonomy to ELA assessment items helped illuminate whether the items called upon "higher order" skills (i.e., the skills within the inferential level) and understandings that are valued within the discipline.

Item Analysis

To determine the placement of items on the Cognitive Rigor Matrices, we evaluated each item along the two axes separately, determining the appropriate DOK level and Bloom's Revised Taxonomy level(s) to describe the item. We then compared the item to the descriptors and examples in the relevant matrix to confirm our placement of the item at the correct "intersection" of DOK and Bloom's.

To assign the items with a content focus on reading literary texts a level on the Hillocks and Ludlow Taxonomy, we evaluated each item against the descriptions and examples of the seven taxonomy levels found in Hillocks and Ludlow's original 1984 publication as well as in the 2009 NCTE publication "Writing about Literature."

In our qualitative analysis of each item, we evaluated the item's potential to measure conceptual understandings, skills, and modes of thinking that are authentic and significant to the discipline of English Language Arts.

This evaluation was not directly tied to any single framework, but it was informed by several sources:

1. **The NCTE/IRA Standards for the English Language Arts**
2. **The PISA Reading Literacy Framework**
3. **The Key Shifts in English Language Arts from the Common Core State Standards**

From these sources, we distilled a set of criteria and guiding questions to assess the quality and rigor of each item from a disciplinary standpoint:

Qualitative Criteria and Guiding Questions for Item Analysis

Criterion	Guiding Question
Disciplinary Value	What significant disciplinary skills and understandings does this item call upon and provide evidence of?
Text Complexity	Does the item require students to read and grapple with texts characterized by complex language and multiple layers of meaning and purpose? If an item involves multiple texts, do students have to think critically about the relationship(s) among the texts?
Text Dependence	Is the item truly text-dependent? Does it require students to draw upon and justify responses with evidence from texts, both literary and informational?
Application of Cognitive and Metacognitive Strategies to Informational Text	Does the item provide opportunities for students to use a variety of strategies to find, select, evaluate, and interpret information from informational text?
Range of Authentic Texts	Does the item provide opportunities to construct meaning from texts that are authentic to personal, academic, civic, and/or professional purposes?
Format and Design Features	How do the format and design features of the item contribute to its cognitive complexity and its potential to measure significant disciplinary concepts and skills?
Integration and Application of Understandings and Skills	Does the item integrate conceptual understanding and application of skills rather than target discrete knowledge/skills in isolation?
Cognitive Rigor	What other characteristics of the item justify its placement at a particular level on the appropriate cognitive rigor taxonomy (i.e., Cognitive Rigor Matrix; Taxonomy of Skills in Reading and Interpreting Fiction)?

Summary of Findings

In our examination of a wide range of items using the above criteria and guiding questions, we identified four core features of high-quality ELA assessment items, which we will explain and illustrate through our item analyses. The four features are as follows:

Feature 1
Items that are text-dependent and measure more than literal comprehension

Feature 2
Items that measure cognitive and metacognitive competencies for reading literacy

Feature 3
Items that require evidence-based writing

Feature 4
Items that call for authentic disciplinary uses of literacy

In the section that follows, we explain each feature, present one or two examples of items that reflect that feature, and then narrate our analysis of the item's quality. Each item is accompanied by a profile that identifies the item type, content focus, grade level, scoring method (computer or human), and cognitive rigor as represented by the item's DOK level, Bloom's Revised Taxonomy level, and Hillocks & Ludlow Taxonomy of Skills level (where applicable).

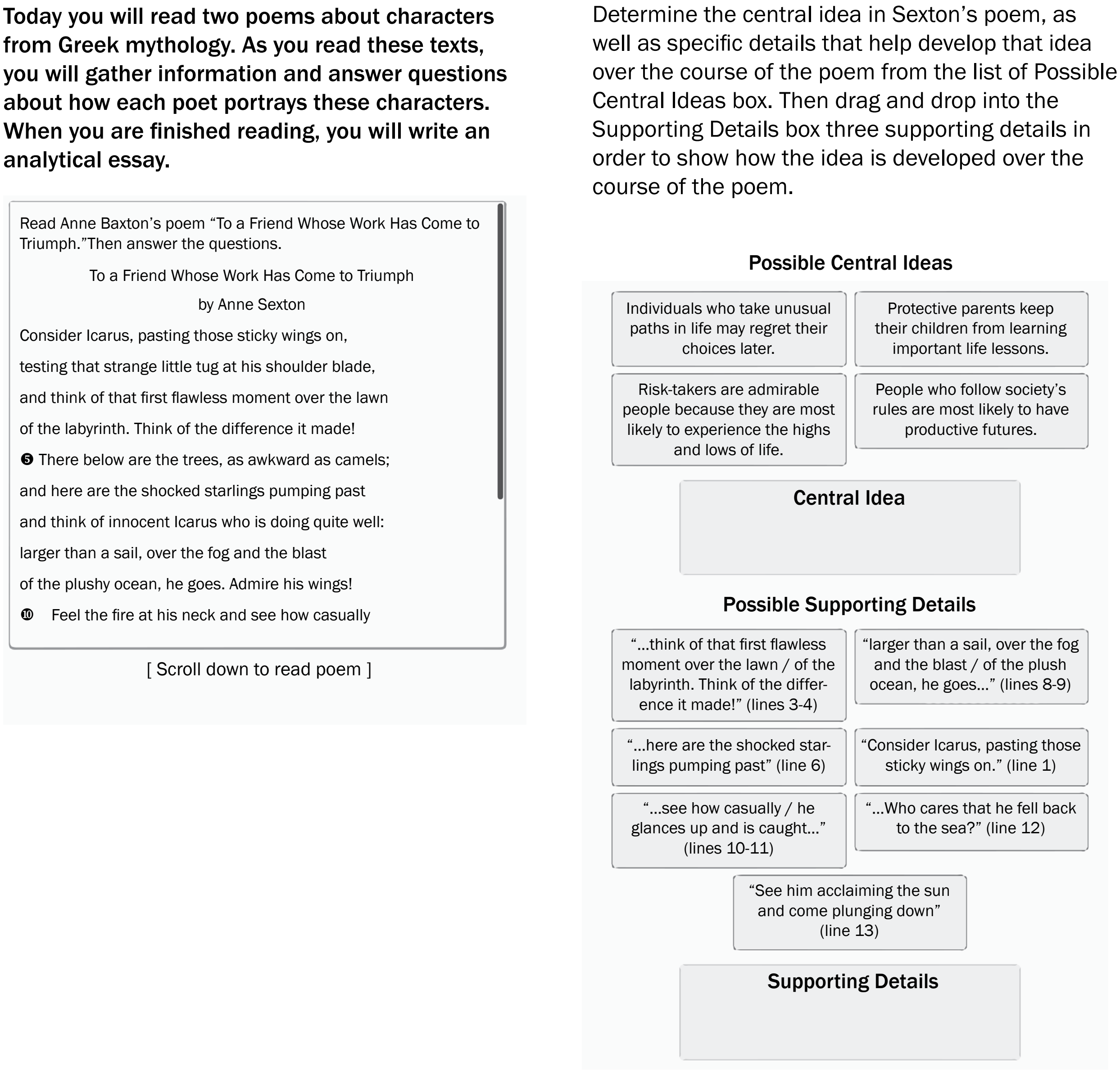
Feature 1: Items that are text-dependent and measure more than literal comprehension

This feature reflects the current push, both within and across disciplines, for students to read and make meaning from more complex texts. An implied criterion of this feature is that students actually encounter rich, complex texts in the assessment, as it is "virtually impossible to develop a sequence of intellectually demanding tasks for a shallow and simplistic text" (Mihalakis & Petrosky, 2015). However, although the presence of complex texts is necessary, it is not sufficient. By definition, complex texts have multiple layers of meaning and/or purpose, and to assess students' ability to work with complex text, we must ask them to read for those layers and to construct meaning across whole texts rather than "strip-mining texts to find clearly stated ideas . . . and to recognize literary devices" (Mihalakis & Petrosky, 2015).

This feature actually has two components: first, the item must be truly text-dependent—i.e., answering the question posed by the item depends on reading (and often rereading) and closely examining the text(s). Additionally, the item must require at least simple inter-sentence analysis or inference; preferably, it requires inference across a significant passage or an entire text.

We found that a wide variety of item types, from selected response (i.e., multiple choice) to performance task, have the potential to measure high-level interpretive skills, though items that require students to write (i.e., constructed-response items and performance tasks) can provide more valid, authentic, and detailed evidence of students' skills. In the section that follows, we present two examples of items that reflect Feature 1. The first example is a high school level, technology-enhanced, selected-response item. The second is a high school level performance task.

Item Example 1 - Item Reflecting Feature 1



Today you will read two poems about characters from Greek mythology. As you read these texts, you will gather information and answer questions about how each poet portrays these characters. When you are finished reading, you will write an analytical essay.

Read Anne Baxton's poem "To a Friend Whose Work Has Come to Triumph."Then answer the questions.

To a Friend Whose Work Has Come to Triumph

by Anne Sexton

Consider Icarus, pasting those sticky wings on,
testing that strange little tug at his shoulder blade,
and think of that first flawless moment over the lawn
of the labyrinth. Think of the difference it made!
❺ There below are the trees, as awkward as camels;
and here are the shocked starlings pumping past
and think of innocent Icarus who is doing quite well:
larger than a sail, over the fog and the blast
of the plushy ocean, he goes. Admire his wings!
❿ Feel the fire at his neck and see how casually

[Scroll down to read poem]

Determine the central idea in Sexton's poem, as well as specific details that help develop that idea over the course of the poem from the list of Possible Central Ideas box. Then drag and drop into the Supporting Details box three supporting details in order to show how the idea is developed over the course of the poem.

Possible Central Ideas

Individuals who take unusual paths in life may regret their choices later.

Protective parents keep their children from learning important life lessons.

Risk-takers are admirable people because they are most likely to experience the highs and lows of life.

People who follow society's rules are most likely to have productive futures.

Central Idea

Possible Supporting Details

"...think of that first flawless moment over the lawn / of the labyrinth. Think of the difference it made!" (lines 3-4)

"larger than a sail, over the fog and the blast / of the plush ocean, he goes..." (lines 8-9)

"...here are the shocked starlings pumping past" (line 6)

"Consider Icarus, pasting those sticky wings on." (line 1)

"...see how casually / he glances up and is caught..." (lines 10-11)

"...Who cares that he fell back to the sea?" (line 12)

"See him acclaiming the sun and come plunging down" (line 13)

Supporting Details

Figure 1. Released selected-response item from the Partnership for Assessment of Readiness for College and Careers ELA Assessment. Reproduced with fair-use permission granted by PARCC for educational purposes.

Item Example 1 - Item Profile	
Source: Partnership for Assessment of Readiness for College and Careers (PARCC) English Language Arts High School Sample Item Sets, accessed on on March 15, 2016 from: http://parcc.pearson.com/sample-items/	
Item Type	Selected Response (Technology-Enhanced)
Content Focus	Reading Literary Text
Grade Level	High School
Scoring	Computer
DOK Level	3
Bloom's Revised Taxonomy	Understand (2) and Analyze (4)
Hillocks & Ludlow Taxonomy of Skills (Fiction Only)	Author's Generalization (6) and Complex Implied Relationship (5)

Item Analysis

In this two-part, technology-enhanced, selected-response item, students must select from four choices the central idea of an entire text (Anne Sexton's poem "To a Friend Whose Work Has Come to Triumph"), and then select, from six choices, three supporting details that show how the selected central idea is developed over the course of the poem. The second part of the item is dependent on the first, requiring students to connect ideas and reason about the relationship between central ideas and supporting details in a text.

This item is part of a sequence of selected- response items that builds toward an analytical essay performance task (which will be featured later in this chapter). After answering two traditional selected-response items about Sexton's portrayal of the mythological character Icarus, students encounter this item in which they "drag and drop" their selected central idea from the list of "Possible Central Ideas" into the blank box, and then they repeat this process to select three supporting details from a list of "Possible Supporting Details." This item illustrates how the interaction of item format and item content can allow for measurement of DOK 3 skills and understanding while remaining computer scorable. While a constructed-response version of this item (in which no possible answers are provided) would be even more rigorous, and would eliminate the possibility of students guessing the correct answers, this item still calls for complex thinking. As Webb describes, students at DOK 3 "explain, generalize, or connect ideas. Standards and items at Level 3 involve reasoning and planning. Students must be able to support their thinking. Items may involve abstract theme identification, inference across an entire passage, or students' application of prior knowledge" (Webb, 2002). The drag-and-drop format used here, along with the requirement for a multi-part answer in which students select their three supporting details independently (rather than selecting from pre-formed

This item illustrates how the interaction of item format and item content can allow for measurement of DOK 3 skills and understanding while remaining computer scorable.

triad answer choices), provides more decision points and thus greater cognitive complexity for students than does the more traditional multiple-choice format. While we have described the item as DOK 3, some might view the item as actually straddling the fence between DOK 2 and 3 because each part of the item has only one correct answer. However, other plausible interpretations of this text *are* possible—they are simply not available as “correct” answers here because of the item format. What carries this item to DOK 3 is that it asks students to identify an abstract theme in a heavily figurative text, which increases the complexity in comparison to an item that asks students to identify the central idea of a more straightforward, informational text. This illustrates that the DOK level of an item is predicated on the complexity of both the content (e.g., interpreting literal vs. figurative language) and the required task (Webb, 2002; Hess, Carlock, Jones, and Walkup, 2009).

The identification of a central idea and supporting details is an important disciplinary skill on its own, but this item’s value is increased because of its relationship to the performance task (an essay) that follows it. PARCC’s ELA item sets are sequenced so that students analyze and interpret two related texts, first separately, then in relation to each other. The sequence culminates in a performance task that asks students to analyze some aspect of the relationship between the two texts. The early items in the sequence typically fall at DOK 2 or 3, while the culminating performance task measures the complex, extended thinking characteristic of DOK 4. This scaffolding allows students to gradually develop their understanding of each text before applying that depth of understanding in the analytical essay. In addition, this test construction strategy provides greater equity as well as greater measurement potential; without the scaffolding that this type of sequence provides, it would be difficult to assess true DOK 4 performances, which require extended strategic planning and thinking, in a standardized/on-demand setting.

Item Example 2 - Item Reflecting Feature 1

This item can be viewed on the CollegeBoard website at the following URL: https://secure-media.collegeboard.org/digitalServices/pdf/ap/ap15_frq_english_literature.pdf.

The scoring guidelines for this item can be found at: https://secure-media.collegeboard.org/digitalServices/pdf/ap/ap15_english_literature_sg.pdf

Item Example 2 - Item Profile	
Source: Advanced Placement Literature and Composition Free Response Question 1 (Poetry), 2015. The College Board, AP Central, Released Items. Accessed on March 15, 2016 from http://apcentral.collegeboard.com/apc/members/exam/exam_information/2002.html	
Item Type	Performance Task
Content Focus	Reading Literary Text
Grade Level	High School
Scoring	Human
DOK Level	3
Bloom's Revised Taxonomy	Analyze (4) and Create (6)
Hillocks & Ludlow Taxonomy of Skills (Fiction Only)	Structural Generalization (7)

Item Analysis

In this performance task, students must analyze a contemporary poem by the Nobel Prize winner Derek Walcott. They must identify the use of poetic devices (which are not named in the prompt) and analyze how those devices convey the significance of the narrator's experience.

This is a fairly typical example of an "author's craft" literary analysis question (Bloom's level 4) that also calls upon a student to write a well-developed essay in response (Bloom's level 6). Tackling this kind of complex literary analysis or close reading of a single, dense text is a core competency at higher levels of English language arts. There are multiple factors of complexity in this item:

- The text itself is challenging and requires analysis of poetic, figurative language that portrays a potentially unfamiliar cultural context (the Caribbean island of Saint Lucia).
- An analysis of the prompt reveals several implied questions a student must answer, all of which require that the student connect complex ideas: What is the experience represented in the poem? What is the significance of that experience? What poetic devices are used? How do they convey the significance of the experience?
- The evaluative criteria (which are not shown here but can be found at the same College Board URL listed above) reveal that the expectations for what constitutes a

"well-developed essay" are high and require the student to undertake a multi-step reading and writing process in a limited amount of time. The student must generate and narrow a compelling thesis in response to an open-ended prompt, marshal textual evidence, explain the significance of the evidence, and organize and express the explanation fluently.

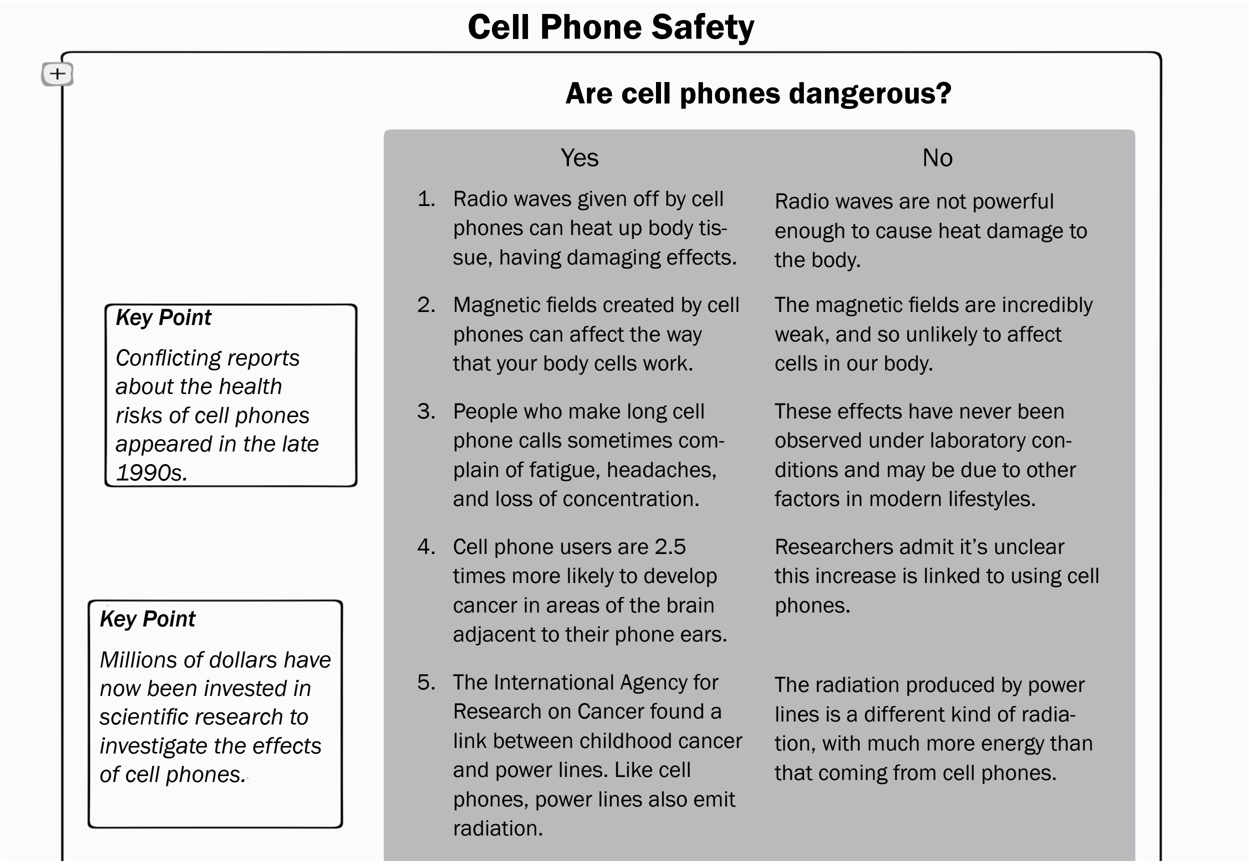
- This task demonstrates that high difficulty and even high complexity does not always correspond to DOK level 4. While the task does have multiple conditions (as outlined above), it does not require "non-routine manipulations across discipline/content areas/ multiple sources" (Webb, 2002). Therefore, this highly difficult and complex item represents DOK level 3.
- The level of difficulty in this task can be mediated without losing the disciplinary value of the item. An item like this could be adapted for lower grade levels by reducing the difficulty of the text itself, by pre-identifying some poetic devices that could be targeted for analysis, or by providing more scaffolding around what the prompt means by "the significance of the experience."

Feature 2: Items that measure cognitive and metacognitive competencies for reading literacy

We use the term "reading literacy" here rather than "reading" to emphasize that reading assessments must go beyond measures of students' ability to decode. This term is used similarly in the 2013 PISA (Program for International Student Assessment) Reading Literacy Framework, which explains that reading literacy includes a wide range of cognitive competencies such as finding, selecting, evaluating, and interpreting information, as well as "metacognitive competencies: the awareness of and ability to use a variety of appropriate strategies when processing texts. Metacognitive competencies are activated when readers think about, monitor and adjust their reading activity for a particular goal" (OECD, 2013).

The development and measurement of these megacognitive competencies supports the Common Core State Standards' third Key Shift in English Language Arts: Building knowledge through content-rich nonfiction. If students are to "build knowledge through texts so they can learn independently," they need much more than "extensive opportunities" to read those texts (CCSSO, 2010)—they need strategies and metacognitive awareness that will enable them to construct meaning from a wide range of texts for personal, academic, civic, and professional purposes.

Item Example 3 - Item Reflecting Feature 2



Cell Phone Safety

Are cell phones dangerous?

Key Point

Conflicting reports about the health risks of cell phones appeared in the late 1990s.

Key Point

Millions of dollars have now been invested in scientific research to investigate the effects of cell phones.

	Yes	No
1.	Radio waves given off by cell phones can heat up body tissue, having damaging effects.	Radio waves are not powerful enough to cause heat damage to the body.
2.	Magnetic fields created by cell phones can affect the way that your body cells work.	The magnetic fields are incredibly weak, and so unlikely to affect cells in our body.
3.	People who make long cell phone calls sometimes complain of fatigue, headaches, and loss of concentration.	These effects have never been observed under laboratory conditions and may be due to other factors in modern lifestyles.
4.	Cell phone users are 2.5 times more likely to develop cancer in areas of the brain adjacent to their phone ears.	Researchers admit it's unclear this increase is linked to using cell phones.
5.	The International Agency for Research on Cancer found a link between childhood cancer and power lines. Like cell phones, power lines also emit radiation.	The radiation produced by power lines is a different kind of radiation, with much more energy than that coming from cell phones.

Question intent: Reflect and evaluate

Text format: Non-continuous

Item Stem:

"It is difficult to prove that one thing has definitely caused another."

What is the relationship of this piece of information to the Point 4 ***Yes and No*** *statements in the table* ***Are cell phones dangerous?***

A It supports the Yes argument but does not prove it.
B It proves the Yes argument.
C It supports the No argument but does not prove it.
D It shows that the No argument is wrong.

SCORING:
Correct
Answer C. It supports the No argument but does not prove it.
Incorrect
Other responses.

Figure 2. Released ELA selected-response item from the Program for International Student Assessment, Reading Literacy 2009. Item in the public domain. Reproduced from nces.ed.gov, permission to excerpt OECD copyrighted materials for fair use purposes granted by OECD.

Item Example 3 - Item Profile	
Source: © 2009. Organization for Economic Co-operation and Development (OECD), PISA Released Paper-Based Assessment Items, Reading Literacy 2009, accessed on March 15, 2016 from https://nces.ed.gov/surveys/pisa/educators.asp	
Item Type	Selected Response
Content Focus	Reading Informational Text
Grade Level	8
Scoring	Computer
DOK Level	2
Bloom's Revised Taxonomy	Analyze (4)
Hillocks & Ludlow Taxonomy of Skills (Fiction Only)	N/A

Item Analysis

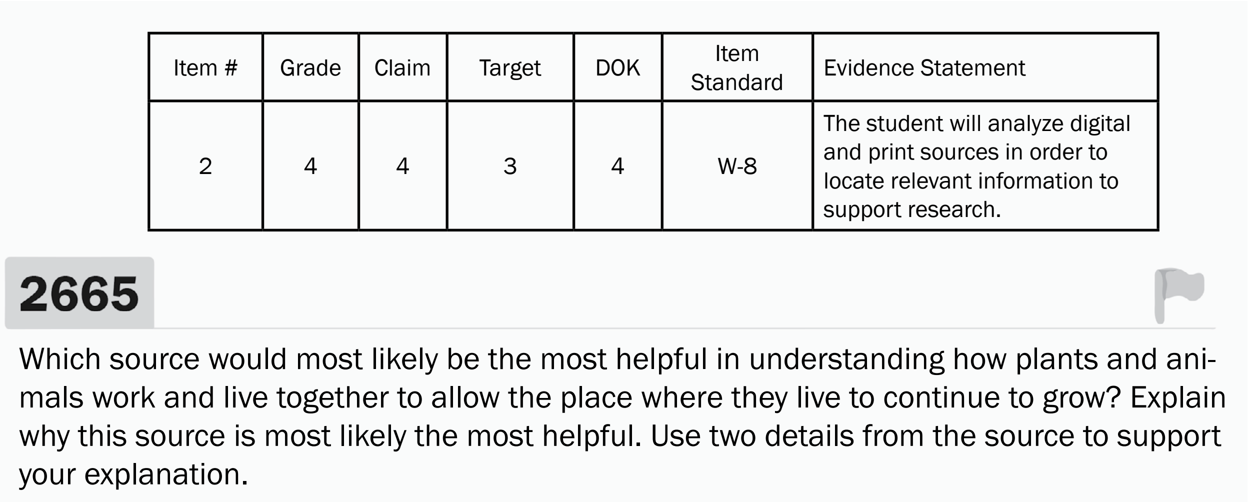
This selected-response item focuses on assessing students' literacy skills rather than their understanding of the content of a particular text. This item is part of a block of items with a similar focus—for example, one related item asks students to consider the purpose of textual features like the "Key Points" in the left margin of the text. This item asks students to reason about and explain the relationship among multiple pieces of information in a mixed-format text, calling upon conceptual understanding of the difference between support and proof for a claim. This emphasis on metacognitive competencies rather than simply decoding the text lends this item its cognitive rigor, while the specific pairing of skill and understanding measured in the item lends it authenticity. The skill of determining how ideas interact in a text to produce meaning, as well as an understanding of what constitutes support versus proof, are key to processing and drawing conclusions from complex informational texts. The skills addressed in this item are authentic and relevant to "the full scope of situations in which reading literacy plays a role, from private to public, from school to work, from formal education to lifelong learning and active citizenship" (OECD, 2013).

The skill of determining how ideas interact in a text to produce meaning, as well as an understanding of what constitutes support versus proof, are key to processing and drawing conclusions from complex informational texts.

This item could be stronger and more coherent still if the "piece of information" presented in the item stem ("It is difficult to prove that one thing has definitely caused another") were taken directly from the text or otherwise contextualized as an important principle for drawing conclusions from evidence. The lack of context could lead to some confusion on the part of students trying to answer the question. Also, students may find the use of the word "point" in the item stem confusing considering that there are text features labeled "Key Points." Nevertheless, in its content, the item is a strong example of a selected-response item that measures key disciplinary skills that are transferable across many texts and contexts.

The item is also potentially educative: it highlights that drawing conclusions from informational text is not simply a process of summarizing or comprehending the text as a whole but rather of making sense of differing and potentially conflicting information using reasoning and evidence.

Item Example 4 — Item Reflecting Feature 2



Item #	Grade	Claim	Target	DOK	Item Standard	Evidence Statement
2	4	4	3	4	W-8	The student will analyze digital and print sources in order to locate relevant information to support research.

2665

Which source would most likely be the most helpful in understanding how plants and animals work and live together to allow the place where they live to continue to grow? Explain why this source is most likely the most helpful. Use two details from the source to support your explanation.

Figure 3. Released ELA constructed-response item from a Smarter Balanced Assessment Consortium performance task. Reproduced with permission granted by the Regents of the University of California.

Example Item 4 — Item Profile	
Source: Smarter Balanced Assessment Consortium (SBAC), ELA Grade 4 Research Question (Item 2665), Practice Performance Task, accessed on March 15, 2016 from http://sbac.portal.airast.org/wp-content/uploads/2013/08/ELA_Practice_Test_Scoring_Guide_Grade_4_PT.pdf	
Item Type	Constructed Response
Content Focus	Reading Informational Text
Grade Level	4
Scoring	Human
DOK Level	3
Bloom's Revised Taxonomy	Evaluate (5)
Hillocks & Ludlow Taxonomy of Skills (Fiction Only)	N/A

Item Analysis

This constructed-response item asks students to select the most relevant source for a stated purpose and to justify their selection with multiple pieces of evidence from the source. There are several layers of complexity in this item. While all three of the sources (not shown here) provided in the task address the same topic, each has a slightly different purpose and focus. The student must consider the relevance of the specific information in each source to determine which would be most appropriate for the given purpose.

Requiring students to justify their selection in writing with textual evidence increases the cognitive demand significantly and provides a much more accurate measurement of the student's ability to perform this extremely important cross-disciplinary research skill.

To answer this question successfully, students must compare multiple sources, evaluating each against the criteria in the item stem. Once they have determined which source would be most helpful for the given purpose, they must select and explain textual details from that source to justify their selection. This item measures a metacognitive competency that is extremely important in the context of research, and the work a student must do to complete the item mimics, to some degree, the act of writing an annotated bibliography, which is an authentic application of the skill of evaluating sources.

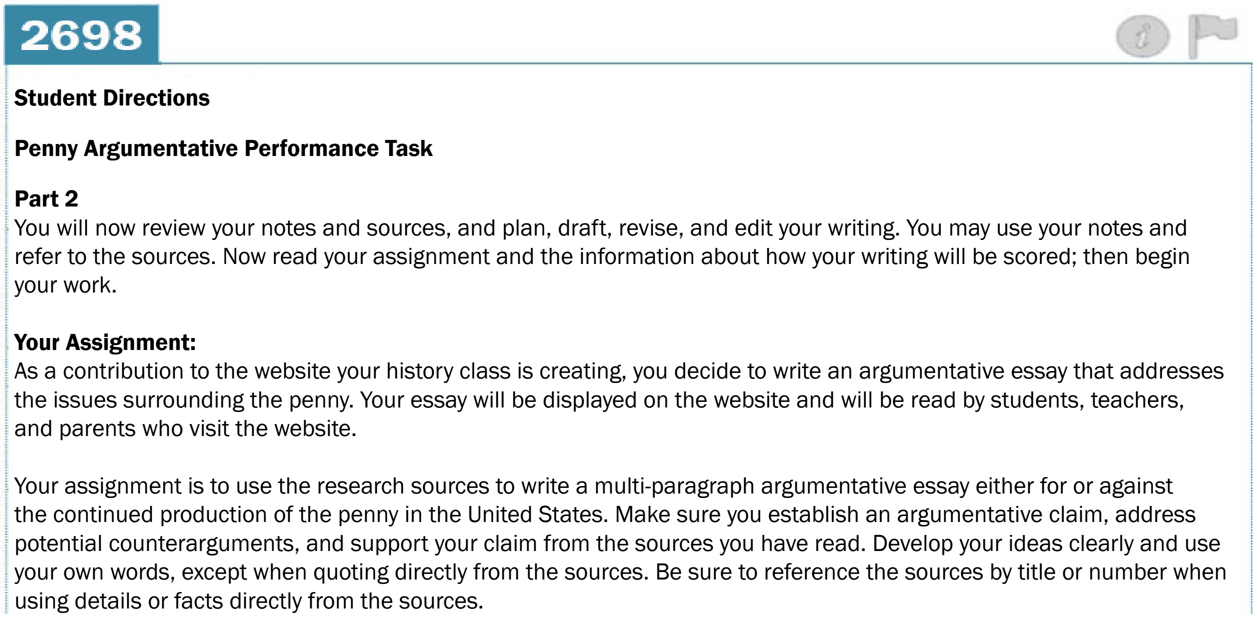
Evaluation of sources could be assessed with a selected-response format, but such an item would be limited to DOK 2. Requiring students to justify their selection in writing with textual evidence increases the cognitive demand significantly and provides a much more accurate measurement of the student's ability to perform this extremely important cross-disciplinary research skill. While the Smarter Balanced metadata identifies this as a DOK level 4 item because it involves multiple sources, we believe it is truly DOK level 3 since students are not synthesizing across sources but rather selecting from among multiple sources.

Feature 3: Items that Require Evidence-Based Writing

For items with a focus on writing, we looked for prompts that called upon students to use textual evidence to develop an explanation, argument, or narrative rather than drawing solely on students' opinions or experiences. We also sought items in which student work was evaluated with criteria that examine the substance of the student's composition (e.g., strength of claims, logical reasoning, use and explanation of textual evidence), going beyond surface-level and structural components of writing (e.g., syntactic variety, use of transitions).

Item Example 5 — Item Reflecting Feature 3

Item #	Grade	Claim	Target	DOK	Item Standard	Evidence Statement
4	8	2	7	4	W-1.a	The student will write full arguments about topics or texts, attending to purpose and audience: establish and support a claim, organize and cite supporting (text) evidence from credible sources, and develop a conclusion that is appropriate to purpose and audience and follows and supports the argument(s) presented.



Student Directions

Penny Argumentative Performance Task

Part 2
You will now review your notes and sources, and plan, draft, revise, and edit your writing. You may use your notes and refer to the sources. Now read your assignment and the information about how your writing will be scored; then begin your work.

Your Assignment:
As a contribution to the website your history class is creating, you decide to write an argumentative essay that addresses the issues surrounding the penny. Your essay will be displayed on the website and will be read by students, teachers, and parents who visit the website.

Your assignment is to use the research sources to write a multi-paragraph argumentative essay either for or against the continued production of the penny in the United States. Make sure you establish an argumentative claim, address potential counterarguments, and support your claim from the sources you have read. Develop your ideas clearly and use your own words, except when quoting directly from the sources. Be sure to reference the sources by title or number when using details or facts directly from the sources.

Figure 5. Released ELA performance task from the Smarter Balanced Assessment Consortium. Reproduced with permission granted by the Regents of the University of California.

Item Example 5 — Item Profile	
Source: Smarter Balanced Practice Performance Tasks, accessed via http://sbac.portal.airast.org/wp-content/uploads/2013/08/ELA_Practice_Test_Scoring_Guide_Grade_8_PT.pdf	
Item Type	Performance Task
Content Focus	Writing
Grade Level	8
Scoring	Human
DOK Level	4
Bloom's Revised Taxonomy	Analyze (4), Evaluate (5), and Create (6)
Hillocks & Ludlow	N/A

Item Analysis

This simulated research performance task asks students to develop and support an argumentative claim based on evidence from multiple sources. This DOK level 4 task goes beyond summarizing information from multiple sources to address a topic (which would be DOK 3) and asks students to synthesize information from multiple sources, examine and refute alternate perspectives, and develop a substantiated claim in response to an open-ended prompt (DOK 4).

Although this task requires research skills, the evaluative criteria (not shown here but available at the item source URL above) focus primarily on writing. However, the rubrics acknowledge that the quality of writing is not solely related to whether students can apply a standard organizational structure and express ideas clearly; the rubric requires that the writing be both substantive, (i.e., consisting primarily of ideas based on research, evidence, and elaboration/explanation), and organized in a way that supports the purpose.

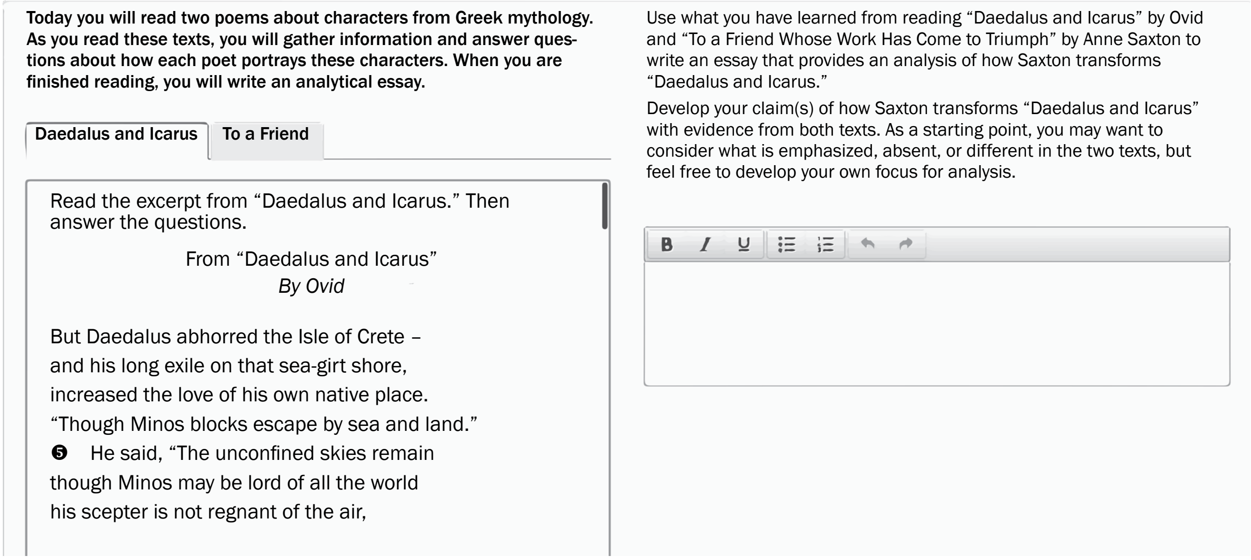
> **To develop a substantiated claim, students must reason from the information in the sources rather than simply repeat the opinions of others. For argumentative tasks to genuinely measure students' ability to develop and substantiate a claim, it is extremely important that the sources provide this balance of perspectives and that they do not "provide the answer" for students...**

This item is also strong because of its authenticity; even though some students may not be personally interested in the topic, the task addresses a question that is current and genuinely debatable. The sources also contribute to the item's authenticity as an argumentation task; while some of the sources express opinions on the topic (both for and against the idea of eliminating the penny), most simply provide information that could be used to support a claim on either side of the debate. To develop a substantiated claim, students must reason from the information in the sources rather than simply repeat the opinions of others. For argumentative tasks to genuinely measure students' ability to develop and substantiate a claim, it is extremely important that the sources provide this balance of perspectives and that they do not "provide the answer" for students, thus turning a synthesis task into a summary task and significantly reducing the item's cognitive complexity. It is also important to note that considering the prompt alone does not reveal whether the task is genuinely open-ended; the sources must be evaluated as well.

Feature 4: Items that Call for Authentic Disciplinary Uses of Literacy

The world of education is now experiencing significant agreement that literacy should be supported and applied across the disciplines and for a wide variety of contexts and purposes, both academic and non-academic. There is less agreement, however, about what types of questions, tasks, and topics are considered authentic or central to the discipline of English language arts. As such, many ELA items focus on highly transferable skills and concepts rather than those that are unique to the discipline. While this is not inherently a bad thing, we believe ELA assessments should also represent core ELA concepts and ways of thinking—e.g., that language has cultural, social, and personal power; that literature both reflects and plays a role in shaping culture; and that readers construct meaning from both text and context, including relationships among texts.

Item Example 6 — Item Reflecting Feature 4



Today you will read two poems about characters from Greek mythology. As you read these texts, you will gather information and answer questions about how each poet portrays these characters. When you are finished reading, you will write an analytical essay.

Read the excerpt from “Daedalus and Icarus.” Then answer the questions.

From “Daedalus and Icarus”
By Ovid

But Daedalus abhorred the Isle of Crete –
and his long exile on that sea-girt shore,
increased the love of his own native place.
“Though Minos blocks escape by sea and land.”
❺ He said, “The unconfined skies remain
though Minos may be lord of all the world
his scepter is not regnant of the air,

Use what you have learned from reading “Daedalus and Icarus” by Ovid and “To a Friend Whose Work Has Come to Triumph” by Anne Saxton to write an essay that provides an analysis of how Saxton transforms “Daedalus and Icarus.”

Develop your claim(s) of how Saxton transforms “Daedalus and Icarus” with evidence from both texts. As a starting point, you may want to consider what is emphasized, absent, or different in the two texts, but feel free to develop your own focus for analysis.

[Scroll down to read poem, use tabs to toggle between poems]

Figure 5. Released PARCC ELA performance task. Reproduced with fair-use permission granted by the PARCC for educational purposes.

Item Example 6 — Item Profile	
Source: Partnership for Assessment of Readiness for College and Careers ELA Literary Analysis Task, High School Sample Item Sets, accessed on March 16, 2016 from http://parcc.pearson.com/sample-items/	
Item Type	Performance Task
Content Focus	Writing
Grade Level	High School
Scoring	Human
DOK Level	4
Bloom’s Revised Taxonomy	Analyze (4) and Create/synthesize across multiple texts (6)
Hillocks & Ludlow Taxonomy of Skills (Fiction Only)	Complex Implied Relationship (5) and Structural Generalization (7)

Item Analysis

This item, a performance task related to Item 1 in this paper, represents a type of conceptual thinking that is deeply central to the discipline of English language arts—intertextuality. The item highlights how the meaning of texts is sometimes predicated upon their relationship with other texts. The two texts represented in this performance task demonstrate a common form of intertextuality, in which a contemporary text builds upon and transforms a classical text. An item such as this honors the humanities dimension of ELA by inviting students to explore how the archetypes and myths that underlie our culture remain constant but also how they change with time.

This item also demonstrates how a performance task with a focus on literary interpretation can draw on ways of thinking about text other than the more common analysis of author's craft, such as that required by Item Example 2 in this paper. That kind of fine-grain analysis of textual elements is certainly valued in the discipline, but it is not the only way to think about how meaning is constructed in literary text. This item more closely represents the kinds of discussion and writing assignments one would see in a college classroom.

The cognitive complexity of the item is high. It requires analysis of complex implied relationships (identifying key similarities and differences across the two texts), synthesis (making a claim about how the more contemporary text "transforms" the classic text), and integration of evidence from both sources to justify and elaborate the response. The prompt is open-ended; while it provides some suggestions for narrowing the focus of the response, it ultimately requires students to develop their own focus for analysis, which adds a layer of decision-making and cognitive complexity.

The two texts represented in this performance task demonstrate a common form of intertextuality, in which a contemporary text builds upon and transforms a classical text. An item such as this honors the humanities dimension of ELA by inviting students to explore how the archetypes and myths that underlie our culture remain constant but also how they change with time.

Note that the prompt includes the instruction to "use what you have learned from reading 'Daedalus and Icarus' by Ovid and 'To a Friend Whose Work Has Come to Triumph' by Anne Sexton. Part of what lends this item its deeper measurement capacity is the way it builds on earlier items that ask students to analyze aspects of each text separately (as in Item Example 1 in this paper) before gradually considering their relationship to each other (for example, an earlier item asks, "Which statement summarizes a key difference between the excerpt from the poem by Ovid and the poem by Sexton?"). The sequence culminates in this analysis of how the more contemporary text transforms some aspect of the classical text. This careful sequencing, designed to support students in developing a coherent understanding of the texts, differs significantly from the common testing practice of using a single text as the stimulus for a wide variety of unrelated questions (e.g., a few vocabulary-in-context questions, a few literal comprehension questions, and a few inference questions). While the latter practice is more efficient for measuring numerous standards with a single stimulus, it misses the opportunity to engage students in a progression of increasingly complex items. Furthermore, it does not reflect how we engage with texts in real life and in English classrooms; it places text in a subservient position to the test itself rather than using the test questions to support coherent understanding of text.

References

Applebee, A. N. (1993). *Literature in the secondary school: Studies of curriculum and instruction in the United States.* Urbana, IL: National Council of Teachers of English.

Applebee, A. N., & Langer, J. A. (2013). *Writing instruction that works: Proven methods for middle and high school classrooms.* New York, NY: Teachers College Press.

Anderson, L. W., and Krathwohl, D. R., et al. (Eds.) (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives.* New York, NY: Longman.

Calkins, L. M. (1980). When children want to punctuate: Basic skills belong in context. *Language Arts*, *57*, 567-73.

DiStefano, P., & Killion, J. (1984). Assessing writing skills through a process approach. *English Education*, *16*(4), 203-7.

Harris, R. J. (1962). *An experimental enquiry into the functions and value of formal grammar in the teaching of written English, with special reference to the teaching of correct written English to children aged twelve to fourteen (Doctoral dissertation)*. Library of Institute of Education, London University.

Hess, K. K. (2009, updated 2013). *Linking research with practice: A local assessment toolkit to guide school leaders*. Underhill, VT: Author.

Hess, K. K., Carlock, D., Jones, B. S., & Walkup, J. R. (2009). *What exactly do "fewer, clearer, and higher standards" really look like in the classroom? Using a cognitive rigor matrix to analyze curriculum, plan lessons, and implement assessments*. Retrieved from http://www.nciea.org/cgi-bin/pubspage.cgi?sortby=pub_date

Hillocks, G. (1986). *Research on written composition: New directions for teaching*. Urbana, IL: National Conference on Research in English/Educational Resources Information Center.

Hillocks, G., & Ludlow, L. H. (1984). A taxonomy of skills in reading and interpreting fiction. *American Educational Research Journal*, *21*(1), 7–24. Retrieved from http://www.jstor.org/stable/1162351.

Hillocks, G., & Smith, M. W. (1991). Grammar and usage. In J. Flood, J.M. Jensen, D. Lapp, & J.R. Squire (Eds.), *Handbook of research on teaching the English language arts* (pp. 591-603). New York, NY: Macmillan.

Lynch, J. J. & Evans, B. (1963). *High school English textbooks: A critical examination*. Boston, MA: Little, Brown, and Company.

Mihalakis, V. & Petrosky, T. (2015). *Collaborative professional development to create cognitively demanding tasks in English language arts.* In J. A. Supovitz & J. Spillane (Eds.), Challenging standards: Navigating conflict and building capacity in the era of the common core. Lanham, MD: Rowman & Littlefield.

National Council of Teachers of English & the International Reading Association. (1996). *Standards for the English Language Arts.* Urbana, IL: National Council of Teachers of English.

National Governors Association Center for Best Practices & Council of Chief State School Officers. (2010). *Common Core State Standards for English language arts and literacy in history/social studies, science, and technical subjects*. Washington, DC: Authors.

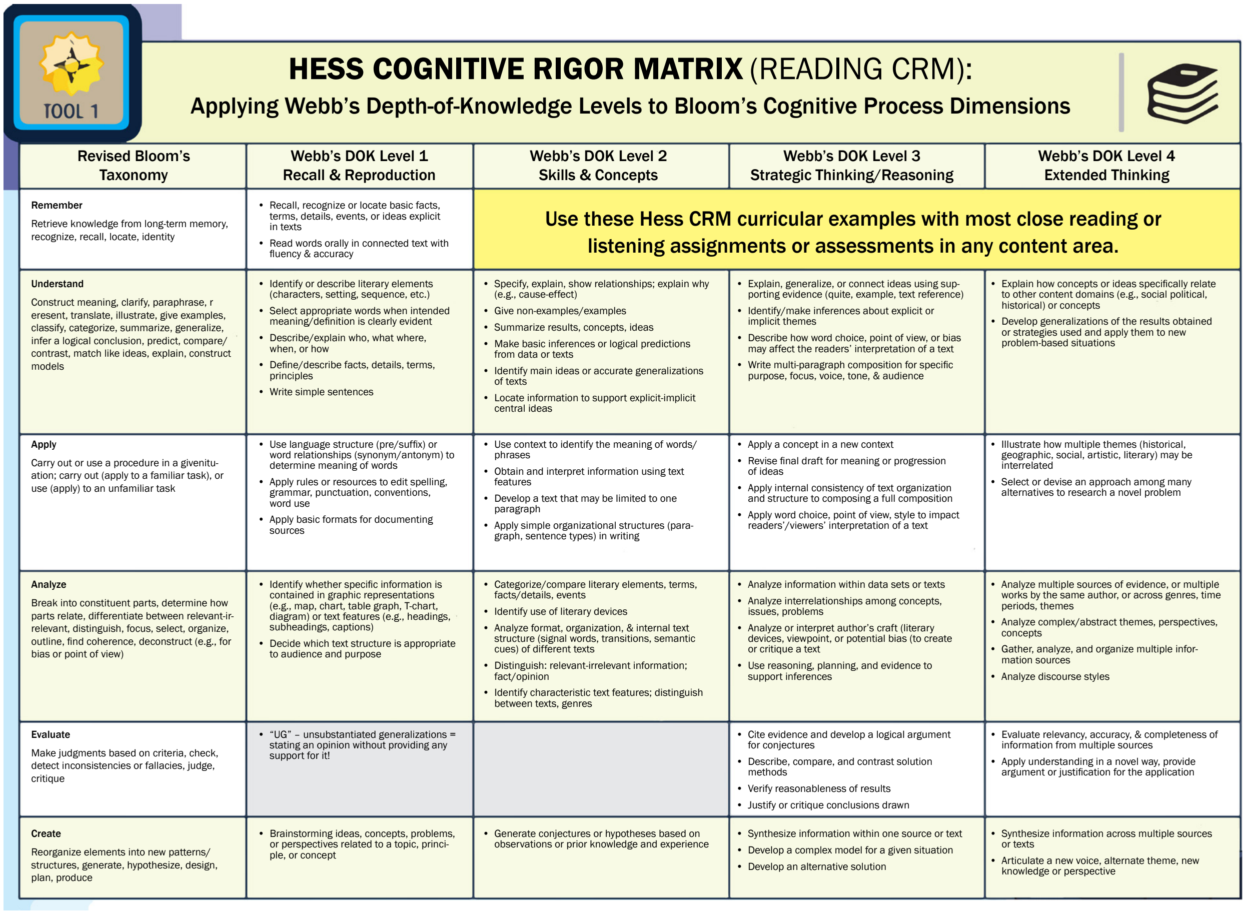
Organization for Economic Cooperation and Development (OECD) (2013). *PISA 2012 Assessment and analytical framework: Mathematics, reading, science, problem solving and financial literacy*, OECD Publishing.

Weaver, C. (1996). *Teaching grammar in context.* Portsmouth, NH: Boynton/Cook.

Webb, N. L. (2002). *Depth-of-knowledge levels for four content areas.* Unpublished paper.

Appendix A

Hess Cognitive Rigor Matrix for Close Reading Across Content Areas



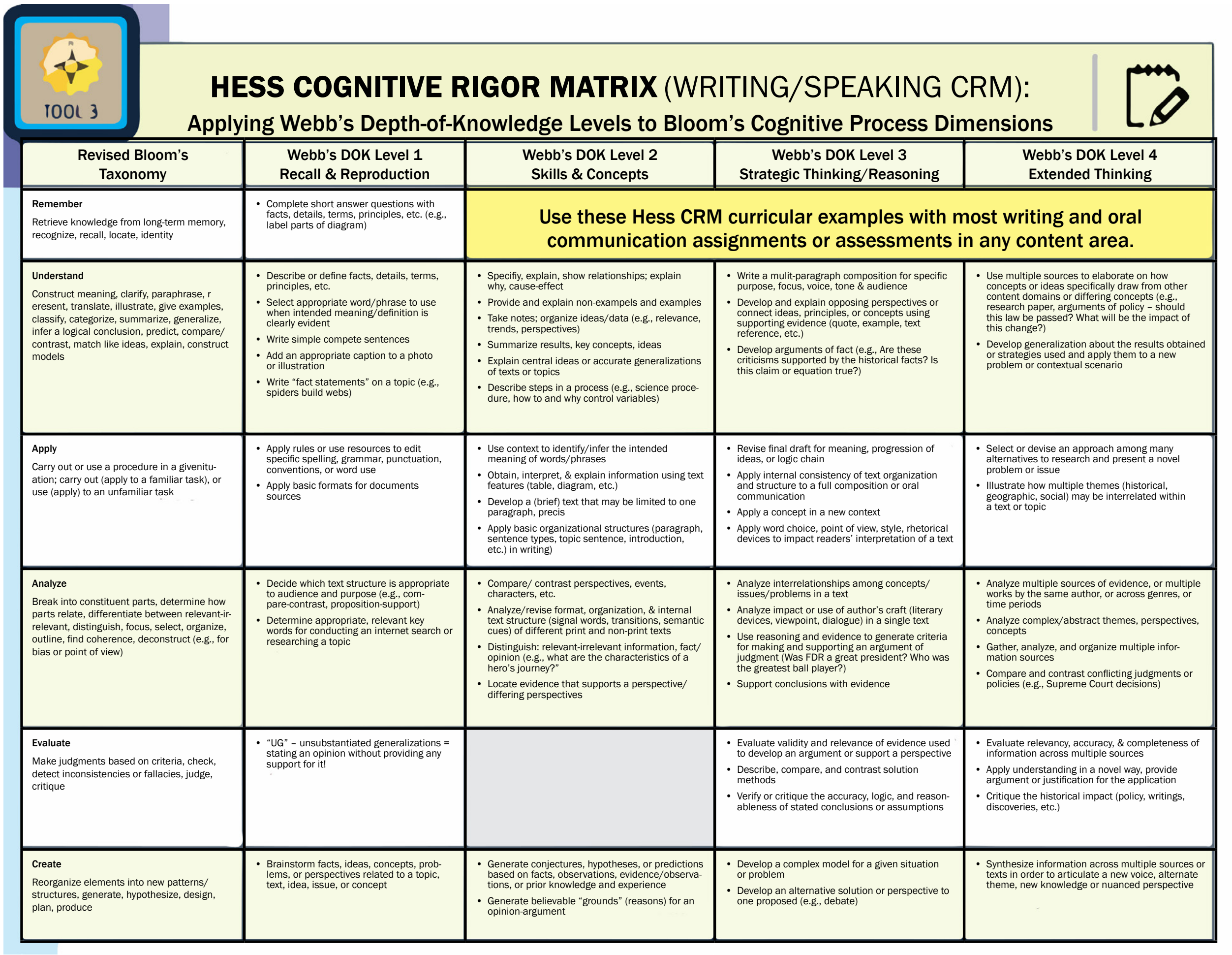
TOOL 1

HESS COGNITIVE RIGOR MATRIX (READING CRM):

Applying Webb's Depth-of-Knowledge Levels to Bloom's Cognitive Process Dimensions

Revised Bloom's Taxonomy	Webb's DOK Level 1 Recall & Reproduction	Webb's DOK Level 2 Skills & Concepts	Webb's DOK Level 3 Strategic Thinking/Reasoning	Webb's DOK Level 4 Extended Thinking
Remember Retrieve knowledge from long-term memory, recognize, recall, locate, identity	• Recall, recognize or locate basic facts, terms, details, events, or ideas explicit in texts • Read words orally in connected text with fluency & accuracy	**Use these Hess CRM curricular examples with most close reading or listening assignments or assessments in any content area.**		
Understand Construct meaning, clarify, paraphrase, r eresent, translate, illustrate, give examples, classify, categorize, summarize, generalize, infer a logical conclusion, predict, compare/contrast, match like ideas, explain, construct models	• Identify or describe literary elements (characters, setting, sequence, etc.) • Select appropriate words when intended meaning/definition is clearly evident • Describe/explain who, what where, when, or how • Define/describe facts, details, terms, principles • Write simple sentences	• Specify, explain, show relationships; explain why (e.g., cause-effect) • Give non-examples/examples • Summarize results, concepts, ideas • Make basic inferences or logical predictions from data or texts • Identify main ideas or accurate generalizations of texts • Locate information to support explicit-implicit central ideas	• Explain, generalize, or connect ideas using supporting evidence (quite, example, text reference) • Identify/make inferences about explicit or implicit themes • Describe how word choice, point of view, or bias may affect the readers' interpretation of a text • Write multi-paragraph composition for specific purpose, focus, voice, tone, & audience	• Explain how concepts or ideas specifically relate to other content domains (e.g., social political, historical) or concepts • Develop generalizations of the results obtained or strategies used and apply them to new problem-based situations
Apply Carry out or use a procedure in a givenituation; carry out (apply to a familiar task), or use (apply) to an unfamiliar task	• Use language structure (pre/suffix) or word relationships (synonym/antonym) to determine meaning of words • Apply rules or resources to edit spelling, grammar, punctuation, conventions, word use • Apply basic formats for documenting sources	• Use context to identify the meaning of words/phrases • Obtain and interpret information using text features • Develop a text that may be limited to one paragraph • Apply simple organizational structures (paragraph, sentence types) in writing	• Apply a concept in a new context • Revise final draft for meaning or progression of ideas • Apply internal consistency of text organization and structure to composing a full composition • Apply word choice, point of view, style to impact readers'/viewers' interpretation of a text	• Illustrate how multiple themes (historical, geographic, social, artistic, literary) may be interrelated • Select or devise an approach among many alternatives to research a novel problem
Analyze Break into constituent parts, determine how parts relate, differentiate between relevant-irrelevant, distinguish, focus, select, organize, outline, find coherence, deconstruct (e.g., for bias or point of view)	• Identify whether specific information is contained in graphic representations (e.g., map, chart, table graph, T-chart, diagram) or text features (e.g., headings, subheadings, captions) • Decide which text structure is appropriate to audience and purpose	• Categorize/compare literary elements, terms, facts/details, events • Identify use of literary devices • Analyze format, organization, & internal text structure (signal words, transitions, semantic cues) of different texts • Distinguish: relevant-irrelevant information; fact/opinion • Identify characteristic text features; distinguish between texts, genres	• Analyze information within data sets or texts • Analyze interrelationships among concepts, issues, problems • Analyze or interpret author's craft (literary devices, viewpoint, or potential bias (to create or critique a text • Use reasoning, planning, and evidence to support inferences	• Analyze multiple sources of evidence, or multiple works by the same author, or across genres, time periods, themes • Analyze complex/abstract themes, perspectives, concepts • Gather, analyze, and organize multiple information sources • Analyze discourse styles
Evaluate Make judgments based on criteria, check, detect inconsistencies or fallacies, judge, critique	• "UG" – unsubstantiated generalizations = stating an opinion without providing any support for it!		• Cite evidence and develop a logical argument for conjectures • Describe, compare, and contrast solution methods • Verify reasonableness of results • Justify or critique conclusions drawn	• Evaluate relevancy, accuracy, & completeness of information from multiple sources • Apply understanding in a novel way, provide argument or justification for the application
Create Reorganize elements into new patterns/structures, generate, hypothesize, design, plan, produce	• Brainstorming ideas, concepts, problems, or perspectives related to a topic, principle, or concept	• Generate conjectures or hypotheses based on observations or prior knowledge and experience	• Synthesize information within one source or text • Develop a complex model for a given situation • Develop an alternative solution	• Synthesize information across multiple sources or texts • Articulate a new voice, alternate theme, new knowledge or perspective

Hess Cognitive Rigor Matrix for Written and Oral Communication



TOOL 3

HESS COGNITIVE RIGOR MATRIX (WRITING/SPEAKING CRM):

Applying Webb's Depth-of-Knowledge Levels to Bloom's Cognitive Process Dimensions

Revised Bloom's Taxonomy	Webb's DOK Level 1 Recall & Reproduction	Webb's DOK Level 2 Skills & Concepts	Webb's DOK Level 3 Strategic Thinking/Reasoning	Webb's DOK Level 4 Extended Thinking
Remember Retrieve knowledge from long-term memory, recognize, recall, locate, identity	• Complete short answer questions with facts, details, terms, principles, etc. (e.g., label parts of diagram)	**Use these Hess CRM curricular examples with most writing and oral communication assignments or assessments in any content area.**		
Understand Construct meaning, clarify, paraphrase, r eresent, translate, illustrate, give examples, classify, categorize, summarize, generalize, infer a logical conclusion, predict, compare/ contrast, match like ideas, explain, construct models	• Describe or define facts, details, terms, principles, etc. • Select appropriate word/phrase to use when intended meaning/definition is clearly evident • Write simple compete sentences • Add an appropriate caption to a photo or illustration • Write "fact statements" on a topic (e.g., spiders build webs)	• Specifiy, explain, show relationships; explain why, cause-effect • Provide and explain non-exampels and examples • Take notes; organize ideas/data (e.g., relevance, trends, perspectives) • Summarize results, key concepts, ideas • Explain central ideas or accurate generalizations of texts or topics • Describe steps in a process (e.g., science procedure, how to and why control variables)	• Write a mulit-paragraph composition for specific purpose, focus, voice, tone & audience • Develop and explain opposing perspectives or connect ideas, principles, or concepts using supporting evidence (quote, example, text reference, etc.) • Develop arguments of fact (e.g., Are these criticisms supported by the historical facts? Is this claim or equation true?)	• Use multiple sources to elaborate on how concepts or ideas specifically draw from other content domains or differing concepts (e.g., research paper, arguments of policy – should this law be passed? What will be the impact of this change?) • Develop generalization about the results obtained or strategies used and apply them to a new problem or contextual scenario
Apply Carry out or use a procedure in a givenituation; carry out (apply to a familiar task), or use (apply) to an unfamiliar task	• Apply rules or use resources to edit specific spelling, grammar, punctuation, conventions, or word use • Apply basic formats for documents sources	• Use context to identify/infer the intended meaning of words/phrases • Obtain, interpret, & explain information using text features (table, diagram, etc.) • Develop a (brief) text that may be limited to one paragraph, precis • Apply basic organizational structures (paragraph, sentence types, topic sentence, introduction, etc.) in writing)	• Revise final draft for meaning, progression of ideas, or logic chain • Apply internal consistency of text organization and structure to a full composition or oral communication • Apply a concept in a new context • Apply word choice, point of view, style, rhetorical devices to impact readers' interpretation of a text	• Select or devise an approach among many alternatives to research and present a novel problem or issue • Illustrate how multiple themes (historical, geographic, social) may be interrelated within a text or topic
Analyze Break into constituent parts, determine how parts relate, differentiate between relevant-irrelevant, distinguish, focus, select, organize, outline, find coherence, deconstruct (e.g., for bias or point of view)	• Decide which text structure is appropriate to audience and purpose (e.g., compare-contrast, proposition-support) • Determine appropriate, relevant key words for conducting an internet search or researching a topic	• Compare/ contrast perspectives, events, characters, etc. • Analyze/revise format, organization, & internal text structure (signal words, transitions, semantic cues) of different print and non-print texts • Distinguish: relevant-irrelevant information, fact/ opinion (e.g., what are the characteristics of a hero's journey?" • Locate evidence that supports a perspective/ differing perspectives	• Analyze interrelationships among concepts/ issues/problems in a text • Analyze impact or use of author's craft (literary devices, viewpoint, dialogue) in a single text • Use reasoning and evidence to generate criteria for making and supporting an argument of judgment (Was FDR a great president? Who was the greatest ball player?) • Support conclusions with evidence	• Analyze multiple sources of evidence, or multiple works by the same author, or across genres, or time periods • Analyze complex/abstract themes, perspectives, concepts • Gather, analyze, and organize multiple information sources • Compare and contrast conflicting judgments or policies (e.g., Supreme Court decisions)
Evaluate Make judgments based on criteria, check, detect inconsistencies or fallacies, judge, critique	• "UG" – unsubstantiated generalizations = stating an opinion without providing any support for it!		• Evaluate validity and relevance of evidence used to develop an argument or support a perspective • Describe, compare, and contrast solution methods • Verify or critique the accuracy, logic, and reasonableness of stated conclusions or assumptions	• Evaluate relevancy, accuracy, & completeness of information across multiple sources • Apply understanding in a novel way, provide argument or justification for the application • Critique the historical impact (policy, writings, discoveries, etc.)
Create Reorganize elements into new patterns/ structures, generate, hypothesize, design, plan, produce	• Brainstorm facts, ideas, concepts, problems, or perspectives related to a topic, text, idea, issue, or concept	• Generate conjectures, hypotheses, or predictions based on facts, observations, evidence/observations, or prior knowledge and experience • Generate believable "grounds" (reasons) for an opinion-argument	• Develop a complex model for a given situation or problem • Develop an alternative solution or perspective to one proposed (e.g., debate)	• Synthesize information across multiple sources or texts in order to articulate a new voice, alternate theme, new knowledge or nuanced perspective

Hess, K.K. (2009, updated 2013). Linking research with practice: A local assessment toolkit to guide school leaders. Underhill, VT: author.

Appendix B

Taxonomy of Skills in Reading and Interpreting Fiction

This excerpt from the 2009 NCTE publication "Writing about Literature" explains and illustrates each level of Hillocks and Ludlow's taxonomy with sample questions based on Chapter 1 of *The Pearl* (1972) by John Steinbeck. These questions comprise one of the four question sets examined in Hillocks and Ludlow's 1984 study.

Literal Level of Comprehension

1. **Basic Stated Information**—Identifying frequently stated information that presents some condition crucial to the story.
   Example: What happened to Coyotito?

2. **Key Detail**—Identifying a detail that appears at some key juncture of the plot and bears a causal relationship to what happens.
   Example: Where did Coyotito sleep?

3. **Stated Relationship**—Identifying a statement that explains the relationship between at least two pieces of information in the text.
   Example: What was the beggars' reason for following Kino and Juana to the doctor's house?

Inferential Level of Comprehension

4. **Simple Implied Relationship**—Inferring the relationship between two pieces of information usually closely juxtaposed in the text.
   Example: What were Kino's feelings about the pearls he offers the doctor? Explain how you know.

5. **Complex Implied Relationship**—Inferring the relationship(s) among many pieces of information spread throughout large parts of the text. A question of this type might concern, for example, the causes of character change. This would involve relating details of personality before and after a change and inferring the causes of the change from the same details and intervening events.
   Example: In this chapter, Kino appears at home and in town. He feels and acts differently in these two places. Part of the difference is the result of what happened to Coyotito. Part is the result of other things.
   (a) What are the differences between the way Kino acts and feels at home and in town? (b) Apart from what happened to Coyotito, explain the causes of those differences.

6. **Author's Generalization**—Inferring a generalization about the world outside of the work from the fabric of the work as a whole. These questions demand a statement of what the work suggests about human nature or the human condition as it exists outside the text.
   Example: What comment or generalization does the chapter make on the way "civilization" influences human behavior and attitudes? Give evidence from the story to support your answer.

CHAPTER THREE

Evaluating Item Quality in Mathematics Assessments

by Vinci E. Daro, Ph.D. and Kari Kokka, Doctoral Candidate

Introduction

Current Context for Assessing Mathematical Proficiency

The adoption of the Common Core State Standards has drawn both public and expert attention to the role and purpose of large-scale, standardized assessments in mathematics. One of the driving principles of the development and state adoption of new standards has been to support students and teachers in exploring mathematical ideas with more rigor, depth, and coherence across the grade levels. Yet dominant culture in the era of standards-based accountability (still) includes enormous pressure on teachers to "cover" the "right" standards, meaning those that will be on the test. As a result, curricula and assessment measures continue to be atomized into bite-sized math, with the persistent unintended consequence of deeply fragmented mathematical experiences for students and teachers alike.

Within the currently active processes of development and adoption of standards-aligned assessments, using test items that align with rigorous learning and coherent instruction is an urgent priority.

Within the currently active processes of development and adoption of standards-aligned assessments, using test items that align with rigorous learning and coherent instruction is an urgent priority. In this environment, long-standing design challenges inherent in large-scale assessment development are intensified. In particular, the challenge of assessing conceptual mathematical understanding–together with procedural fluency–has persisted through decades of large-scale test development efforts, including those of NAEP, TIMSS, PISA, and state assessments (Yuan and Le, 2014). Recent efforts at developing higher quality assessments that might help steer instructional time toward more coherent and rigorous learning experiences have been productive, but this challenge remains.

The purpose of this review is to highlight item design features that offer promise for meaningfully measuring conceptual understanding. The review attends to common design dilemmas related to this challenge. For example, in order to elicit evidence of students' conceptual understanding, wording of items must be mathematically precise, but also student-friendly. Furthermore, what students are expected to do must be clearly and fully presented, but directions must also be concise enough that students can process them under pressure. Items must focus on mathematical ideas that are not only teachable and learnable, but worth teaching and learning.

Items must focus on mathematical ideas that are not only teachable and learnable, but worth teaching and learning.

In addition to managing these and other dilemmas, assessment developers are faced with the challenge of crafting items that are cognitively rigorous without relying on inconsiderate (hard to decode) stimuli and prompts.[1] When assessment items are difficult for construct-irrelevant

[1] Inconsiderately crafted stimuli and prompts are those with construct-irrelevant layering of information, grammatical and/or diagrammatical complexity, misleading cues representations that undermine student reasoning, and/or questions that are inconsistent with the items' context (contextual distortion).

reasons, inconsiderate presentation is very often the culprit.

Conducting this review involved identifying items with clear, concise stems or prompts that give students opportunities to reason about important mathematical ideas, without having to struggle to understand what is expected or what the focus of an item is. The assessment items we have included illustrate design decisions that support high levels of disciplinary rigor – for example, extended chains of reasoning, non-routine problem-solving, student decision-making, and coordinating across multiple mathematical representations – while simultaneously supporting students' access to core disciplinary ideas.

The sample items we have selected are not perfect (and some suggestions for improvement are included in the review), but they offer promising ways of capturing evidence of robust student understanding, that is, evidence of learning that is conceptually coherent and cognitively rigorous. The selected items are all publicly available from current (or recent) large-scale assessments, and therefore reflect approaches to standardized assessment that are very much within reach.

Methods

Item Selection

The item selection process was conducted in three rounds. To find examples of promising items among those publicly available from existing large-scale mathematics assessments, we convened a team of experts to review a wide range of released and practice items from all publicly available state and new consortium high-stakes assessments. This team—which consisted of three mathematical content experts, with support from a team of five interdisciplinary assessment experts—selected items that represent a variety of item types, grade levels, and content foci. Because the purpose of this review is to highlight items that illustrate design features and prompt structures that offer strong potential for meaningfully measuring student understanding in the context of large-scale standardized assessment, the selection included only item types that are typically included in large-scale, standardized testing. The math item selection process focused first on assessments currently or recently used in the United States for accountability purposes, such as student promotion or graduation, teacher evaluation, and/or school quality reports.

The selected items are all publicly available from current (or recent) large-scale assessments, and therefore reflect approaches to standardized assessment that are very much within reach.

Regardless of item type or content focus, the goal of the review process was to identify cognitively rigorous items that assess core disciplinary skills, such as extended reasoning, problem-solving, and coordinating across mathematical representations. The content of each item was considered, together with the response format, prompt wording and structure, and any specific design features that support item quality.

For the first round of item review, practice tests and released items from all states with publicly available items and from both test consortia (Smarter Balanced Assessment Consortium and Partnership for Assessment of Readiness for College and Careers) were combed for promising items, using an DOK framework as an initial filter ("Cognitive Complexity Framework," see below and Appendix A). The numerous assessments were divided amongst three mathematics content experts to read and select items that met an initial cut criterion of cognitive complexity Level 2 or higher for selected-response and constructed-response items, and Level 3 or higher for performance tasks. Each math context expert selected two to four of each item type. This first round yielded approximately twelve selected-response items, ten constructed-response items, and six performance tasks, each with an initial quantitative score and rationale for the score.

Math content experts reconvened for a second round of the selection process. Each item was discussed as a team and assigned a quantitative cognitive complexity level agreed upon by all three math experts. The candidate items were narrowed down to six selected-response items, six constructed-response items, and three performance tasks. All of these nominated items were then analyzed qualitatively (see below).

The third and final round of selection involved cognitive complexity and qualitative analyses of all nominated items and tasks. Two math content experts drafted the cognitive complexity and qualitative analyses for this set of items, which were then brought back to the team for consideration. The team's final selection yielded three selected-response (multiple-choice) items; three short constructed-response items (including two that are technology-enhanced); a performance task; and one hybrid task that blends selected- and constructed-response items to function as a (limited) performance task.

Item Analysis - Conceptual Frameworks

Cognitive Complexity Analysis

The primary tool used in the cognitive complexity analysis of assessment items was an adapted version of Webb's Depth of Knowledge framework (adapted by Herman, Buschang, & La Torre Matrundola, 2014). This DOK framework has been used widely to identify the cognitive demand of assessment items, and serves in the present review as a starting point for identifying item design features that support meaningful measurement of conceptual understanding and mathematical proficiency (NRC, 2001). To evaluate the cognitive complexity of the selected math assessment items, we applied the adapted Webb's Depth of Knowledge framework to arrive at a quantitative measure – Cognitive Complexity Level 1, 2, 3, or 4—for each item. The descriptors for each level are shown in Appendix A and also included (verbatim from the framework) in each item analysis. We also include a narrative description of the rationale for the assigned level [1-4] of each item.

Qualitative Analysis

In addition to the framework for quantitative evaluation, the review relied upon two tools for qualitative analysis of the selected items: (1) the PISA 2015 Draft Mathematics Framework[2], which describes three mathematical processes at the heart of mathematical literacy (*formulating*, *employing*, and *interpreting* mathematics), and seven mathematical capabilities that underlie these processes (*communication*, *mathematizing*, *representation*, *reasoning and argument*, *devising strategies*, *using symbols*, and *using mathematical tools*); and (2) the Cognitive Rigor Matrix/Depth of Knowledge[3] table developed by Hess, Carlock, Jones, and Walkup (2009), which provides descriptions of each Depth of Knowledge level, 1 through 4, for each of Webb's and Bloom's dimensions (*remember*, *understand*, *apply*, *analyze*, *evaluate*, and *create*). These two tools were used as references when considering the following categories addressed in the qualitative analysis of each item:

What the item assesses: The mathematics content of the item is described.

Approaches to the problem: If the item allows for multiple approaches to solving the problem, these approaches are discussed.

Design features that support item quality: Specific features of the item that distinguish it as a high quality or promising item are discussed. For example, the item may allow for multiple entry points or multiple solution strategies, or the item may include diagrams or technological enhancements that support student entry to the task and reasoning about important mathematical ideas.

Suggested improvements to the item: No item is perfect, and the items in this paper are presented as samples for discussion. Thus, suggestions are provided to improve many of the items reviewed. Suggestions may include ways to allow for multiple entry points or multiple solution strategies, or ways to make the item more clear, mathematically precise, or more relevant and authentic for the intended student audience.

Summary of Findings

The review process led our team to identify several patterns among the promising items across the grade levels and content foci. These patterns suggest a way forward from where we are, with respect to the design dilemmas described above.

In order to craft cognitively rigorous items that are both student-friendly and mathematically precise, are both clear and concise, and focus on mathematical ideas that are teachable, learnable, and important, we propose focusing on the following design considerations: **a focus on core disciplinary processes and ideas, support for multiple entry points, support for multiple solution strategies, considerate presentation, technological enhancements that support student reasoning, and engaging contexts.**

Our intention is to support those who want to prioritize assessments that can help steer in-

[2] http://www.oecd.org/pisa/pisaproducts/pisa2015draftframeworks.htm

[3] http://www.nciea.org/publications/rigorpresentation_KH11.pdf

structional time toward more coherent, more cognitively rigorous and conceptually robust learning experiences. Toward this end, we suggest the following questions for considering the design features that offer promise:

Core disciplinary processes and ideas: Does the item target student understanding of central and important mathematical processes and idea(s), and is the treatment of the idea(s) mathematically coherent?

Multiple entry points: Does the item offer visual representations, technology enhancements, or response formats that provide enough content for students to reason with, even if they have not memorized a certain procedure or computation?

Multiple solution strategies: Does the item support multiple approaches to a problem, and is the response format open enough to allow for more than one correct answer?

Considerate presentation: Is the stimulus/prompt presented in a way that minimizes construct-irrelevant difficulty by being concise, clear, mathematically and contextually coherent, and student-friendly?

Technology enhancements: Do the enhancements either (1) offer a productive scaffold for student reasoning without reducing the disciplinary rigor of the item, or (2) elevate the rigor of the item?

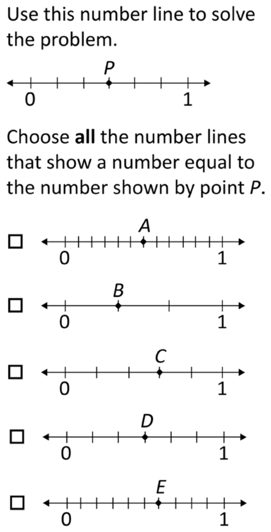
Engaging context: Is the item or task situated in a context likely to be meaningful and sensible to students from diverse socioeconomic, cultural, and language backgrounds?

Review of Selected Assessment Items

As noted earlier, the items we selected illustrate design features and prompt structures that offer promise for meaningful measurement of student understanding and skills in the context of large-scale standardized assessment. These promising items are presented below in three parts: Part I, Selected-Response Samples; Part II, Short Constructed-Response Samples; Part III, Performance Task Samples. Each sample item is presented together with cognitive complexity and qualitative analyses of the item.

Part I. Selected-Response Samples

Item Example 1



Use this number line to solve the problem.

Choose **all** the number lines that show a number equal to the number shown by point *P*.

Figure 1. Released selected-response item from the Smarter Balanced Assessment Consortium Grade 3 Mathematics Practice Test (Item#7). Reproduced with the permission of the Regents of the University of California.

Item Example 1 - Item Profile	
Source: Smarter Balanced Assessment Consortium, Grade 3 Mathematics Computer Adaptive Test Practice Test (Item #7), accessed on March 9, 2016 from http://sbac.portal.airast.org/practice-test/	
Grade level	3
Response type	Selected response
Core disciplinary processes and ideas	Interpreting number lines; interpreting values between 0 and 1; using number lines to identify equivalent values; recognizing equivalent fractions
Multiple entry points	Yes
Multiple solution strategies	Yes
Considerate presentation	Yes
Technology enhancements	No
Engaging context	No
Cognitive Complexity Level	2

Complexity Level (1-4)

Level 2

- Task requires some mental processing and more than rote application of skill, concept or procedural and/or algorithmic tasks.
- Students often make decisions about how to approach the problem

Rationale

This item has a cognitive complexity level of 2 because memorized procedures are not likely to support a student's success on the item and there are different approaches students might take to solving the problem. The task requires students to make sense of values between 0 and 1 using number lines on which only 0 and 1 are labeled, and partitions of different equal intervals are marked.

What the item assesses

This item assesses students' understanding of how a number line is used to represent numerical values, including the understanding that identical positions on the number line represent equal numerical values. The item requires students to use number lines to reason specifically about the values of fractions between 0 and 1. A student's ability to interpret number lines that are marked with different intervals will contribute to success on the item.

Approaches to the problem

There is room for slightly different approaches to this problem, depending on how students understand and use number lines. Some students may directly read the position of Point *P* visually and conclude that it is equidistant from 0 and 1, and then visually identify Points *A* and *D* as also each equidistant from 0 and 1. Students using this strategy are likely to also identify the value represented by *P* as one half, but need not do so to be successful. Other students might use the tick marks to identify *P* as 3⁄6 or 'three sixths' and then find points with equal or equivalent values, using the tick marks to determine the numerator and denominator in each case: Point *A* represents 6⁄12 or 'six twelfths,' Point *D* represents 3⁄6 or 'three sixths,' and all of the other points represent values not equivalent or equal to 3⁄6. Variations on these two approaches are also possible: (1) relying on a conceptual understanding of the structure of a number line, including the significance of equal distances, and (2) translating between number lines and numerical values, specifically values between 0 and 1.

Design features that support item quality

The primary design feature that makes this a quality assessment item is the minimal labeling of numerical values: although this item is ostensibly 'about fractions,' there are no fraction values provided or expected. This serves to focus students' attention very efficiently on the relevant concepts and the core disciplinary skills of using number lines to reason about fractions and using fractions to reason about number lines.

The item prompt is concise and sets clear expectations. The response choices are also well designed: with one correct choice identical to the initial number line and just one other correct choice, the item is likely to provide useful information about the understandings students bring to the problem.

An additional strength of this item lies in its consequential validity (as a practice test item). Insofar as practice test items guide instruction intended to prepare students for success on a summative test, this item supports instruction focused on conceptual understanding of the power and purpose of the number line. Students who have experience using number lines flexibly to represent, order, and operate on whole values will be better prepared for an item like this, as will students who have opportunities to become familiar with number lines marked with different intervals, both labeled and unlabeled. The item calls for learning experiences in which students connect their understanding of concepts that matter for fractions (order, the meaning of equal denominators, relative value, and the infinite divisibility of numbers) with concepts that matter for number lines (order, the meaning of equal intervals, relative position, and, even in third grade, the density of the number line).

...although this item is ostensibly 'about fractions,' there are no fraction values provided or expected. This serves to focus students' attention very efficiently on the relevant concepts and the core disciplinary skills...

Suggestions for improving the item

It could be stated that the intervals between 0 and 1 have equal length. The final part of the prompt *could* be made more mathematically correct:

> *Choose **all** the number lines that show a labeled point with value equal or equivalent to the number shown by Point P.*

However, for Grade 3, adding the word 'equivalent' may be not only unnecessary, but also difficult to decode. Pilot testing would be required to determine whether this wording is an improvement or detraction.

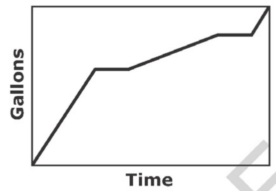
Item Example 2:

23 Mr. Jones filled his swimming pool with water.

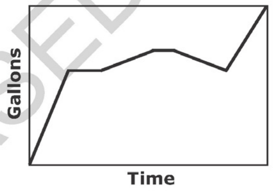
- Mr. Jones began filling the pool at a constant rate.
- He turned off the water for a while.
- He then turned the water back on at a slower constant rate.
- Mr. Jones turned off the water again for a while.
- He then turned the water back on at the first rate.

Which graph ***best*** represents Mr. Jones filling the pool?

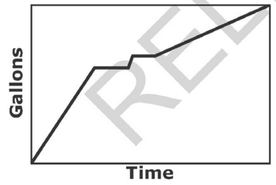
A



B



C



D

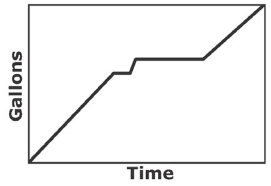


Figure 2. Released selected-response item from the North Carolina READY End-of-Grade Mathematics Assessment, Grade 8, p.18. Reproduced with permission of the North Carolina Department of Public Instruction.

Item Example 2 - Item Profile	
Source: North Carolina READY End-of-Grade Mathematics Assessment (Item #23), Grade 8, Revised 2015. © 2013 by the North Carolina Department of Public Instruction. Accessed on March 9, 2016 from http://www.ncpublicschools.org/docs/accountability/testing/releasedforms/g8mathpp.pdf	
Grade level	8
Response type	Selected response
Core disciplinary processes and ideas	Interpreting graphs; interpreting slope; using graphs to model relationships between variable quantities
Multiple entry points	Yes
Multiple solution strategies	Yes
Considerate presentation	Yes
Technology enhancements	No
Engaging context	Minimal
Cognitive Complexity Level	2.5

Cognitive Complexity Level (1-4)

Level 3

- Involves developing a solution strategy, and may have more than one possible answer.
- Task often requires significant departure from traditional application of concepts and skills.
- Solution strategy often involves working with multiple mathematical objects (numbers, expressions, equations, diagrams, graphs) or problem structures.

Rationale

This item has a cognitive complexity level that lies closest to 3. While there is only one possible correct answer, selecting the correct response requires developing a solution strategy for translating between a verbal description of a situation and a graph that models the situation. This is a significant departure from more procedural approaches to graphing insofar as there are no values to consider, no specific points to plot or read off of a graph, and no equations to parse.

What the item assesses

The item requires interpretation of a described situation, interpretation of given piecewise linear graphs, and translation between these two representations. The item requires a conceptual understanding of how a graph can be used to model a relationship between variable quantities, in this case, time and volume of water. It is not a technical exercise in identifying the slope value or identifying coordinates of any points; instead, it is designed to assess students' understanding of how different slope values appear on a graph, and how greater and lesser values of each quantity are represented graphically.

Approaches to the problem

There are multiple ways of approaching this problem. Students could read the described situation and begin to identify which aspects of the situation can be represented by a mathematical model, and how. Toward this end, they might ask themselves, What are the quantities? How are they related? or How would this relationship look on a graph? Students might recognize that the 'constant rate' mentioned in certain parts of the description will translate to linear graphs, or they might visualize the situation or draw a diagram to represent the alternately growing and enduring volume of water in the pool. Alternatively, students could look at the graphs and begin to identify significant features (e.g., the relative slopes of each piece of the function), and then match these features to the described situation. In all of these approaches, the reasoning involved requires conceptual linking between a situation and a mathematical model, without the scaffold—or the interference—of numerical values.

Design features that support item quality

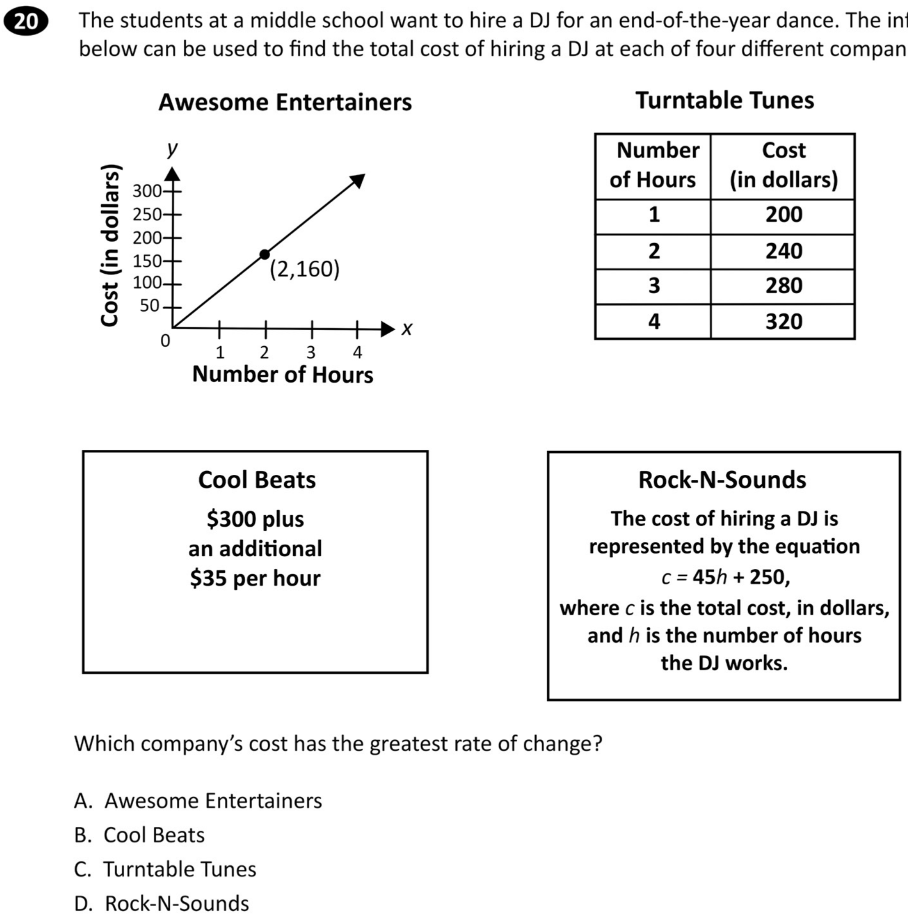
The description of the situation is straightforward and clear, and the bullet point structure sets up a useful correspondence with the shape of the graphs. Both of these features support the interpretive work students are expected to do. While the situation has no context or purpose (why would Mr. Jones behave in this way?), the intent of the problem is clear: students are to coordinate the verbal description of the two related quantities—amount of water in the pool, and time—with the graph showing this relationship. The fact that there are no values included in the description or on the graph supports a focus on the conceptual connection between the phrase, “a slower constant rate,” and a relatively less steep slope for part of the graph. This focus is not obscured by excess information to process or calculations to trip over, so as an assessment item, a correct response is directly within reach for any student who understands that this kind of connection-making ‘counts’ as important mathematical work and/or who has experience interpreting what the slope of a linear graph means in various particular contexts.

> **...a correct response is directly within reach for any student who understands that this kind of connection-making ‘counts’ as important mathematical work and/or who has experience interpreting what the slope of a linear graph means in various particular contexts.**

Suggestions for improving the item

The final question would be more accurately posed by asking, “Which graph best represents the amount of water in the pool as Mr. Jones fills it?”

Item Example 3



20 The students at a middle school want to hire a DJ for an end-of-the-year dance. The information below can be used to find the total cost of hiring a DJ at each of four different companies.

Awesome Entertainers

y
300
250
200
150
100
50
0
Cost (in dollars)
(2,160)
1
2
3
4
x
Number of Hours

Turntable Tunes

Number of Hours	Cost (in dollars)
1	200
2	240
3	280
4	320

Cool Beats

$300 plus an additional $35 per hour

Rock-N-Sounds

The cost of hiring a DJ is represented by the equation $c = 45h + 250$, where c is the total cost, in dollars, and h is the number of hours the DJ works.

Which company's cost has the greatest rate of change?

A. Awesome Entertainers
B. Cool Beats
C. Turntable Tunes
D. Rock-N-Sounds

Figure 3. Released selected-response item from the Massachusetts Comprehensive Assessment System, Grade 8 Mathematics, Spring 2015 (Item #20). Reproduced with the permission of the Massachusetts Department of Education.

Item Example 3 - Item Profile	
Source: Massachusetts Comprehensive Assessment System (MCAS), Grade 8 Mathematics, Spring 2015 Released Items, p. 235 (Item #20). Accessed on March 9, 2016 from http://www.doe.mass.edu/mcas/2015/release/default.html	
Grade level	8
Response type	Selected response
Core disciplinary processes and ideas	Using graphs, tables, and equations to model relationships between variable quantities; coordinating across different mathematical representations; interpreting and comparing constant rates when represented in verbal descriptions of situations, graphs, tables, and equations
Multiple entry points	Yes
Multiple solution strategies	Yes
Considerate presentation	Yes
Technology enhancements	No
Engaging context	Adequate
Cognitive Complexity Level	2

Cognitive Complexity Level (1-4)

Level 2

- Task requires some mental processing and more than rote application of skill, concept or procedural and/or algorithmic tasks.
- Students often make decisions about how to approach the problem.

Rationale

Students are asked to compare four different linear relationships between hours and cost, represented in four different ways: table, graph, verbal description, and equation. Students must decide how to read the constant rate off of each of these representations, and then apply the appropriate concept and procedure in each case in order to derive the value of each rate for comparison. The item requires students to understand that they are expected to compare the rate, and not the overall cost, for a particular number of hours (or another reasonable comparison that might be made across the DJ companies). Although there is rote application of skills involved, the interpretive work required to identify how the rate of change is represented in each case and the decisions needed about how to make a comparison warrant this item being designated a Level 2.

What the item assesses

This item challenges students to apply and connect their understanding of linear relationships across multiple representations. Students are likely to be successful on this item if they know that a constant rate of change is represented on a graph as the slope of a line, represented in a table as the constant amount of increase or decrease in one variable over equal intervals of the other variable, and represented in an equation that is written in slope-intercept form as the coefficient of the independent variable. Knowing how to interpret a description of a situation with a constant rate also supports success on this item.

Approaches to the problem

There are multiple approaches to this problem. In order to select the correct response, students must find the hourly rate of each DJ, and then compare these rates to identify which has the greatest value. For each of the given representations, there are multiple valid approaches to identifying and/or deriving a constant rate and multiple valid approaches for comparing rates. Possible approaches for each representation are presented below.

Graph: For the graph, students can calculate the slope between any two points, using any method they know, or they can identify the unit rate by finding the vertical increase for a horizontal increase of 1 unit. There are other approaches, but all require understanding that the slope, 80, represents the constant rate of change between time and cost. The fact that the DJ in this case does not have a fixed initial cost is an important feature of the design of this item. This design decision supports access to the mathematics of the item, as does the decision to label a single point with coordinates that are relatively straightforward to work with in order to derive the slope.

Because there are multiple methods that work to identify and derive the hourly rates from each representation, students who do not successfully recall a taught procedure can still reason their way to the correct rate if they understand what they are looking for (a constant rate of change).

Table: For the table, students can find the ratio between change in cost and change in hours, 40, and confirm that this ratio is constant over different intervals of the two variables. There are other approaches to interpreting the table, but all approaches require understanding that the constant rate is a constant ratio of differences between the two variables, time and cost. The values in the table are relatively easy to think with, which is important for making the item both accessible and more conceptually oriented than computationally challenging.

Equation: For the equation, students can read the constant rate of change, 45, directly off of the equation, or they can substitute different hour values into the equation and use the pattern to find the constant increase in cost; again, they need to understand what they are looking for (a constant rate of change).

Description: For the description, students can read the constant rate of $35 per hour directly if they understand that this is the constant rate of change to compare with other DJs; otherwise, they can test values, as in the equation, to find the pattern in the increase in cost with the increase in hours, and reason that the constant rate of change is reflected in this pattern.

Design features that support item quality

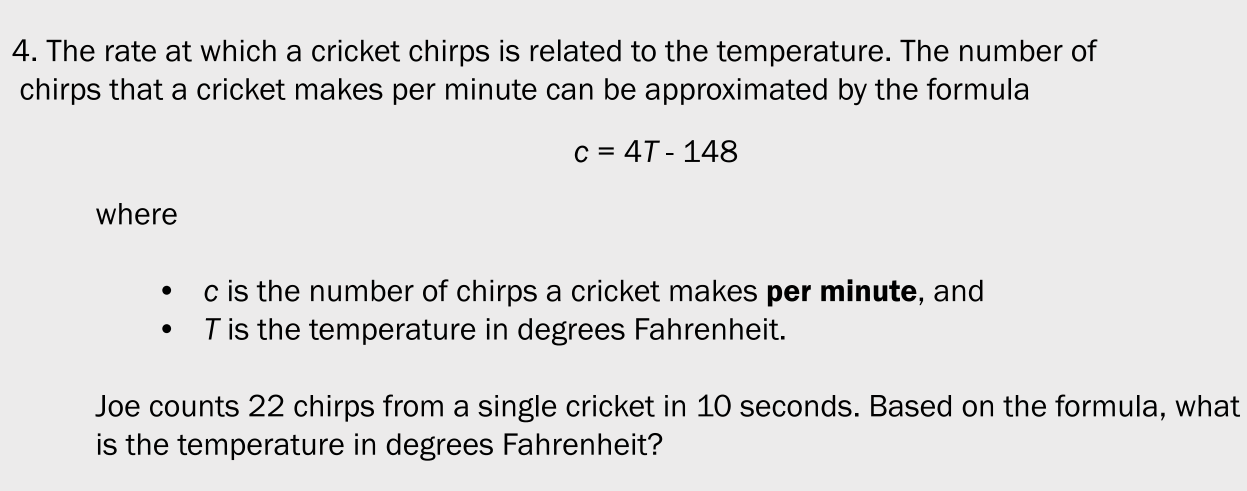
Students with a solid understanding of linear relationships will still be challenged procedurally to derive the correct rate from each representation and then make the comparisons correctly across representations. For students with a more fragile grasp of how to interpret linear relationships, both the context and the variety of representations support reasoning about each DJ's costs. Because there are multiple methods that work to identify and derive the hourly rates from each representation, students who do not successfully recall a taught procedure can still reason their way to the correct rate *if they understand what they are looking for* (a constant rate of change).

Because there are multiple methods that work to identify and derive the hourly rates from each representation, students who do not successfully recall a taught procedure can still reason their way to the correct rate if they understand what they are looking for (a constant rate of change).

Suggestions for improving the item

While mathematically precise, the final question would be more student-friendly if it were tied more coherently to the context. For example, "Which DJ company charges the greatest hourly rate?"

Item Example 4



4. The rate at which a cricket chirps is related to the temperature. The number of chirps that a cricket makes per minute can be approximated by the formula

$$c = 4T - 148$$

where

- c is the number of chirps a cricket makes **per minute**, and
- T is the temperature in degrees Fahrenheit.

Joe counts 22 chirps from a single cricket in 10 seconds. Based on the formula, what is the temperature in degrees Fahrenheit?

Figure 4. Released constructed-response item from the 2013 Connecticut Academic Performance Test, Grade 10 Mathematics (Item #4). Reproduced with the permission of the Connecticut State Department of Education.

Item Example 4 - Item Profile	
Source: Connecticut Academic Performance Test (Grade 10), 2013 Mathematics Released Items and Scored Student Responses, p.196 (Item #4). Accessed on March 14, 2016 from http://www.csde.state.ct.us/public/csde/cedar/assessment/capt/released_items.htm#7	
Grade level	10
Response type	Constructed response
Core disciplinary processes and ideas	Interpreting linear functions that model relationships between variable quantities; using linear equations to find solutions
Multiple entry points	Yes
Multiple solution strategies	Yes
Considerate presentation	Yes
Technology enhancements	None
Engaging context	Minimal
Cognitive Complexity Level	2.5

Cognitive Complexity Level (1-4)

Level 2

- Task requires some mental processing and more than rote application of skill, concept or procedural and/or algorithmic tasks.
- Students often make decisions about how to approach the problem.

Rationale

This item has a cognitive complexity level of 2. The item does not reflect a significant departure from traditional application of concepts, nor does the item have more than one possible answer, but the item challenges students to make decisions about how to approach a problem that involves working with a verbal description, an equation, and a pair of values that cannot themselves be directly substituted into the equation. A student's computations and algebraic manipulation will only be successful if the student can develop a solution strategy that is based on an accurate interpretation of the quantities, their relationship, and the given values.

What the item assesses

The sense-making this item requires is about the quantities, including their units, and the algebraic representation of the relationship between the quantities. A solid grounding in algebraic manipulation will support students' success on this item, as will experience with modeling linear, or approximately linear, relationships between real-world quantities.

Approaches to the problem

Students can use the ratio of ‘22 chirps to 10 seconds’ to identify a unit rate of 2.2 chirps per second, and then multiply 2.2 by 60 seconds to find a *c*-value of 132 chirps per minute. This value can be substituted into the equation to solve for the corresponding *T*-value. There are several paths from this point forward that would work to find *T*, including, for example, a purely algebraic process, mental calculation, and guess-and-check.

Students could also multiply ‘22 chirps per 10 seconds’ by 6 to get ‘132 chirps per 60 seconds’, and then follow one of the paths described above.

Design features that support item quality

There are several important mathematical challenges in this item: students must first understand the situation and then understand how the answer they are asked to provide relates to the equation they are given. The values provided in the situation students must interpret have units that are different from the variables represented in the equation, which means the provided values must not be directly substituted into the equation to find an answer. These challenges generate cognitive complexity that is appropriate not only for Grade 10, but also for the kinds of real-world situations that are typically modeled by linear functions. While cognitively complex, the prompt is considerately worded: the language is concise and mathematically coherent, and what students are expected to do is clear.

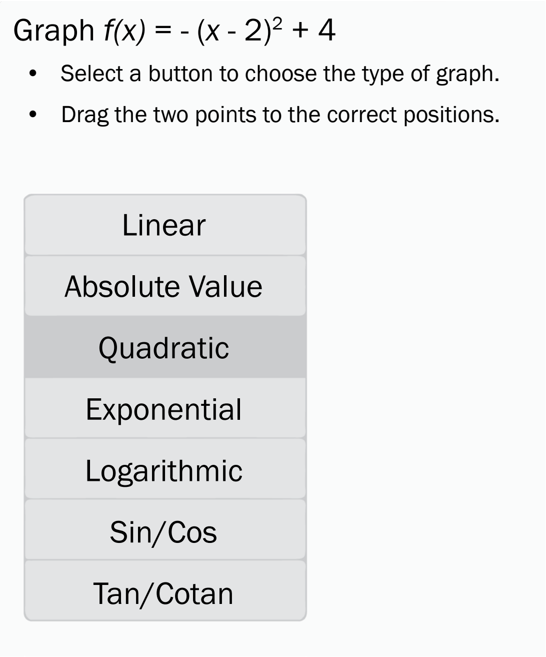
While cognitively complex, the prompt is considerately worded: the language is concise and mathematically coherent, and what students are expected to do is clear.

Suggestions for improving the item

The item would be improved by a motivating purpose for computing the current temperature.

Item Example 5

(technology-enhanced*)



Graph $f(x) = -(x - 2)^2 + 4$

- Select a button to choose the type of graph.
- Drag the two points to the correct positions.

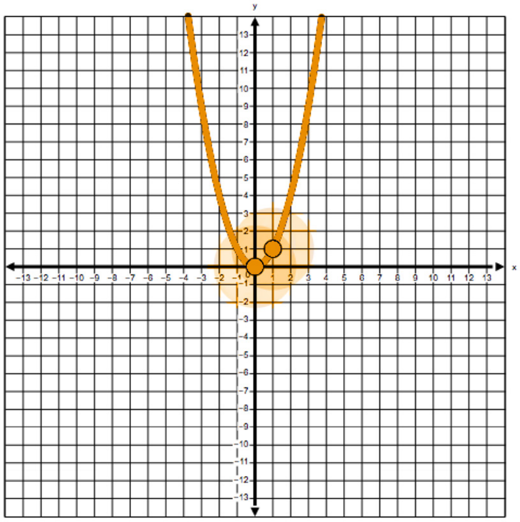


Figure 5. Released item from Partnership for Assessment of Readiness for College and Careers (PARCC) End-of-Year Algebra 1 Test (Computer-Based Practice Test, Part II, Item #6). Reproduced with fair-use permission from PARCC.

***About the technological enhancement:** The coordinate plane provided for this item has no graph until students select a function type. Once a type is selected, the parent function of that type appears on the plane with two marked points, at $x = 0$ and $x = 1$. Students can then transform the graph by dragging either or both of these points. The entire (shown) graph transforms automatically when either point is dragged to a new position.

Item Example 5 - Item Profile	
Source: Partnership for Assessment of Readiness for College and Careers (PARCC) End-of-Year Algebra 1 Test (Computer-Based Practice Test, Part II, Item #6). Accessed on March 14, 2016 from http://parcc.pearson.com/practice-tests/math	
Grade level/course	Algebra
Response type	Constructed response, technology enhanced
Core disciplinary processes and ideas	Interpreting the structure of algebraic expressions; recognizing a quadratic function represented algebraically; making connections between mathematical representations; awareness of function types; graphing quadratic functions; transformations in the coordinate plane.
Multiple entry points	Yes
Multiple solution strategies	Yes
Considerate presentation	Yes
Technology enhancements	Yes
Engaging context	No
Cognitive Complexity Level	2-3

Cognitive Complexity Level (1-4)

Level 2

- Task requires some mental processing and more than rote application of skill, concept or procedural and/or algorithmic tasks.
- Students often make decisions about how to approach the problem.

Level 3

- Involves developing a solution strategy, and may have more than one possible answer
- Task often requires significant departure from traditional application of concepts and skills.
- Strategy often involves working with multiple mathematical objects (numbers, expressions, equations, diagrams, graphs) or problem structures.

Rationale

This item has a cognitive complexity level that lies between 2 and 3: all of the criteria for Level 2 are met, and although there is only one possible answer, students must work with both algebraic and graphical representations of a function to develop a solution strategy (Level 3). Moreover, the technology enhancement supports a departure from traditional application of concepts in that the graph facilitates an investigative approach to the connection between input/output values of a function and the points on its graph.

What the item assesses

This item assesses students' basic familiarity with function "families," and the specific ability to recognize and graph a quadratic function. The first part of the item depends on a student's understanding that there are families, or types, of functions, and how these are named. This part also depends on students having a strategy to identify the function as quadratic. The second part of the item challenges students to represent the function graphically, using a tool that facilitates the process.

Approaches to the problem

There are several strategies that will work to solve each part of the problem. To successfully complete the first part of the problem, students might directly recognize the structure of the algebraic expression as quadratic, or might decide to manipulate the expression into a more familiar form (e.g., $ax^2 + bx + c$) to identify or confirm the function's type. Students could also use the graphing tool to investigate possible function types before making a selection.

For the second part of the problem, once students are working with a given parabola in the coordinate plane, some students may draw on their knowledge of quadratic functions to directly identify key features of the graph (vertex and intercepts). Students with less experience with – or less memorized knowledge about – quadratic functions might substitute input values into the function to determine a few output values, and then transform the graph accordingly. Still other students might use a guess-and-check approach to transforming the graph, trying out different positions, orientations, and dilations of the parent graph and then testing coordinate values by substituting them into the equation.

For both parts of the problem, students might also create a table of values to (1) identify the change in output values as quadratic, and (2) identify sufficient coordinate pairs to establish the graph.

Design features that support item quality

Two features make this a high quality item. First, the technological enhancement makes it relatively easy to manipulate each graph, so the graphing tool supports the process of connecting corresponding pairs of input and output values with corresponding points on the graph. The ability to freely transform the graph using either or both of the points facilitates an investigative approach to the problem. This provides important opportunities for success to students who have forgotten (or never memorized) how to directly read the vertex off of the given form of the quadratic expression but have a solid understanding of the concept of a function—that a function determines one output value for each input value, that each corresponding pair of values can be represented as a point in the coordinate plane, and

...the technological enhancement makes it relatively easy to manipulate each graph, so the graphing tool supports the process of connecting corresponding pairs of input and output values with corresponding points on the graph. The ability to freely transform the graph using either or both of the points facilitates an investigative approach to the problem.

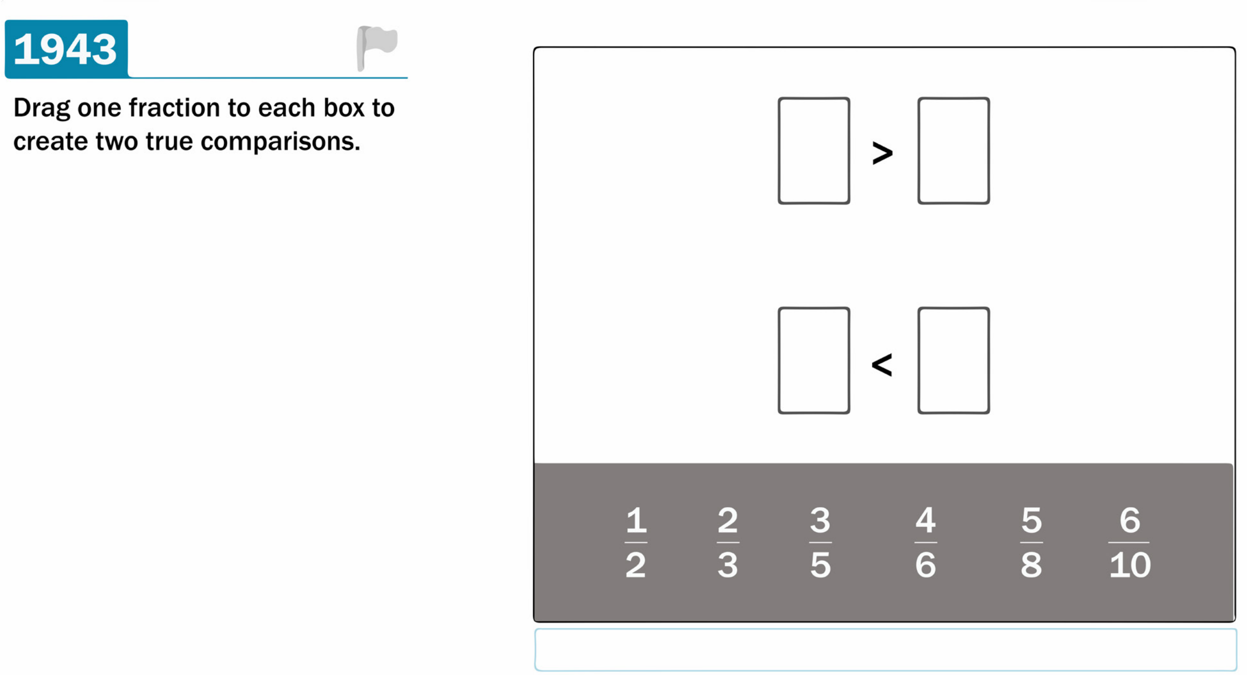
that this full set of points is the graph of the function. Second, student success does not depend on memorization (though the item's presentation of the function in vertex form makes memorized knowledge about this form directly relevant). Instead, conceptual understanding, together with an investigative approach, will yield a correct response, and both are supported by the item's design.

Suggestions for improving the item

The second bullet of the item prompt, "Drag the two points to the correct positions," may not be clear enough to some students. While brevity is usually a positive feature of short assessment items in math, a slight addition to this bullet's instructions may provide more clarity about what is expected. For example, the instructions could be revised as follows: "Create the graph of the function by dragging the two marked points to the correct positions."

Item Example 6

(technology-enhanced)



1943

Drag one fraction to each box to create two true comparisons.

□ > □

□ < □

$\frac{1}{2}$ $\frac{2}{3}$ $\frac{3}{5}$ $\frac{4}{6}$ $\frac{5}{8}$ $\frac{6}{10}$

Figure 6. Released item from the Smarter Balanced Assessment Consortium (SBAC), Grade 4 Math Computer Adaptive Test Practice Test (Item #1943). Reproduced with the permission of the Regents of the University of California.

<table>
<tr><th colspan="2">Item Example 6 - Item Profile</th></tr>
<tr><td colspan="2"><b>Source:</b> Smarter Balanced Assessment Consortium (SBAC) Grade 4 Computer Adaptive Test Practice Test (Item #26). Accessed on March 14, 2016 from: http://sbac.portal.airast.org/practice-test/</td></tr>
<tr><td><b>Grade level/course</b></td><td>4</td></tr>
<tr><td><b>Response type</b></td><td>Constructed response, technology enhanced</td></tr>
<tr><td><b>Core disciplinary processes and ideas</b></td><td>Identifying equivalent fractions; ordering fractions</td></tr>
<tr><td><b>Multiple entry points</b></td><td>Yes</td></tr>
<tr><td><b>Multiple solution strategies</b></td><td>Yes</td></tr>
<tr><td><b>Considerate presentation</b></td><td>Yes</td></tr>
<tr><td><b>Technology enhancements</b></td><td>Yes</td></tr>
<tr><td><b>Engaging context</b></td><td>No</td></tr>
<tr><td><b>Cognitive Complexity Level</b></td><td>2-3</td></tr>
</table>

Cognitive Complexity Level (1-4)

Level 2

- Task requires some mental processing and more than rote application of skill, concept or procedural and/or algorithmic tasks.
- Students often make decisions about how to approach the problem.

Level 3

- Involves developing a solution strategy, and may have more than one possible answer.
- Task often requires significant departure from traditional application of concepts and skills.
- Solution strategy often involves working with multiple mathematical objects (numbers, expressions, equations, diagrams, graphs) or problem structures.

Rationale

The cognitive complexity level of this item lies between 2 and 3. The item meets the requirements for Level 2, and has more than one possible answer, qualifying it partially for Level 3. Students must decide which fractions to use in each pair, and can use any strategy they are comfortable with for ordering or comparing fractions. The item design represents a departure from traditional fraction assessment items, enabled in part by the “drag and drop” technology.

What the item assesses

This item assesses students' ability to compare fractions, as well as students' fluency with notation for inequality statements. Also, a student's understanding of equivalent fractions would directly support their success on the item.

Approaches to the problem

This item allows for a variety of approaches. Students might directly recognize two pairs of unequal fractions, know which is greater in each case, and know how to set up an inequality statement using notation correctly. Others might use the drag and drop technology to visually stabilize a possible inequality statement for consideration, and then compute or reason through a determination of which of the two fractions has greater value (or determine that they are equivalent). Students may also draw on their familiarity with certain "benchmark" fractions (e.g., ½ and ⅔), and then compute or reason through a determination of other fractions greater or less than these familiar fractions. Students could also draw on their understanding or recognition of equivalent fractions in the given set, and then use those to anchor the two inequalities they must produce, building from the same value each time.

Students who are more comfortable with a procedural approach to comparing fractions might find a common denominator of all or some of the given fractions, and then compare numerators (though this would be less efficient than relying on familiarity with benchmark fractions or on conceptual understanding of the meaning of smaller and greater denominators). Alternatively, or additionally, students might use a number line strategy to order all or some of the given fractions before completing the statements. Other approaches are also possible (e.g., using tape diagrams or area models to compare values.)

Design features that support item quality

The technological enhancement of this item is minimal — and some readers might dispute that this is a constructed-response item, since students are selecting from a given set of options. However the drag-and-drop technology is more than a convenient way of selecting answer choices. It directly supports, and even elevates, the cognitive rigor of the problem. The simple technology, combined with the design decision to provide more fractions than are needed to solve the problem, enable a much wider range of correct responses than traditional selected-response fraction items; this supports student choice and increases rigor. At the same time, the item's design allows students who have developed proficiency with a limited number of fractions to work with those familiar fractions, giving more students access to core disciplinary ideas. The technology does nothing to 'give away' answers, but students can use the drag-and-drop format to visually arrange and stabilize possible inequalities in order to systematically consider whether they are true or not.

The simple technology, combined with the design decision to provide more fractions than are needed to solve the problem, enable a much wider range of correct responses than traditional selected-response fraction items; this supports student choice and increases rigor.

The fraction choices also reflect important design decisions: two choices are considered 'benchmark' fractions (students typically have more experience with these than others), and there are two pairs of equivalent fractions, which provide the opportunity to bring experience with equivalence to bear on the problem.

Additionally, the fact that students must produce two (as opposed to only one) true comparisons is likely to yield useful information about the understandings students have about the relative values of fractions, without exhausting students for whom this problem is computationally or otherwise challenging.

Suggestions for improving the item

The item prompt could be slightly revised for clarity as follows: "Fill in each box with a fraction from the list to make the comparisons true. Drag the fractions into the boxes."

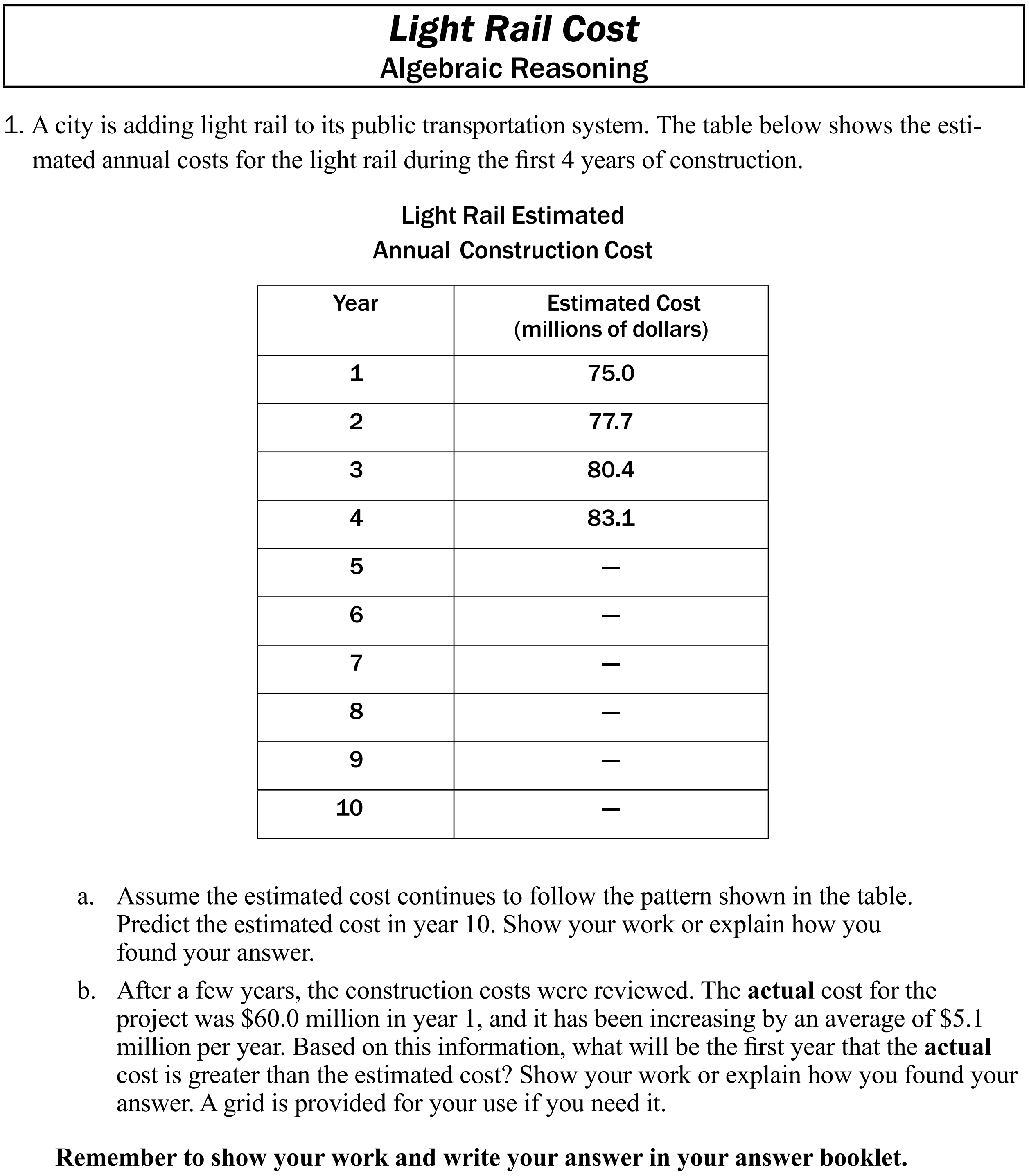
Part III. Sample Performance Tasks

Engagement is important because of its relationship to student performance on performance assessment tasks, especially for students who have been typically less advantaged in school settings such as English Language Learners, students of historically marginalized backgrounds, etc.

Performance tasks should invite students to engage in cognitively demanding work through analyzing and synthesizing various sources and representations of information. Performance tasks usually ask students to represent scenarios mathematically and/or understand how given mathematical representations relate back to a scenario. Performance tasks often focus on mathematics content that is below the grade level of the overall assessment because they are designed not to assess grade-level concepts and skills, but rather students' abilities to analyze, synthesize, communicate, and represent mathematical ideas relevant to a cognitively demanding problem or scenario. In our review of items, we looked for performance tasks that are engaging because they reflect authentic real-world situations and/or because they encourage student agency by giving students a role with a clear purpose and choices to make. Engagement is important because of its relationship to student performance on performance assessment tasks, especially for students who have been typically less advantaged in school settings such as English Language Learners, students of historically marginalized backgrounds, etc. (Arbuthnot, 2011; Darling-Hammond et al., 2008; SCOPE/SCALE, 2015; Walkington, 2013).

The two performance tasks shown below are presented as samples for discussion. The first, a sample Connecticut Academic Performance Task, is included because it demonstrates how a performance task may invite students to engage in cognitively complex mathematics. While item is not perfect, we believe discussion of this sample item may help educators and assessment directors understand promising and important features of performance tasks for inclusion in standardized summative assessments.

Item Example 7



Light Rail Cost

Algebraic Reasoning

1. A city is adding light rail to its public transportation system. The table below shows the estimated annual costs for the light rail during the first 4 years of construction.

Light Rail Estimated
Annual Construction Cost

Year	Estimated Cost (millions of dollars)
1	75.0
2	77.7
3	80.4
4	83.1
5	—
6	—
7	—
8	—
9	—
10	—

a. Assume the estimated cost continues to follow the pattern shown in the table. Predict the estimated cost in year 10. Show your work or explain how you found your answer.

b. After a few years, the construction costs were reviewed. The **actual** cost for the project was $60.0 million in year 1, and it has been increasing by an average of $5.1 million per year. Based on this information, what will be the first year that the **actual** cost is greater than the estimated cost? Show your work or explain how you found your answer. A grid is provided for your use if you need it.

Remember to show your work and write your answer in your answer booklet.

Figure 7. Released Grade 10 performance task from the 2013 Connecticut Academic Performance Test Released Items (Item #1). Reproduced with permission from Connecticut State Department of Education.

Item Example 7 - Item Profile	
Source: Connecticut Academic Performance Test (Grade 10), 2013 Mathematics Released Items and Scored Student Responses, p.171 (Item #1). Accessed on March 14, 2016 from http://www.csde.state.ct.us/public/csde/cedar/assessment/capt/released_items.htm#7	
Grade level/course	High school
Response type	Constructed response
Core disciplinary processes and ideas	Recognizing aspects of a real world situation that correspond with the structure of linear relationships between variable quantities; switching between mathematical representations to solve problems; using systems of linear equations to find solutions
Multiple entry points	Yes
Multiple solution strategies	Yes
Considerate presentation	Yes
Technology enhancements	No
Engaging context	Adequate
Cognitive Complexity Level	4

Cognitive Complexity Level (1-4)

Level 4

- Task requires extended reflection, including complex problem solving, abstract reasoning, an investigation, processing of multiple conditions of the problem, and non-routine manipulations.
- Task often requires extended time.

Rationale

This performance task has a cognitive complexity level of 4 because students must engage in complex problem solving and abstract reasoning while considering two conditions of the problem to arrive at a solution. Students can develop a solution strategy using a table, graph, or equation (or even possibly guess-and-check). These solution strategies involve working with and among multiple mathematical representations (e.g., table, graph, equation, and scenario), requiring extended time to navigate between multiple representations and strategies.

What the item assesses

This item requires students to create a mathematical model using their knowledge of linear functions. Students are asked to *decontextualize* and *contextualize,* navigating between the scenario and the mathematical representation of the scenario. They must understand multiple representations of linear functions (table, data points, graph, and scenario stated in words) in order to construct their model. Students are presented with a data table indicating the *estimated* annual construction costs of the light rail. They are then given the *actual* annual construction costs of the light rail, in words and with data points, and are asked to reevaluate and compare their initial model to the actual costs. After being presented with the *actual* annual construction costs data, students are asked, "Based on this information, what will be the first year that the ***actual*** cost is greater than the estimated cost?" which requires them to solve a system of linear functions.

Approaches to the problem

In order to create the mathematical models necessary to solve the problem, students must use their understanding of functions, and they must realize that both functions are linear functions, represented in two different ways. The first function is represented as a table, and the second function is represented in words and with data points.

There are multiple approaches to this problem. Students may create two tables representing each function to find the point of intersection, or they may create equations and solve the system algebraically. Students might also graph both functions and find the point of intersection; although a graphical approach is not as accurate, it is a useful starting point for some students who may then check their answer using an additional solution strategy (e.g., algebraic approach). Finally, students may also use a guess-and-check strategy to find the first year that the actual cost is greater than the estimated cost.

Design features that support item quality

This task is a promising assessment item because it allows students to use multiple representations of linear functions: table, graph, equation, and words/scenario. This flexibility invites students to demonstrate their conceptual understanding of mathematical modeling and linear functions using one of multiple approaches to the problem. Not only does the task invite flexibility, but it draws on students' higher order thinking skills to understand and navigate between representations.

Not only does the task invite flexibility, but it draws on students' higher order thinking skills to understand and navigate between representations.

Importantly, although the rubric for this task is not shown, it explicitly states that scorers should give credit to these various solution approaches: graphical, algebraic, and guess-and-check responses. The rubric also specifies that a range of responses should be granted full credit (i.e., "7.00-8.00 years with sufficient supporting work").

Analysis of scoring rubrics is outside the scope of this paper, but scoring rubrics are of special interest when searching for or developing quality performance tasks. Not only should the rubric specify the variety of methods that can be used to arrive at an answer, but the expectations defined in a rubric must align very closely with the task prompt – meaning that the rubric should evaluate what the task clearly and explicitly asks students to complete. If, for example, the task itself does not clearly ask for specific steps or a particular method, the scoring rubric should not be written to anticipate these, and students should not be penalized for failing to provide what they have not been explicitly asked to include.

Suggestions for improving the item

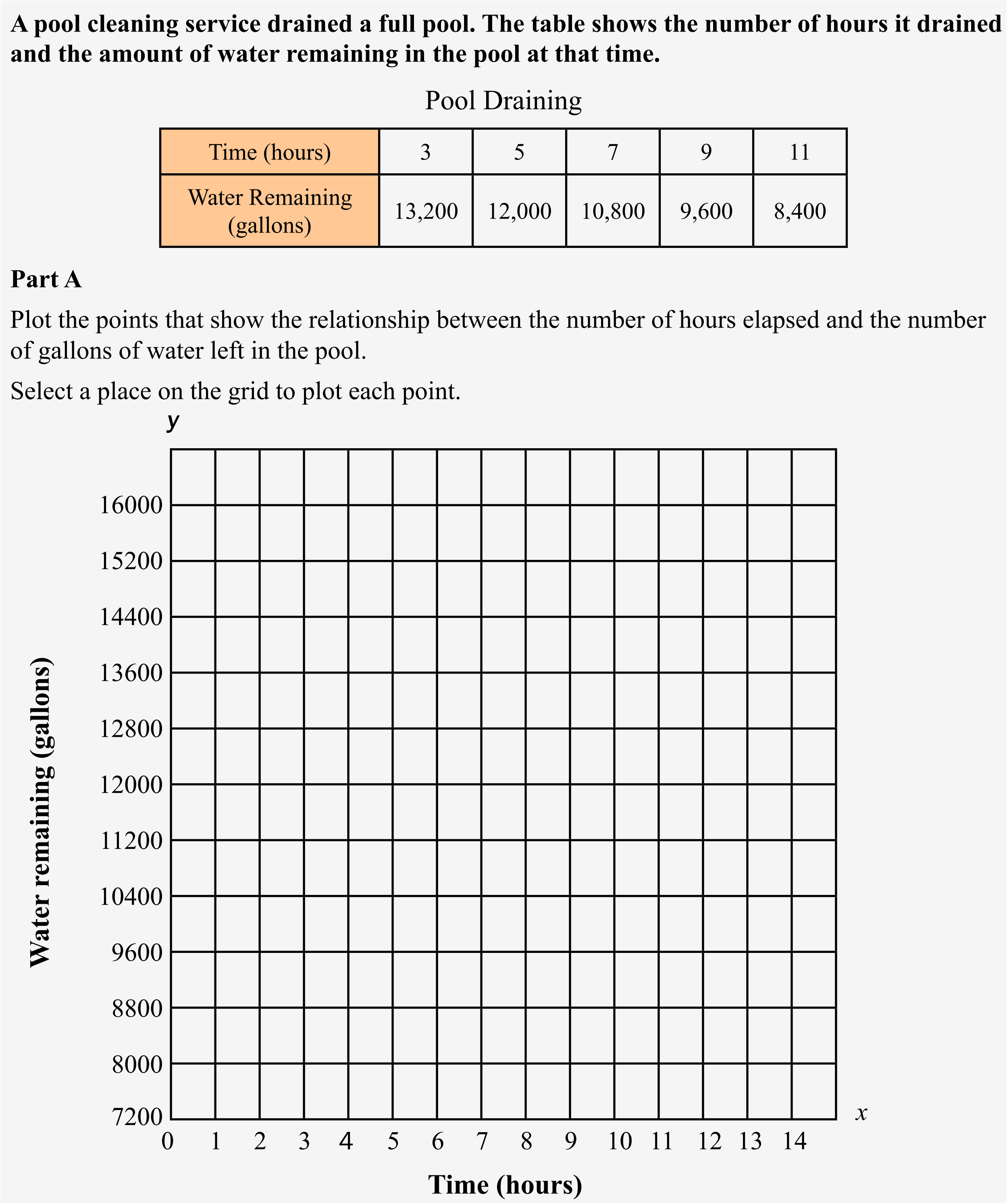
This task may be improved by adding a visual image of a light rail vehicle (e.g., streetcar, subway car) for students to imagine the scenario, especially for students who are English Language Learners. An image of a light rail transit system under construction would be even more illustrative. In case the intended student audiences are unlikely to have experiences with a light rail transit system, contextualized items like this should, whenever feasible, be reframed around a more familiar transportation system (e.g., a new bus line), in order to make the task relevant and authentic to the student audience.

The rubric could be improved to also include students' use of a table as an acceptable and credit-worthy solution strategy.

Item Example 8

(Hybrid)

The following sample PARCC task offers one possible way to develop a performance task that does not require hand scoring. This is an example of a hybrid task that blends selected- and constructed-response items to function as a (limited) performance task. Hand scoring may not always be a viable option (due to costs, etc.), and this task presents one way for students to engage in a performance task that is essentially a connected set of constructed-response and selected-response (multiple-choice) items. This task is a promising sample because it is NOT a set of disjointed multiple-choice or constructed-response items. An ill-designed performance task may be a set of items that have a common theme (e.g., a sporting event or a science experiment) but are not mathematically connected. In this task, the selected-response and constructed-response items are both thematically and mathematically connected.



A pool cleaning service drained a full pool. The table shows the number of hours it drained and the amount of water remaining in the pool at that time.

Pool Draining

Time (hours)	3	5	7	9	11
Water Remaining (gallons)	13,200	12,000	10,800	9,600	8,400

Part A

Plot the points that show the relationship between the number of hours elapsed and the number of gallons of water left in the pool.

Select a place on the grid to plot each point.

Figure 8a. Released item from Partnership for Assessment of Readiness for College and Careers (PARCC) Grade 8 End-of-Year Mathematics Computer-Based Practice Test, Part II Calculator Part (Item #7). Reproduced with fair-use permission from PARCC.

Part B

The data suggests a linear relationship between the number of hours the pool had been draining and the number of gallons of water remaining in the pool. Assuming the relationship is linear, what does the rate of change represent in the context of this relationship?

○ A. the number of gallons of water in the pool after 1 hour
○ B. the number of hours it took to drain 1 gallon of water
○ C. the number of gallons drained each hour
○ D. the number of gallons of water in the pool when it is full

Part C

What does the *y*-intercept of the linear function represent in the context of this relationship?

○ A. the number of gallons of water in the pool after 1 hour
○ B. the number of hours it took to drain 1 gallon of water
○ C. the number of gallons drained each hour
○ D. the number of gallons of water in the pool when it is full

Part D

Which equation describes the relationship between the time elapsed and the number of gallons of water remaining in the pool?

○A. $y = -600x + 15{,}000$
○B. $y = -600x + 13{,}200$
○C. $y = -1{,}200x + 13{,}200$
○D. $y = -1{,}200x + 15{,}000$

Figure 8b. Released item from Partnership for Assessment of Readiness for College and Careers (PARCC) Grade 8 End-of-Year Mathematics Computer-Based Practice Test. Reproduced with fair-use permission from PARCC.

<table>
<tr><th colspan="2">Item Example 8 - Item Profile</th></tr>
<tr><td colspan="2"><b>Source:</b> Partnership for Assessment of Readiness for College and Careers (PARCC) Grade 8 Mathematics Computer-Based Practice Test, Part II Calculator Part (Item #7). Accessed on March 14, 2016 from: http://parcc.pearson.com/practice-tests/math</td></tr>
<tr><td><b>Grade level</b></td><td>8</td></tr>
<tr><td><b>Response type</b></td><td>Performance task, computer enhanced</td></tr>
<tr><td><b>Core disciplinary processes and ideas</b></td><td>Recognizing patterns and their correspondence to the structure of linear relationships; using graphs and equations to represent linear relationships between variable quantities; interpreting mathematical representations</td></tr>
<tr><td><b>Multiple entry points</b></td><td>Yes</td></tr>
<tr><td><b>Multiple solution strategies</b></td><td>Yes</td></tr>
<tr><td><b>Considerate presentation</b></td><td>Yes</td></tr>
<tr><td><b>Technology enhancements</b></td><td>Yes</td></tr>
<tr><td><b>Engaging context</b></td><td>Minimal</td></tr>
<tr><td><b>Cognitive Complexity Level</b></td><td>3</td></tr>
</table>

Cognitive Complexity Level (1-4)

Level 3

- Involves developing a solution strategy, and may have more than one possible answer.
- Task often requires significant departure from traditional application of concepts and skills.
- Solution strategy often involves working with multiple mathematical objects (numbers, expressions, equations, diagrams, graphs) or problem structures.

Rationale

This task has a cognitive complexity level of 3 because students must work with multiple mathematical objects (numbers, expressions, equations, diagrams, graphs) and problem structures. Students are asked to create a mathematical model using a graph and an equation. The task invites students to contextualize and decontextualize, navigating between the scenario and the mathematical representations of the scenario; in other words, the task asks student to go beyond the traditional, rote application of concepts and skills.

What the item assesses

This task assesses students' understanding of multiple representations of linear functions to create a viable solution strategy. Students are asked to create a specific representation of a model and demonstrate understanding of how the parts of the mathematical model relate to the scenario (e.g., What does the *y*-intercept of the linear function represent in the context of this relationship?) Part A asks students to interpret the scenario (presented in words and with a table) and to display the data on the graph, using computer-enhanced technology where students can plot points on the graph through the online testing platform. Parts B and C ask students to understand the slope and *y*-intercept in the context of the pool draining scenario, assessing students' ability to navigate between understanding the scenario and the mathematical representations of the scenario. Part D asks students to find the equation that models change in gallons of water remaining over time in hours.

Approaches to the problem

This item offers a limited number of approaches to solving the problem. In Part A, students must plot points from the given table to represent the function graphically. For Parts B and C, students may arrive at an answer by using either the table representation or the graphical representation of the function to identify what the slope and *y*-intercept represent about the scenario. In Part D, which asks students to select the equation that represents the relationships between the time elapsed and the number of gallons remaining in the pool, students may use a table or graphical representation to help them understand the algebraic representation that correctly models the given function.

Design features that support item quality

This is a promising performance task because, although it is composed of a constructed-response item (the graph) and several selected-response items, the mathematics "hangs together" across the items of the task. In part A, when students create a graph using the online testing platform, they have the opportunity to reason with the data displayed both as a graph and in a table, encouraging flexibility of understanding and connecting between representations. This scaffold helps students to get a start on the problem by considering multiple representations of the scenario. The topic of this task (draining a pool) is also authentic, although it may not be comprehensible by all students, particularly those who live in colder climates or who do not have access to a pool.

This is a promising performance task because, although it is composed of a constructed-response item (the graph) and several selected-response items, the mathematics "hangs together" across the items of the task.

Suggestions for improving the item

This item can be improved by opening it up to allow for responses beyond the formats included, for example plotting points on a graph, which may or may not necessitate hand scoring. The item could also be strengthened by the addition of a visual image of a pool for students to better picture the scenario, especially for students who are English Language Learners. Like the previously discussed sample performance task from the Connecticut Academic Performance Test, this task could be improved by ensuring that the context is appropriate and relevant for its audience. For assessments in states where students may have less exposure to swimming pools, a more suitable scenario may be draining a bath tub.

References

Arbuthnot, K. (2011). *Filling in the blanks: Understanding standardized testing and the black-white achievement gap*. Charlotte, NC: Information Age Publishing.

Darling-Hammond, L., Barron, B., Pearson, D. P., et al. (2008). *Powerful learning: What we know about teaching for understanding*, pp. 74-76. San Francisco, CA: Jossey-Bass.

Herman, J., Buschang, R., La Torre Matrundola, D., & Wang, J. (2014). *An explanation of the ELA and Math Cognitive Complexities Frameworks.* Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing.

Herman, J. L., La Torre Matrundola, D., & Wang, J. (2015). *On the road to assessing deeper learning: What direction do test blueprints provide?* (CRESST Report 849). Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Herman, J.L. & Linn, R.L. (2013). *On the road to assessing deeper learning: The status of Smarter Balanced and PARCC assessment consortia.* (CRESST Report 823). Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

SCOPE/SCALE (2015). *Engagement toolkit for assessment item writers: Dimensions of engagement definitions and ways to incorporate; Task development guidelines with dimensions of engagement for item writers; Review tool with engagement considerations for item writers/evaluators.* Stanford, CA: Stanford Center for Opportunity Policy in Education/Stanford Center for Assessment, Learning & Equity.

Walkington, C. A. (2013). Using adaptive learning technologies to personalize instruction to student interests: The impact of relevant contexts on performance and learning outcomes. *Journal of Educational Psychology*, *105*(4), 932-945.

Yuan, K. & Le, V. (2014). *Measuring deeper learning through cognitively demanding test items: Results from the analysis of six national and international exams.* Santa Monica, CA: RAND Corporation. http://www.rand.org/pubs/research_reports/RR483.html.

Appendix A

Hewlett PISA Study
CRESST Mathematics Framework: Cognitive Complexity

Level	Descriptors
Level 1	Task is primarily rote or procedural, requiring recall, recognition, or direct application of a basic concept, routine computation, algorithm or representation
Level 2	Task requires some mental processing and more than rote application of skill, concept or procedural and/or algorithmic tasks. Students often make decisions about how to approach the problem.
Level 3	Involves developing a solution strategy, and may have more than one possible answer Task often requires significant departure from traditional application of concepts and skills Solution strategy often involves working with multiple mathematical objects (numbers, expressions, equations, diagrams, graphs) or problem structures
Level 4	Task requires extended reflection, including complex problem solving, abstract reasoning, an investigation, processing of multiple conditions of the problem, and non-routine manipulations Task often requires extended time

Note: Webb's DOK framework (2007) as adapted by Herman, J., Buschang, R., La Torre Matrundola, D., & Wang, J. (2014)

CHAPTER FOUR

Evaluating Item Quality in Science Assessments

Jill Wertheim, Ph.D.
Nicole C. Holthuis, Ph.D.
Susan E. Schultz, Ph.D.*

* This work was completed in conjunction with the Stanford NGSS Assessment Project (SNAP) Team: Professor Jonathan Osborne (Principal Investigator), Professor Raymond Pecheone, Professor Emerita Helen Quinn, and Paolo Martin, doctoral candidate. See the SNAP homepage for more information: https://web.stanford.edu/group/ngss_assessment/cgi-bin/snapgse/

In recent years there have been fundamental changes in how K-12 science education is conceptualized. Specifically, expectations of what it means to be competent in doing science and understanding science have broadened.

> *"Beyond skillful performance and recall of factual knowledge, contemporary views of learning prize understanding and application of knowledge in use. Learners who understand can use and apply novel ideas in diverse contexts, drawing connections among multiple representations of a given concept. They appreciate the foundations of knowledge and consider the warrants for knowledge claims. Accomplished learners know when to ask a question, how to challenge claims, where to go to learn more, and they are aware of their own ideas and how these change over time." (National Research Council, 2007, p. 19)*

Many states are now translating this more contemporary view of science education into standards such as the Next Generation Science Standards (currently adopted by 15 states and numerous districts) or modified versions of the NGSS to accommodate individual state concerns.[1]

This shift in what it means to be science literate raises questions about what kinds of evidence would show that students have met the new expectations and how to elicit such evidence. With the new standards, mastery of science concepts no longer focuses solely on demonstrating factual or conceptual science knowledge. Instead the new standards prioritize science practices such as reasoning about phenomena using scientific evidence, drawing on scientific and engineering principles to solve problems, and reflecting on common themes that underpin big scientific ideas. As such, learning outcomes should be expressed as expectations that integrate both the competencies (i.e., skills and practices) and content understandings that students will be expected to demonstrate. Decades of development of assessments designed to evaluate content knowledge separately from science skills or practices have left teachers, education administrators, and assessment developers with few resources suitable for use in this new paradigm (Hannaway & Hamilton, 2008; NRC, 2011, 2014; Pellegrino, 2013).

...the new standards prioritize science practices such as reasoning about phenomena using scientific evidence, drawing on scientific and engineering principles to solve problems, and reflecting on common themes that underpin big scientific ideas.

Assessing mastery of these new conceptions of science learning calls for well-defined specifications for the development of large-scale summative assessment items and tasks. Preparing teachers to instruct and assess students on these newer conceptions of science learning underscores the need for instructional and assessment resources that synthesize 1) disciplinary core ideas (DCIs), 2) scientific practices, and 3) cross-cutting concepts. The development of assessments that attend to these three dimensions of science learning can, however, build

[1] See Heitin, L (2015, May). Districts out ahead of states in adopting science standards. *EdWeek*. May 5, 2015. Accessed on March 15, 2016 from http://www.edweek.org/ew/articles/2015/05/06/districts-out-ahead-of-states-in-adopting.html.)

upon existing banks of assessment items that, to varying degrees, assess student performance on each dimension separately.

The Stanford NGSS Assessment Project (SNAP) team conducted a review of existing assessments to identify robust examples of assessment items and tasks aligned with the NGSS. Many states are in a transitional phase in science assessment in which they in the process of developing new assessment systems and items while phasing out previous ones. Thus our review relied not only on gathering and reviewing released items from large-scale assessments, but also on newer items and tasks—both formative and summative—from smaller scale projects.

The review points to the specific characteristics of each assessment and task that align with the three dimensions of science and which aspects or dimensions need more attention. The outcomes of this work show how new assessments will need to differ from previous assessments and how existing assessments could be used as models for development of new assessments that are more fully aligned to the NGSS and similar frameworks. The full report of both the methodologies we used to conduct this review and our review findings can be obtained from the SNAP homepage.

Model assessments operationalize new standards of performance and communicate some of the fundamental shifts in how competency will be defined by the new standards.

From this review and analysis, we have identified four key features of assessments needed to support the newer vision of science learning and have provided examples of items and tasks to illustrate promising approaches to meeting the goals of the NGSS and similar state science frameworks.

A need for model assessments

Model assessments will be an important source of guidance for teachers, curriculum developers, and assessment designers who are preparing new science instructional and testing materials. Model assessments operationalize new standards of performance and communicate some of the fundamental shifts in how competency will be defined by the new standards. For example, students' understanding of photosynthesis can be assessed by items and tasks in very different ways. On the one hand, an end-of-unit assessment task could ask students to draw on the principles they have learned about photosynthesis and energy flows to construct their own explanation of energy flow through a system. On the other hand, it might ask students to simply select from a list of options to identify the name of the process that transfers heat energy from the sun to another object. These two very different assessment items would likely be supported by very different kinds of learning activities and experiences in the classroom.

Methods

To identify possible model science assessment items and tasks, our SNAP team conducted a three-stage review and analysis. 1) We established a bank of existing assessment items;

2) We developed review criteria, and 3) We conducted an in-depth analysis of a subset of items and tasks. We established a bank of existing assessments from which we could cull potential model items and tasks. The assessments we chose exhibited the following characteristics: they reflected the range of assessment formats that would be part of the new system of assessments (i.e., selected-response items, short constructed-response items, and short and extended performance tasks); they were designed for a variety of purposes (e.g., curriculum-embedded, external summative); and they were aligned to constructs relevant to the newer conceptions of science learning (i.e., disciplinary core ideas, science and engineering practices, and crosscutting concepts). The goal of this bank was not to be exhaustive, but rather to cast as wide net as possible to obtain a sense of what might exist for models of ways to probe each of the dimensions of the NGSS.

Second, we developed a list of criteria to be used for the analysis of the existing assessments. While the NGSS have not been adopted in all states, the ideas within the framework and standards are representative of newer conceptions of science education. Thus, our review criteria were largely based on the goals described in the following sources: the K-12 Framework for Science Education (NRC, 2012), the NGSS (Achieve, 2013), Developing Assessments for the Next Generation Science Standards (NRC, 2014), and the Guide to Implementing the Next Generation Science Standards (NRC, 2015). Additional criteria that derive from general principles for high quality assessment came from sources such as Knowing What Students Know (NRC, 2001) and the Stanford Center for Assessment, Learning, and Equity's (SCALE) criteria for evaluating performance assessments (unpublished). Below we present the criteria, questions, and coding categories that our reviewers compiled to analyze items and tasks in our assessment bank.

Criterion	Guiding Question and Coding Category (if applicable)
Grade Band	What grade band is the item or task designed for? • Elementary • Middle school • High school
Student Response	What does the item or task require the student to do? • Selected response — includes selecting a correct answer from a list of words, diagrams, pictures, or other visuals; matching, ordering, or indicating True/False • Constructed response — includes text responses from one word up to a paragraph or a visual response such as a drawing, map, simulation, diagram, or physical model • Performance task — includes tasks that require students to perform an activity (real or simulated), ranging from following step-by-step instructions to completing unstructured tasks where students have broad parameters within which to choose what to do and how to do it
Disciplinary Core Ideas (DCI)	What scientific content does the item or task address? • Matter and its interactions • Motion and stability: Forces and interactions • Energy • Waves and their applications in technologies for information transfer • From molecules to organisms: Structures and processes • Ecosystems: Interactions, energy, and dynamics • Heredity: Inheritance and variation of traits • Biological evolution: Unity and diversity • Earth's place in the universe • Earth's systems • Earth and human activity • Engineering design • Links among engineering, technology, science, and society
Science and Engineering Practices	What science and engineering practices does the item or task address? • Asking questions (for science) and defining problems (for engineering) • Developing and using models • Planning and carrying out investigations • Analyzing and interpreting data • Using mathematical computational thinking • Constructing explanations (for science) and designing solutions (for engineering) • Engaging in argument from evidence • Obtaining, evaluating, and communicating information
Crosscutting Concepts	What crosscutting concepts does the item or task address? • Patterns • Cause and effect: Mechanism and explanation • Scale, proportion, and quantity • Systems and system models • Energy and matter: Flows, cycles, and conservation • Structure and function • Stability and change
Number and Integration of NGSS Dimensions	How many of the three dimensions of NGSS (i.e., disciplinary core ideas, science and engineering practices, and crosscutting concepts) are being probed and to what extent are they integrated?
Extent of Focus on Big Ideas in Science	To what extent, from low, medium, to high, is the assessment item or task focusing on the big, essential concepts or skills that are central to the discipline and worth learning and evaluating?
Cognitive Demand Level	What is the level of cognitive demand, on a scale of low to high, required to complete the assessment item or task? • *Low* — Carry out a one-step procedure, for example, recall a fact, term, principle or concept or locate a single point of information from a graph or table. • *Medium* — Use and apply conceptual knowledge to describe or explain phenomena; select appropriate procedures involving two or more steps; organize/display data; interpret or use simple data sets or graphs. • *High* — Analyze complex information or data; synthesize or evaluate evidence; justify; reason given various sources; develop a plan or sequence of steps to approach a problem.

With regard to the last criterion in the table, we evaluated the cognitive demand level of each item or task using a construct adopted by developers of the 2015 PISA (Program for International Student Assessment). In the past, various cognitive demand schemes have been developed to evaluate assessment items and curricular tasks, such as Bloom's Taxonomy (Bloom, 1956), revised versions of Bloom's taxonomy (Anderson & Krathwohl, 2001; Marzano & Kendall, 2007), and Webb's Depth of Knowledge Levels (1997). The PISA measure of cognitive demand is an adapted version of Webb's Depth of Knowledge Levels (Webb, 1997). Webb's Depth of Knowledge Levels offer a taxonomy that identifies an item's cognitive demand from the verbal cues that are used—e.g., analyze, arrange, compare—as well as the expectations of the depth of knowledge required.

Thirdly, from the bank of 203 assessments gathered at stage one, we selected approximately 50 assessments that represent a range of grade levels, formats (i.e., selected response, constructed response, performance task), and scientific subject areas. The items and tasks within these assessments were then evaluated using the criteria discussed above.

Summary of Findings

Key Features for Designing NGSS-Aligned Assessments

Based on the NGSS framework and our review of science assessments from our bank of assessments, we identified four key features that should be included in the design of the next generation of science assessments.

Key Feature 1: Assessments should be aligned with and integrate multiple learning dimensions of science and engineering. The NGSS focuses on three learning dimensions: disciplinary core ideas, science and engineering practices, and crosscutting concepts. Using this vision of science learning, student proficiency with any given concept cannot be demonstrated just by knowledge of the relevant facts. Instead, students must be able to demonstrate how they can draw on their disciplinary content knowledge to engage in one or more science and engineering practices and apply or identify common themes that cut across science disciplines. For assessments to support this vision of three-dimensional learning, tasks must, for formative uses, probe each of the three dimensions in a way that exposes developing proficiency, and, for summative uses, tasks must probe how well students are able to integrate the three dimensions by applying their science and engineering knowledge to engage with phenomena using one or more of the practices.

Key Feature 2: Assessments should focus on the big ideas in science. The NGSS emphasize the big ideas and themes in science, and fine details are included only as they are central to making sense of the big ideas. In fact, the writers of the NGSS deliberately excluded some topics that have long been part of science classes because they considered the topics to be non-essential for contributing to students' understanding of the big ideas.

Key Feature 3: Assessments should address the full range of science and engineering practices. Science educators today have conceived of a more robust set of science and engineering practices than in the past. There are now science and engineering practices that have not previously been explicitly taught or assessed in science. For example, in the past, students were often assessed on their ability to conduct an investigation but were rarely given opportunities to demonstrate their ability to design an investigation on their own. Thus, new items and tasks need to be developed to assess this broader conception of what it means to "do" science and engineering.

Key Feature 4: Assessments should require students to demonstrate their reasoning and problem-solving skills. Well-designed performance tasks are able to probe much more deeply into students' reasoning and their ability to draw on their knowledge and skills as they are needed to investigate questions and solve problems. Thus, the next generation of science assessments will require extended response times and more creative and novel use of technology (e.g., computer simulations, video, and interactive platforms).

The following section explores each of the four recommended key features for designing NGSS-aligned assessments and provides some sample items and tasks that illustrate the key features along with analyses of those items and tasks.

Key Feature 1: Assessments should be aligned with multiple learning dimensions of science and engineering.

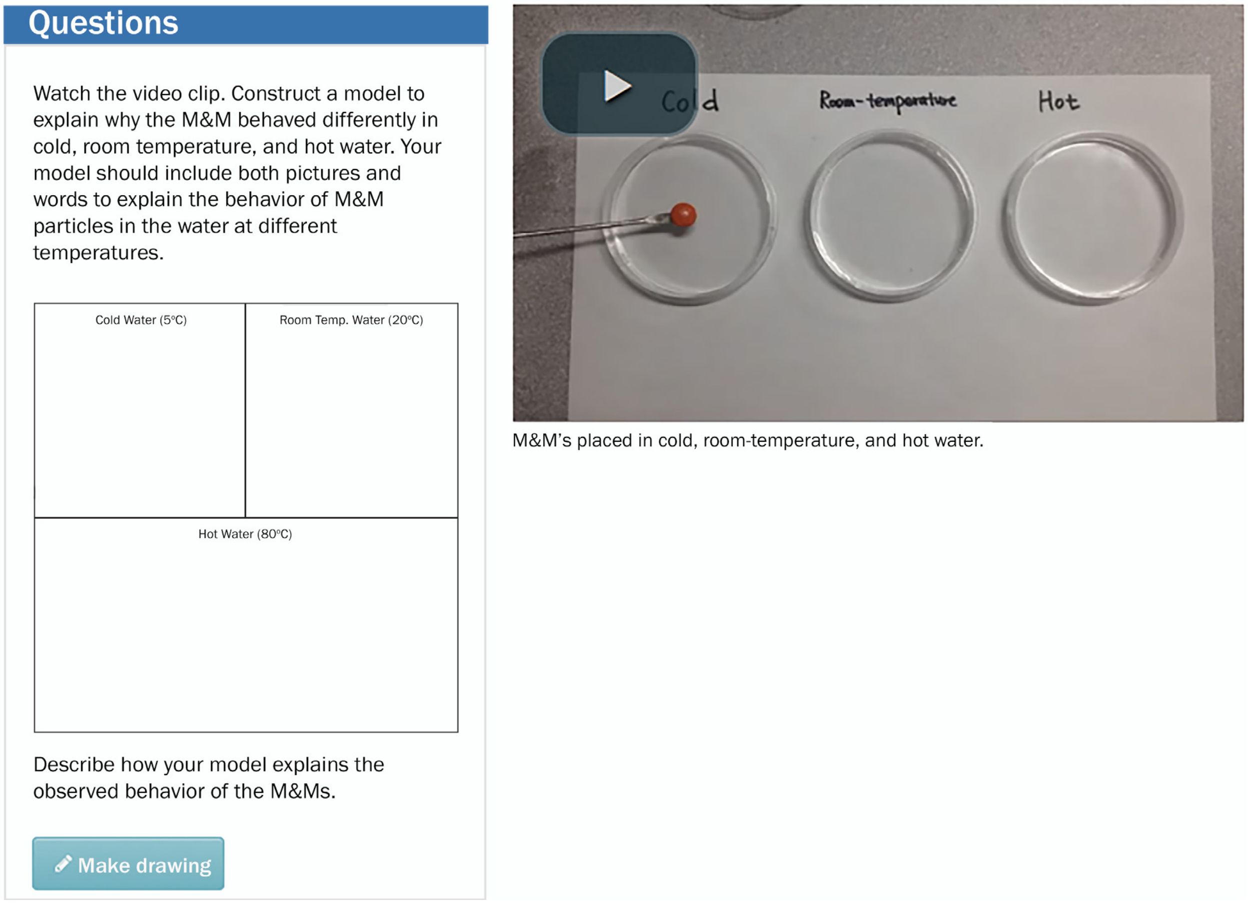
As mentioned earlier, the K-12 Science Framework (NRC, 2012) describes the importance of learning science in three dimensions (disciplinary core ideas, science and engineering practices, and crosscutting concepts):

> *"...in order to facilitate students' learning, the dimensions must be woven together in standards, curricula, instruction, and assessments. When they explore particular disciplinary ideas... students will do so by engaging in [science and engineering] practices... and should be helped to make connections to the crosscutting concepts." (p. 29)*

Certain contexts might call for assessing the dimensions separately, particularly in the case of formative assessments. But, in general, to evaluate the kind of engagement with scientific concepts in the manner described in the Framework, assessment tasks should be aligned to the three dimensions so that the targeted knowledge (disciplinary core idea) is integrated with a science/engineering practice and a crosscutting concept. A task that consists of multiple interrelated items might probe the three dimensions in their entirety, though each of the component items might probe just one or two dimensions (NRC, 2014).

Examples of Multi-dimensional Assessment Items and Tasks

Item Example 1



Questions

Watch the video clip. Construct a model to explain why the M&M behaved differently in cold, room temperature, and hot water. Your model should include both pictures and words to explain the behavior of M&M particles in the water at different temperatures.

Cold Water (5°C)	Room Temp. Water (20°C)
Hot Water (80°C)	

Describe how your model explains the observed behavior of the M&Ms.

M&M's placed in cold, room-temperature, and hot water.

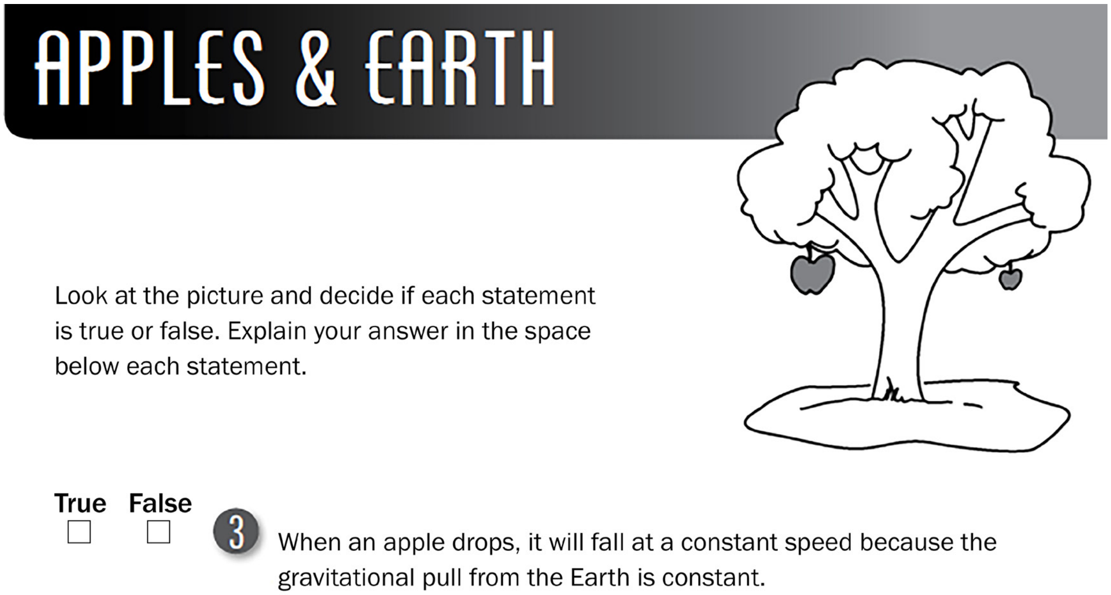
Figure 1. A two-part task for Grades 6-8 from the Concord Consortium's NGSS Assessment Project. Reproduced from States of Matter Task by the Concord Consortium, under Creative Commons Attribution 4.0 license.

Item Example 1 - Item Profile	
Source: Concord Consortium's Next Generation Science Assessment Project, Grades 6-8. Accessed on March 3, 2016 from http://authoring.concord.org/activities/4282	
Grade Band	Middle School
Student Response	Short performance task
Disciplinary Core Idea(s)	Energy
Science and Engineering Practice(s)	Developing and using models; Constructing explanations (for science)
Crosscutting Concepts	Cause and effect (potentially)
Number and Integration of NGSS Dimensions	2 dimensions, highly integrated
Extent of focus on Big Ideas in Science	High
Cognitive Demand Level	Medium

Our first example of a multi-dimensional task (Figure 1) is from the Concord Consortium's Next Generation Science Assessment for the topic "Energy" (Grades 6-8). For this task, students are asked to do the following: after watching a short video of a phenomenon (in which M&Ms are put into water at three different temperatures), they are asked to draw a model and provide an explanation of why there are differences in the way the M&Ms changed dependent on the different temperatures.

This multi-dimensional task requires that students use their physical science knowledge to develop a model that shows the cause of a phenomenon (DCI; science practice) and to construct a written explanation for the phenomenon (DCI; science practice). The scientific content and the science practice are also highly integrated: an understanding of particle motion is critical to developing the model—without such understanding, it is difficult to develop the model.

Notably, the crosscutting concept of "cause and effect" is implicit, but student understanding of this concept is not directly elicited in the task. In a fully three-dimensional task, this dimension would be addressed more explicitly by asking students to identify the mechanism that caused the phenomenon they observed.



APPLES & EARTH

Look at the picture and decide if each statement is true or false. Explain your answer in the space below each statement.

True □ False □

3 When an apple drops, it will fall at a constant speed because the gravitational pull from the Earth is constant.

Figure 2. WestEd, Making Sense of Science: Force and Motion. This one-time reproduction for educational purposes of this copyrighted material is covered by "fair use" guidelines. No rights to further reproduce this copyrighted material should be inferred. For more information, contact WestEd Publications at 562-799-5195 or email permissions@wested.org.

Item Example 2 – Item Profile	
Source: Daehler, K., and Folsom, J. (2014). Making Sense of SCIENCE Force & Motion: Formative Assessment Task Bank for Grades 6–8 (p.31). San Francisco: WestEd. All Rights Reserved. Project homepage: http://we-mss.weebly.com/	
Grade Band	Middle School
Student Response	Selected response (T/F); short constructed response
Disciplinary Core Idea(s)	Motion and stability: Forces and interactions
Science and Engineering Practice(s)	Constructing explanations (for science)
Crosscutting Concepts	None
Number and Integration of NGSS Dimensions	2 dimensions (potentially), highly integrated
Extent of focus on Big Ideas in Science	High
Cognitive Demand Level	Medium (potentially)

This second example item illustrates how sometimes it takes only a little revision to fully incorporate a second or third dimension in an item. Figure 2 above shows a short constructed-response item that is one of a series of items designed to provide insight into students' ideas about a physical science concept.

This short item requires students to both analyze a phenomenon in order to decide if a statement describing the phenomenon is true or false and explain their answer. As such, the cognitive demand level of the item is low to medium. This item can be revised, however, to address the practice of scientific explanation more specifically. If students are asked to specifically "construct an explanation using your knowledge of the earth's gravitational force" rather than "explain your answer" (a general direction that can prompt them to justify their answer based not on scientific ideas but on other reasoning), the item now addresses the DCI and an important scientific practice. Also, by specifically asking for a scientific explanation, the cognitive demand level of the task is raised from a low level to a medium level.

Key Feature 2: Assessments should focus on the big ideas in science.

Recent conceptions of science education have placed a greater focus on the most important and broadly explanatory ideas of science and de-emphasized the details that are not essential to the understanding of those ideas. For example, the NRC K-12 Science Framework states: "Specify big ideas, not lists of facts: Core ideas in the framework are powerful explanatory ideas, not a simple list of facts, that help learners explain important aspects of the natural world" (NRC, 2012, p. 254). Science assessments, in turn, need to reflect this shift by eliciting students' conceptual understanding of the big ideas of science.

Examples of Assessment Items that Focus on Big Ideas

Item Example 3

Brian and Joe are looking at the water boiling in the pan on the stove.

Brian says that the bubbles are made of air that gets pushed out of the water when the water gets hot. He argues that he knows there is air dissolved in water because fish are able to breathe the oxygen in the water.

Joe says that the bubbles are made of water that has turned into a gas — water vapor.

Joe agrees with Brian that fish are able to breathe oxygen in the water. But the pan has been boiling for 10 minutes and it is still bubbling just as much as it was at the beginning. If Brian was right, wouldn't the air be gone by now?

What idea is Joe arguing for? ____

____

What is the reason Joe gives to convince Brian he is right?

a. Fish are able to breathe the oxygen in the water

b. Bubbles are made of air

c. The pan has been boiling for 10 minutes and it is still bubbling

d. Hot water boils

Brian says that he knows that water is made of hydrogen and oxygen. The bubbles are caused by the water breaking down to produce hydrogen and oxygen that are both gases. These form bubbles like the gas in soda.

Joe is unconvinced. He remembers observing that the saucepan lid became covered in water drops as the water continued to boil.

How could he use this observation to convince Brian that he's wrong?

____

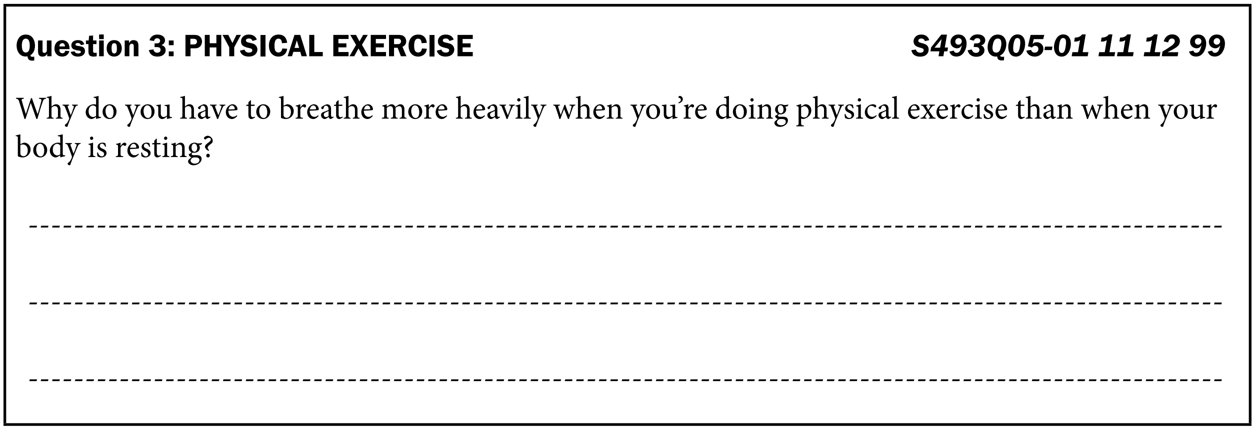
Figure 3. Group of items adapted from Assessments of Argumentation in Science. Reproduced with the permission of Jonathan Osborne.

Item Example 3 – Item Profile	
Source: Stanford Assessments of Argumentation in Science. “Bubbles in Water” task. Accessed on March 30, 2016 from: http://scientificargumentation.stanford.edu/assessments/bubbles-in-water/	
Grade Band	Middle School
Student Response	Selected response; constructed response
Disciplinary Core Idea(s)	Matter and its interactions
Science and Engineering Practice(s)	Engaging in argument from evidence
Crosscutting Concepts	None
Number and Integration of NGSS Dimensions	2 dimensions, highly integrated
Extent of focus on Big Ideas in Science	High
Cognitive Demand Level	High

Figure 3 shows a sequence of items, excerpted from the Assessments of Argumentation in Science. In this set of items, students are asked to clarify an argument about what causes bubbles in boiling water. Students are expected to combine content knowledge with an observation to analyze and evaluate an argument and construct a counter-argument. Together, the items focus on a common phenomenon to elicit students’ understanding of a foundational concept of science: states of matter. In total, the cognitive demand level of the task is considered to be “high.”

In this group of items, students are asked to use their knowledge of states of matter to engage in argumentation. While these items are provided as an example of assessing big ideas in science, they also provide an example of the challenges of doing so. The set of items requires a great deal of reading and reading comprehension. Students who are reading below grade level and/or learning English may have difficulty with these types of items.

Item Example 4



Question 3: PHYSICAL EXERCISE ***S493Q05-01 11 12 99***

Why do you have to breathe more heavily when you're doing physical exercise than when your body is resting?

Figure 4. Released item from a three-question item block on physical exercise, PISA Science Literacy assessments (2000 and 2006). Item in the public domain, reproduced from nces.ed.gov. Permission to excerpt OECD copyrighted materials for fair use purposes granted by OECD.

Item Example 4 – Item Profile	
Source: Program for International Student Assessment, Science Literacy Released Items 2000 and 2006. © PISA 2000, 2006. Accessed from: http://nces.ed.gov/surveys/pisa/educators.asp	
Grade Band	High School
Student Response	Constructed response
Disciplinary Core Idea(s)	From molecules to organisms: Structures and processes
Science and Engineering Practice(s)	None
Crosscutting Concepts	None
Number and Integration of NGSS Dimensions	1 dimension
Extent of focus on Big Ideas in Science	High
Cognitive Demand Level	High

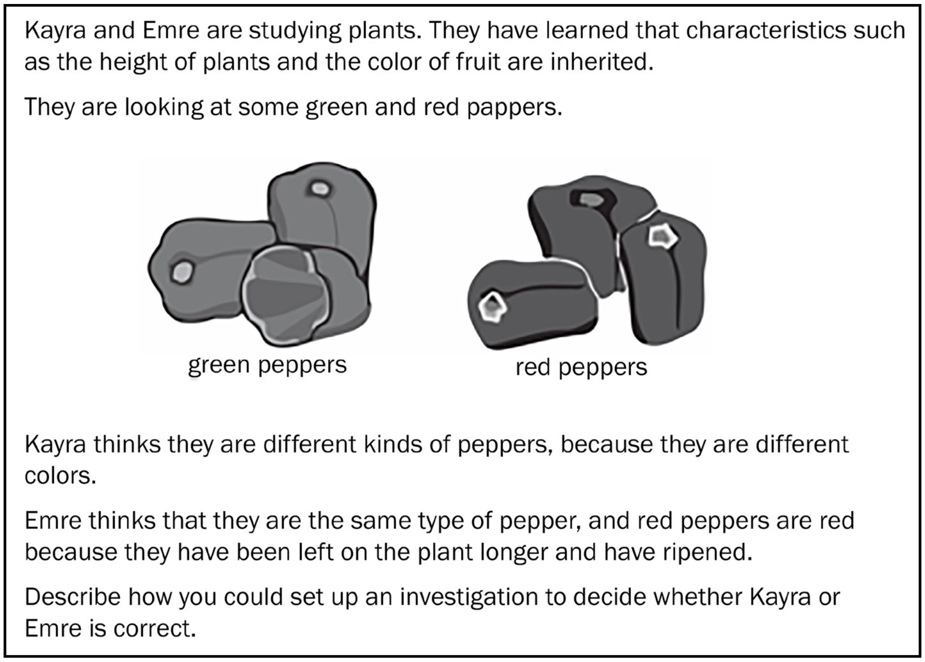
Our second example of an item that focuses on a big idea in science is a concise constructed-response item (Figure 4) that requires students to demonstrate deep conceptual understanding of the role of cellular respiration and its inputs and outputs. Rather than focusing on the discrete steps or molecular reactions that occur during cellular respiration, the open-ended item elicits understanding of an essential big idea in science: animals require oxygen to "produce" energy. In contrast to the task above, the language demands are lower, as the item focuses solely on the disciplinary core idea.

Key Feature 3: Assessments should address the full range of science and engineering practices.

In contrast to previous state and national science standards, the science education community is now envisioning a more robust set of science and engineering practices that 1) include some practices that were not explicitly defined in previous standards; 2) combine engineering practices with those of science; and 3) break the practices down into different component practices from previous standards (e.g., the practice "Planning and carrying out investigations" includes component practices such as evaluating methods or tools for the collection of data and revising an experimental design). Each of these differences produces a gap between existing assessments that are aligned to a set of goals different from those of the NGSS and what is needed for assessing the NGSS.

Indeed, there are abundant existing assessments that evaluate specific details of planning an investigation, such as identifying independent and dependent variables, and there are also many tasks that ask students to evaluate the methods or design of an investigation. Few existing assessments, however, engage students in planning their own investigation, including identifying the data they would need to collect in order to support a scientific claim.

Item Example 5



Kayra and Emre are studying plants. They have learned that characteristics such as the height of plants and the color of fruit are inherited.

They are looking at some green and red pappers.

green peppers red peppers

Kayra thinks they are different kinds of peppers, because they are different colors.

Emre thinks that they are the same type of pepper, and red peppers are red because they have been left on the plant longer and have ripened.

Describe how you could set up an investigation to decide whether Kayra or Emre is correct.

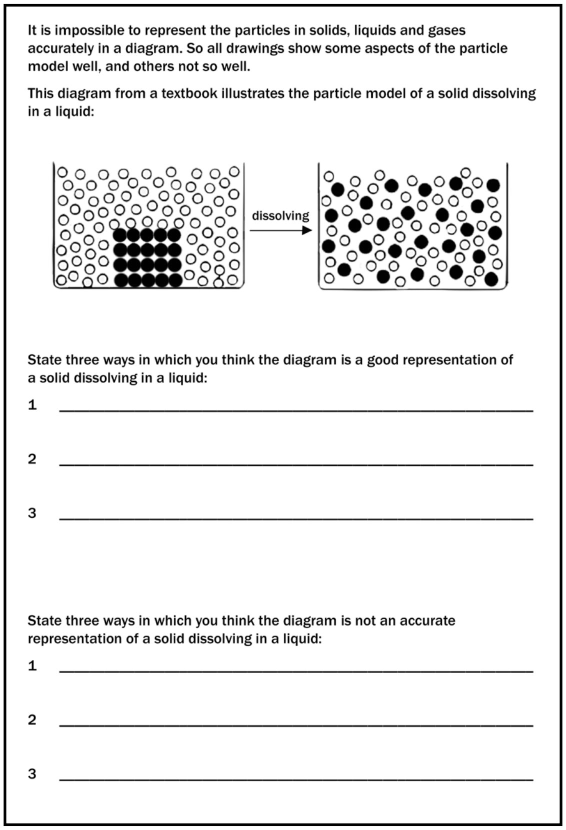
Figure 5. A constructed-response item about designing an investigation to test two ideas "Investigation of Green/Red Peppers" – Item S042297 from TIMSS, 2011). (Item in the public domain, reproduced from nces.ed.gov with fair-use permission of the International Association for the Evaluation of Educational Achievement (IEA).

Item Example 5 – Item Profile	
Source: TIMSS (Trends in International Mathematics and Science Study), 2011 Assessment. *Complete 2011 TIMSS 8 science set*, p.39. “Investigation of Green/Red Peppers” (Item S042297). Accessed from: https://nces.ed.gov/timss/pdf/TIMSS2011_G8_Science.pdf Copyright © 2011 International Association for the Evaluation of Educational Achievement (IEA). Publisher: TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College, Chestnut Hill, MA and International Association for the Evaluation of Educational Achievement (IEA), IEA Secretariat, Amsterdam, the Netherlands.	
Grade Band	Middle School (specifically 8th grade)
Student Response	Constructed response
Disciplinary Core Idea(s)	Heredity: Inheritance and variation of traits
Science and Engineering Practice(s)	Planning and carrying out investigations
Crosscutting Concepts	None
Number and Integration of NGSS Dimensions	2 dimensions, highly integrated
Extent of focus on Big Ideas in Science	High
Cognitive Demand Level	High

This example item (Figure 5) is a constructed-response item that addresses the science practice of planning investigations. The item requires students to design an investigation to determine which of two arguments can be used to correctly explain a common phenomenon (i.e., some peppers are red and some are green). Students design an investigation that would investigate two different explanations for the phenomenon, which means an investigation that requires multiple steps. For these reasons, the item has a high cognitive demand level.

As is, this item is fairly one-dimensional. It focuses on a science practice without requiring significant content understanding. In addition, an item such as this one might be more effective at eliciting correct responses if it were explicit about the elements of an investigation that must be included (e.g., data to collect, variables to control, question to investigate), but the task provides an example of a promising foundation for assessing the practice of designing an investigation.

Item Example 6



It is impossible to represent the particles in solids, liquids and gases accurately in a diagram. So all drawings show some aspects of the particle model well, and others not so well.

This diagram from a textbook illustrates the particle model of a solid dissolving in a liquid:

dissolving

State three ways in which you think the diagram is a good representation of a solid dissolving in a liquid:

1 ____

2 ____

3 ____

State three ways in which you think the diagram is not an accurate representation of a solid dissolving in a liquid:

1 ____

2 ____

3 ____

Figure 6. An assessment item that targets students' ability to evaluate a model (developed by The University of York EPSE Project). Reproduced with permission from Robin Millar.

Item Example 6 – Item Profile	
Source: The University of York EPSE Project	
Grade Band	Grade not indicated, but aligns to Middle School Performance Expectations
Student Response	Constructed response
Disciplinary Core Idea(s)	Matter and its interactions
Science and Engineering Practice(s)	Developing and using models
Crosscutting Concepts	Stability and change (potentially)
Number and Integration of NGSS Dimensions	2 dimensions, highly integrated
Extent of focus on Big Ideas in Science	High
Cognitive Demand Level	High

Our second example item (Figure 6) is a constructed-response item that focuses on another science practice—developing and using models. This practice has been examined and articulated in the research literature, but has rarely been the focus of assessment items. In fact, there are few existing science items that directly elicit students' ability to develop, evaluate, or use a scientific or engineering model. Specifically, this example item targets the science practice of critiquing or evaluating models. The task portrays a very general molecular representation, with few words and a simple visual.

The item could be made richer by adding a scenario in which students need to apply their scientific content knowledge to develop a third model that represents some other change to the state of the matter in the container. The item could also assess another dimension of students' understanding by asking specific questions to elicit the crosscutting concept of stability and change.

Key Feature 4: Assessments should require students to demonstrate their reasoning and problem-solving skills.

As described above, many science educators envision a scientific literacy in which students gain the intellectual tools needed to make sense of scientific phenomena in the world around them.

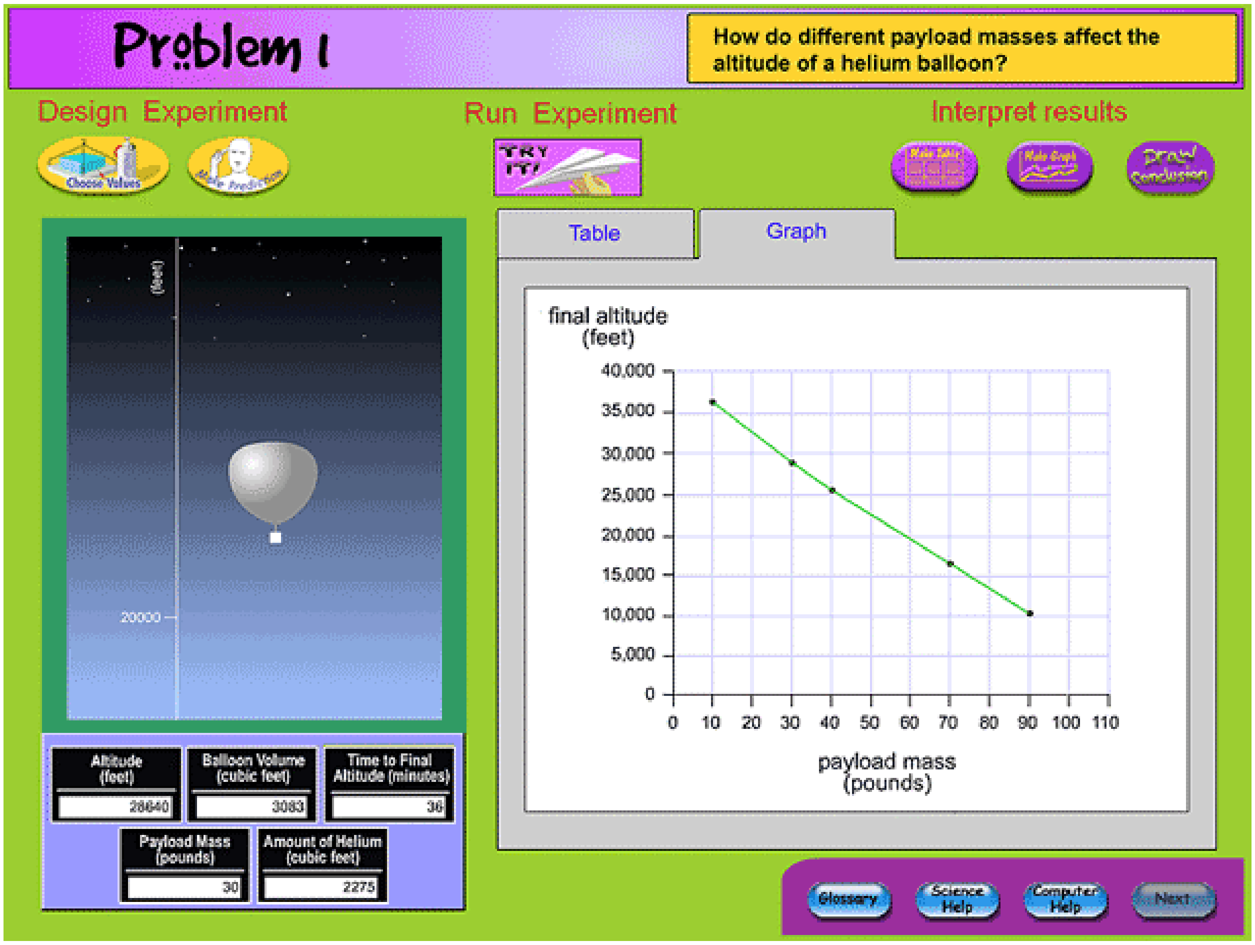
> *"By the end of the 12th grade, students should have gained sufficient knowledge of the practices, crosscutting concepts, and core ideas of science and engineering to engage in public discussions on science-related issues, to be critical consumers of scientific information related to their everyday lives, and to continue to learn about science throughout their lives. They should come to appreciate that science and the current scientific understanding of the world are the result of many hundreds of years of creative human endeavor." (NRC, 2012, p. 9)*

Assessments can make explicit the kind of scientific reasoning that students should be prepared to do and can provide insight into students' progress toward that goal. Assessments, therefore, should require students to engage in scientific reasoning. That is, assessments should present the opportunity for students to perform activities relevant to investigating phenomena and solving problems. For example, students should have opportunities to demonstrate their ability to devise methods to collect and analyze data, use models to evaluate their analyses, and make claims and justify their responses. Clusters of short constructed-response items can provide brief glimpses of how students conduct these activities, and in some settings these are the closest approximation of students' reasoning about phenomena and problems that is feasible. Performance tasks can better elicit this extended reasoning across the three dimensions, and can provide opportunities to observe students drawing on multiple elements of each dimension as they are needed to solve problems and answer questions. However, although investigative performance tasks provide a wealth of evidence about students' learning and their ability to demonstrate science practices, they require additional time and equipment that is not always feasible for large-scale assessment.

...students should have opportunities to demonstrate their ability to devise methods to collect and analyze data, use models to evaluate their analyses, and make claims and justify their responses.

In an attempt to solve this problem, computer-based assessments offer a promising avenue for engaging students in science and engineering practices without requiring much time or equipment.

Item Example 7



Problem 1

How do different payload masses affect the altitude of a helium balloon?

Altitude (feet)	Balloon Volume (cubic feet)	Time to Final Altitude (minutes)
28640	3083	36

Payload Mass (pounds)	Amount of Helium (cubic feet)
30	2275

Figure 7. A computer-based simulation task in which students run a simulation to collect data for three closely related investigations about a helium balloon. Item in the public domain, reproduced from the National Assessment of Educational Progress (NAEP), Technology Rich Environment Study, 2007 with fair-use permission from NCES.

Item Example 7 – Item Profile	
Source: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), Technology Rich Environment Study, 2007, Physical Simulation. Accessed on March 3, 2016 from https://nces.ed.gov/nationsreportcard/studies/tba/tre/sim-description.aspx	
Grade Band	Middle School (specifically 8th grade)
Student Response	Selected response; computer-based performance task
Disciplinary Core Idea(s)	Matter and its interactions
Science and Engineering Practice(s)	Analyzing and interpreting data
Crosscutting Concepts	Patterns (potentially)
Number and Integration of NGSS Dimensions	2 dimensions (potentially 3), somewhat integrated
Extent of focus on Big Ideas in Science	Medium
Cognitive Demand Level	High

The NAEP computer-based task shown in Figure 7 has a structured process for simulating trials to collect data for several variables related to the height to which a helium balloon can rise. Students have to make predictions, run the simulated trials, analyze patterns in data, and select an appropriate explanation in a series of selected-response (multiple-choice) items that require students to make sense of the data they collected. As such, the cognitive demand level of the task is described as medium to high.

The online platform provides opportunities for students to select variables to test, run trials, and analyze a data table and graph to draw conclusions about observed patterns. In doing so, students are able to engage in the practice of investigation in ways that are not afforded by a traditional assessment.

The online platform provides opportunities for students to select variables to test, run trials, and analyze a data table and graph to draw conclusions about observed patterns. In doing so, students are able to engage in the practice of investigation in ways that are not afforded by a traditional assessment.

The task would be even stronger if it addressed a crosscutting concept. The task has the potential to tap into student understanding of the crosscutting concept of patterns if questions are added that explicitly prompt students to identify patterns in the data and the nature of those patterns.

Fire Extinguisher

Some fires can be extinguished by smothering them with carbon dioxide gas (CO^2). A company is designing a fire extinguisher that uses the chemical reaction between vinegar and baking soda to produce carbon dioxide. Since the fire extinguisher must produce the gas quickly in order to put out a fire, the designers need your help in studying variables that affect how much carbon dioxide this reaction produces in a certain amount of time.

There are several variables that may affect the rate of carbon dioxide production in the fire extinguisher, such as the amount of baking soda, the concentration of vinegar solution, and the temperature of the vinegar solution. You will investigate two of these variables using a plastic bottle as a model fire extinguisher.

Your model fire extinguisher should only hold a maximum of 10 cc (cubic centimeters) of vinegar solution. Note: 1 cc=1 mL.

Your task:

Part I: You and your partner will design and conduct an experiment to determine how the *amount of baking soda* affects how much carbon dioxide is produced in a *certain amount of time.*

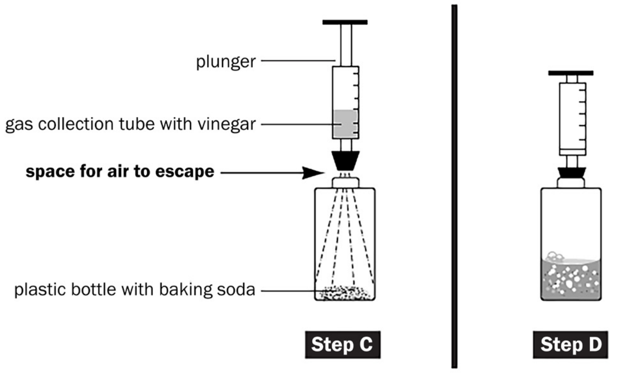
Part II: You and your partner will design and conduct an experiment to determine how *another variable* that you choose affects how much carbon dioxide is produced in a *certain amount of time.*

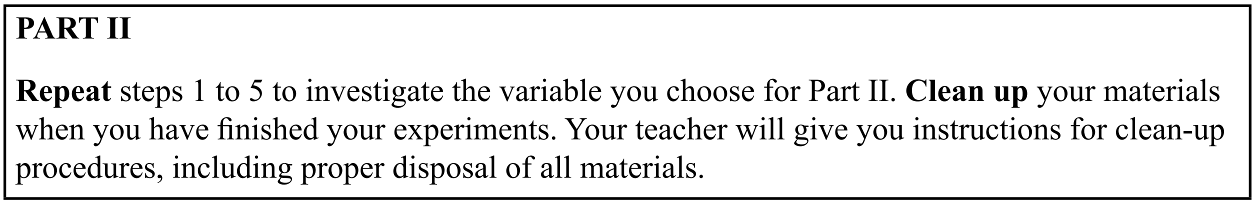
PART I

1. **State** the problem you are going to investigate. Clearly identify the independent and dependent variables that will be studied. Write your problem statement on page 5.

2. **Design** an experiment to solve the problem. Your experimental design should match the statement of the problem, should control for variables and should be clearly described so that someone else could replicate your experiment. Use a control and perform multiple trials, as appropriate. Write your experimental design on page 5.

   Use the diagram below to help you set up your experiment. **Remember, your model fire extinguisher should only hold a maximum of 10 cc of solution. Note: 1 cc = 1 mL.**


plunger
gas collection tube with vinegar
space for air to escape
plastic bottle with baking soda
Step C
Step D



PART II

Repeat steps 1 to 5 to investigate the variable you choose for Part II. **Clean up** your materials when you have finished your experiments. Your teacher will give you instructions for clean-up procedures, including proper disposal of all materials.

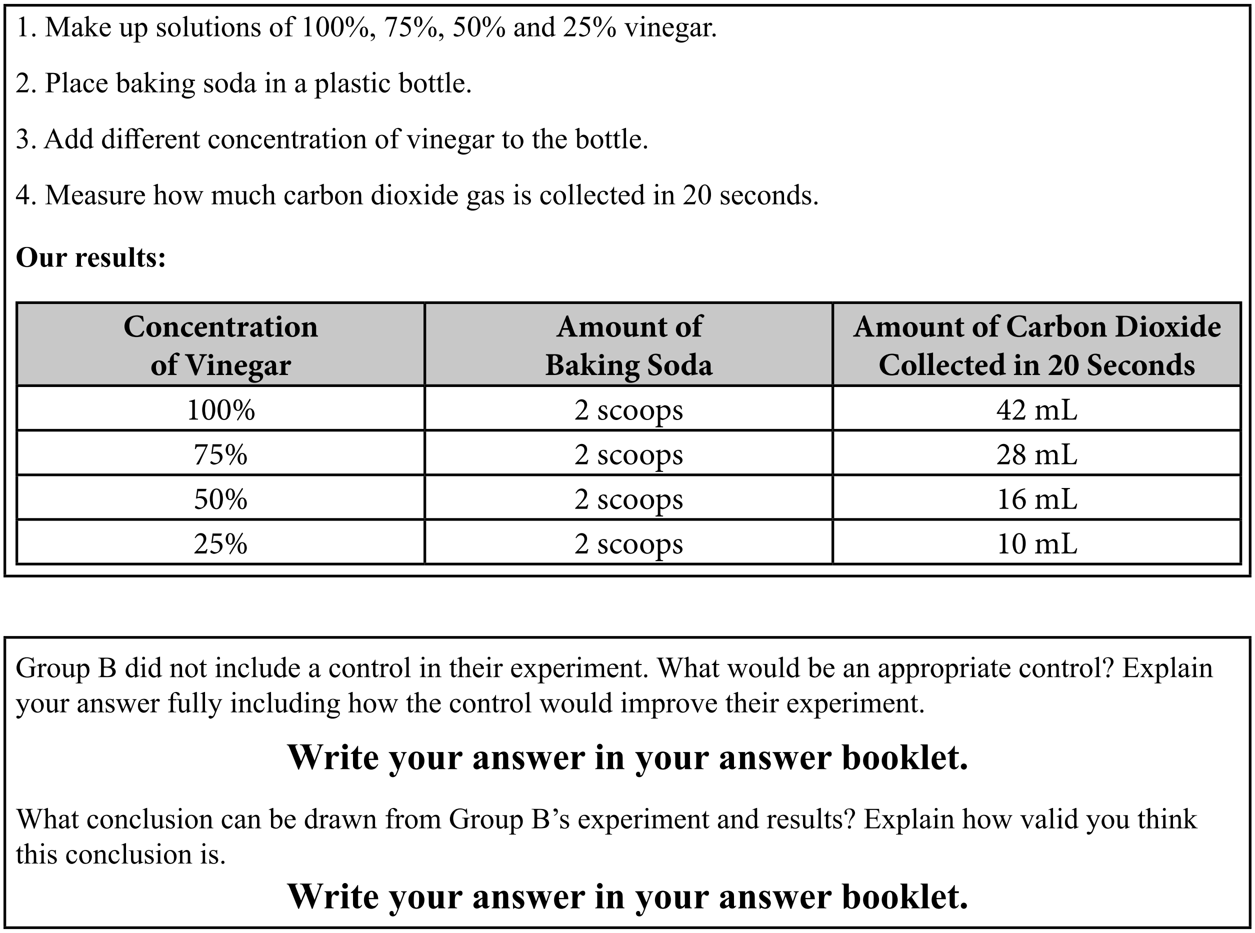
Figure 8a. Parts I and II from a grade 9-10 Science Curriculum-Embedded Task from the Connecticut Academic Performance Testing program © Connecticut State Department of Education. Reproduced with the permission of the Connecticut State Department of Education.

Item Example 8 – Item Profile	
Source: Connecticut Department of Education, Science Curriculum-Embedded Task (Connecticut Academic Performance Testing Science Released Items, 2004, p. 147-152.)	
Grade Band	High School
Student Response	Constructed response; performance task
Disciplinary Core Idea(s)	Matter and its interactions
Science and Engineering Practice(s)	Planning and carrying out an investigation
Crosscutting Concepts	None
Number and Integration of NGSS Dimensions	2 dimensions, highly integrated
Extent of focus on Big Ideas in Science	Medium to high
Cognitive Demand Level	High

This final example demonstrates how a longer performance task gives students the opportunity to use science practices to reason with evidence while balancing the need to give students some structure to guide their investigation. It does so by separating the design of the investigation and the analysis of the results. For example, Figure 8a-c shows a multi-stage item developed by the Connecticut Department of Education. In the first two stages (Figure 8a), students are given a constrained scenario and are asked to describe a problem and design/ conduct investigations to address the problem. In part I they design and conduct an experiment to "determine how the amount of baking soda affects how much carbon dioxide is produced." In part II, they design and conduct an experiment to determine how another variable (that they choose) affects how much carbon dioxide is produced.

After students describe and conduct their own investigations, they complete a more open-ended part of the assessment. They are presented with a data table from the investigations of two other groups (one of which is provided in Figure 8b below). Students are asked to analyze the results, critique the design of the investigation, and draw conclusions based on their analysis.

Group B carried out the following experiment.



1. Make up solutions of 100%, 75%, 50% and 25% vinegar.
2. Place baking soda in a plastic bottle.
3. Add different concentration of vinegar to the bottle.
4. Measure how much carbon dioxide gas is collected in 20 seconds.

Our results:

Concentration of Vinegar	Amount of Baking Soda	Amount of Carbon Dioxide Collected in 20 Seconds
100%	2 scoops	42 mL
75%	2 scoops	28 mL
50%	2 scoops	16 mL
25%	2 scoops	10 mL

Group B did not include a control in their experiment. What would be an appropriate control? Explain your answer fully including how the control would improve their experiment.

Write your answer in your answer booklet.

What conclusion can be drawn from Group B's experiment and results? Explain how valid you think this conclusion is.

Write your answer in your answer booklet.

Figure 8b. Open-ended questions from a grade 9-10 Science Curriculum-Embedded Task from the Connecticut Academic Performance Testing program © Connecticut State Department of Education. Reproduced with the permission of the Connecticut State Department of Education.

This performance task draws on multiple science practices needed to reason through observations made from simple investigations around physical science content. As such, it has a high cognitive demand level. The investigation at the beginning of the task enables students to become familiar with the problem, but the rest of the task is standardized with a common investigation and data set, making sure that students' performance on the rest of the practices being probed is not dependent on their investigation in parts I and II.

References

Achieve, Inc. (2013). *Next Generation Science Standards*. Washington, DC: author.

Anderson, L. W., & Krathwohl, D. R. (Eds.) et al. (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's Taxonomy of Educational Objectives.* Boston, MA: Pearson Allyn & Bacon.

Bloom, B. S. (Ed.), Engelhart, M. D., Furst, E. J., Hill, W. H., Krathwohl, D. R. (1956). *Taxonomy of educational objectives. Handbook I: The cognitive domain.* New York, NY: David McKay Co Inc.

Hannaway, J., & Hamilton, L. (2008). *Performance-based accountability policies: Implications for school and classroom practices.* Washington, D.C.: Urban Institute and RAND Corporation.

Heitin, L. (2015) Districts out ahead of states in adopting science standards. *Education Week*, 34 (29). Retrieved from http://www.edweek.org/ew/articles/2015/05/06/districts-out-ahead-of-states-in-adopting.html.

Marzano, R. J., & Kendall, J. S. (Eds) (2007). *The new taxonomy of educational objectives.* 2nd ed. Thousand Oaks, California: Corwin Press.

National Research Council (2001). *Knowing what students know: The science and design of educational assessment.* Washington, DC: The National Academies Press. doi: 10.17226/10019

National Research Council (2007). *Taking science to school: Learning and teaching science in grades K-8.* Washington, DC: The National Academies Press. doi: 10.17226/11625

National Research Council (2011). *Assessing 21st century skills* (summary of a workshop). Washington, DC: The National Academies Press. doi: 10.17226/13215

National Research Council (2012). *A framework for K-12 science education: Practices, crosscutting concepts, and core ideas.* Washington, DC: The National Academies Press. doi: 10.17226/13165

National Research Council (2014). *Developing assessments for the Next Generation Science Standards.* Washington, DC: The National Academies Press. doi: 10.17226/18409

National Research Council (2015). *Guide to implementing the Next Generation Science Standards.* Washington, DC: The National Academies Press. doi: 10.17226/18802

Pellegrino, J. W. (2013). Proficiency in science: Assessment challenges and opportunities. *Science*, *340*(6130), 320–323. doi: 10.1126/science.1232065

Webb, N. (1997). *Criteria for alignment of expectations and assessments on mathematics and science education* (Research Monograph Number 6). Washington, DC: CCSSO

CHAPTER FIVE

Evaluating Item Quality in History Assessments

by Daisy Martin, Ph.D.

Introduction

History/social studies education in the United States is currently undergoing change as standards, instructional materials, and educators focus more on disciplinary skills and content, rather than solely on knowledge of historical specifics. In history, this is sometimes characterized as a move towards historical thinking rather than memorization, or a focus on not only what we know about the past, but also how we know it. At least 40 states' history/social studies standards include historical thinking, as do national and cross-state frameworks such as the C3 Framework for Social Studies State Standards and the Common Core State Standards for Literacy in History/Social Studies (American Historical Association, 1997; College Board, 2015; Martin, Maldonado, Schneider and Smith, 2011; National Council for the Social Studies, 2013; National Governors Association Center for Best Practices, Council of Chief State School Officers, 2010.) In effect, this shift means that expectations for student learning in history/social studies are becoming more complex and challenging as students are expected to "do" history and social studies (even if in limited ways) rather than only recall subject-specific facts.

Doing history means that students engage with history as an interpretive and analytic discipline.

Doing history means that students engage with history as an interpretive and analytic discipline. They encounter and evaluate historical arguments and narratives, and corroborate and synthesize varied types of evidence to construct legitimate interpretations of past events. To think historically, students learn to question and analyze primary and secondary sources, and to contextualize those sources as well as historical events and specifics. They use concepts such as perspective, significance, change and continuity, and evidence to understand and make sense of the past. All of this "doing" involves both disciplinary skills and knowledge, as students work with specific times, places, themes, sources, and events to craft, tell, or evaluate evidence-based historical arguments and stories. Doing social studies also integrates knowledge and skill as students apply disciplinary concepts and tools to investigate contemporary questions and problems and then communicate their conclusions and recommendations.

This paper aims to help stakeholders know more about how to craft and select standardized history items that align with these complex disciplinary competencies. We focus on history items rather than social studies items because of their greater availability and other reasons noted below. We explain a tool that includes criteria for classifying the cognitive demand of existing history items and take a close look at a small set of those items to help readers identify design features of items that measure important disciplinary competencies.

Methods

We started with the 50 State Assessment Collection compiled by our research team. The collection of available sample released items for large-scale assessments of history and social studies than for the other three core disciplines. This reflects the current state of standardized testing in the U.S.; that is, history/social studies is not assessed by states as frequently as

English language arts, math, and science. In 2010, a year when federal policy encouraged states to use “high-quality assessments” that were tied to college- and career-ready standards (U.S. Department of Education, 2009, p.2) only twenty-six states tested students in history (Martin et al., 2011). Thirteen of those tests contained only multiple-choice questions, and thirteen of those tests asked students to do some writing (Martin et al., 2011). This means that the bank of states’ sample released items is relatively limited in item format. Additionally, unlike reading, science, and mathematics, which are assessed in the Program for International Student Assessment, there is no international test in history/social studies. Nor could we select items from Smarter Balanced and PARCC, the two national Common Core testing consortia, because they do not include history/social studies items as part of their assessments. However, states are not the only ones testing students’ historical understanding. The College Board, a private nonprofit corporation, and the Department of Education’s National Center for Education Statistics (NCES) do so as well, so we were able to include sample released items from their tests (AP U.S. History and the National Assessment of Educational Progress, respectively) in addition to assessments from the 50 State Assessment Collection.

In 2010, a year when federal policy encouraged states to use “high-quality assessments” that were tied to college- and career-ready standards (U.S. Department of Education, 2009, p.2) only twenty-six states tested students in history (Martin et al., 2011).

Next, we worked to identify items in the collection that show promising approaches to assessing historical understanding. We surveyed the available items and chose to focus on items from high school level assessments as these were more available and also could demand more complex thinking. We also decided to focus on U.S. History items as there were more samples for this subject area and it is currently the most frequently assessed topic in the history/social studies scope and sequence across states. Having identified all of the available high school U.S. history items in the bank, we then reviewed each item, looking for items that went beyond testing the recall of single historical specifics and that represented a variety of item formats. Altogether, we identified items from 12 state tests, the NAEP U.S. History test, and the Advanced Placement (AP) U.S. History test that met these criteria. We then narrowed this set to focus on the highest quality items. This included reviewing scoring tools, blueprints, and ancillary item materials as available. These helped us evaluate the alignment between item, scoring tool, and intended measurement target(s). In several cases, a lack of alignment between the scoring tool and an item’s demands disqualified it from our final set of six items.

At the same time that we were reviewing and selecting items, we were seeking a tool to analyze and provide a quantitative score for the cognitive complexity of each test item in the set. While other core subjects use existing tools to code the cognitive complexity of items, a widely used, validated tool is not available in history. We decided to create a discipline-specific tool that would identify criteria for analyzing the cognitive demands and complexity of large-scale history test items.

In the subsequent rounds of selection and analysis of the items, we circled between analysis of the items and the development, application, and revision of our tool to assess an item’s cognitive complexity. Two assessment experts reviewed and applied three prototypes of the tool to sample items to inform the tool's final content. (See Appendix A for the final version of the tool, entitled **Evaluating Cognitive Complexity in History Items**.) This process also helped narrow the set of test items to those that show promising approaches to assessing significant and complex disciplinary understandings.

Selection of Test Items

The final set of items that we selected to present as promising items represents the range of released item types that currently exist on large-scale history tests. At one end of that range is the selected-response item. These items (also known as multiple-choice items) can ostensibly assess different aspects of historical understanding, and can, in more promising cases, include authentic historical materials such as primary source excerpts. But selected-response items do not truly show what and how students understand given that the only thing students produce is a blackened bubble. (For a more extensive discussion on the limitations of multiple-choice history items, see Reich, 2009, and Wineburg, 2004.) At the other end of the spectrum are free-response and document-based essay questions. These extended constructed-response questions allow us to see much more of what and how a student understands. We found few technology-enhanced items in the overall item bank. In one case technology allowed for different kinds of selected-response items, such as using drag and drop to sequence historical events or match historical specifics[1], but otherwise we did not find examples where technology was used to impact the type or demands of an item. For our final set, we deliberately chose released items that had actually been administered to students and, when appropriate, included scored samples of students' responses. The scored samples made it possible to see more of what the item elicited and what was valued in a student response.

At the other end of the spectrum are free-response and document-based essay questions. These extended constructed-response questions allow us to see much more of what and how a student understands.

It is important to note that we found tests from the College Board and NCES to have more high quality, cognitively challenging items than the state tests, which are dominated by multiple-choice recall questions. In addition, it should be noted that the AP U.S. History test was recently redesigned to focus on historical thinking, and so the sample items for that test generally demand more authentic and complex disciplinary thinking. However, this report focuses on a broader sample of assessment items than the AP test represents, as AP is currently directed at only a segment of the student population, and only a few administered items have been released.

Our final set of items includes two multiple-choice items, one short constructed-response

[1] See Pearson Education, Inc. (2014). Colorado Measures of Academic Success: Science and Social Studies. Retrieved January 26, 2016, from http://www.pearsonaccess.com/cs/Satellite?c=Page&childpagename=Colorado%2Fco-PALPLayout_v2&cid=1205794393643&pagename=coPALPWrapper

item, one thematic block, and two extended constructed-response items. We first explain the Evaluating Cognitive Complexity in History Items tool developed for analysis of the cognitive complexity each item. We then provide a brief summary of findings and implications of our item evaluation. The individual item analyses follow, wherein readers will find a score generated by the Evaluating Cognitive Complexity tool for each item, as well as a qualitative analysis. These qualitative analyses investigate the item, scoring system, and, as available, the item's intended measurement targets, to answer three questions: What does this item measure? What are the item's strengths? How could this item be improved?

Evaluating Cognitive Complexity in History Items Tool

This tool, available in Appendix A, was created to analyze the relative cognitive demand of the spectrum of items currently being used on large-scale history assessments in the U.S. to assess students' historical understanding. It is, therefore, focused on the world of what is, rather than what could be. It was built using existing research and theory from the field of teaching, learning, and assessing history and historical thinking.

The Tool - Theoretical Foundation

This tool is built using the following three established principles in that literature:

(1) The integration of historical knowledge, skills, and thinking is at the crux of historical understanding.

First, historical knowledge and skills are both assumed to be necessary to challenge students to demonstrate their historical understanding. While some assessment items may only measure knowledge of a historical specific or time, and others may only test whether a student can employ a historical skill, an item that requires the demonstration of both historical knowledge and skill to produce an answer is more reflective of the authentic demands of the discipline and is likely to be more challenging.

(2) Disciplinary literacy is key to historical understanding.

Second, disciplinary literacy—that is, specific ways of reading and writing in the discipline of history—is integral to understanding and knowing history (Moje et al, 2004; Monte-Sano, 2010, 2011; Shanahan and Shanahan, 2008; Wineburg, 1991, 2001; Wineburg and Martin, 2004). Wineburg (1991) identified three key ways historians approach text that are specific to their training and activities: sourcing, contextualization, and corroboration. Similarly, historians learn to write in specific ways, and making evidence-based arguments is essential to the discipline (Mink, 1987; Monte-Sano, 2010; Schneider and Zakai, 2016).

(3) Assessment item format shapes cognitive demand.

Large-scale history assessment items come in varied formats and the format of an assessment item significantly influences its potential cognitive demand. Answering selected-response items that ask students to blacken an oval will necessarily be less demanding than items that ask students to produce a written response. Likewise, requiring the production of a couple of sentences or a paragraph is likely less demanding than requiring the production of an entire essay. This third principle is not about what is being assessed—it is about how it is being assessed.

These three principles are the theoretical foundation for the Evaluating Cognitive Complexity tool that we developed to generate a quantitative measure of the cognitive complexity of existing large-scale test items in history. The tool focuses on two aspects of an item: Design Features and Disciplinary Demands. Each of these aspects includes three dimensions that are scored. The scores for these six dimensions are combined to generate one quantitative score for the cognitive complexity of each our selected items. Using our tool, an item could receive a total of 1 to 17 points, the higher number reflecting the greatest cognitive complexity. Below, we briefly explain the specific rationale for each of these six dimensions.

The Tool - Dimensions of Item Evaluation

Aspect 1: Design features

The design features we evaluated include 1) the item's format, 2) the types of materials that students encounter and work with in an item, and 3) whether the item measures historical knowledge and skills in an integrated way.

(1) Item format

There were four types of items in the bank of released test items: selected response (aka multiple choice), short constructed response, thematic block, and extended constructed-response. A "thematic block" is a set of items that all address the same historical topic (Lazer, 2015). This block can include any combination of selected-response items and short or extended constructed-response items. Using our tool, a selected-response item received the fewest points given that its design is less cognitively demanding than the other types. Constructed-response and thematic block items received the same number of points. While a thematic block of items could be more challenging and cognitively demanding than a constructed-response item alone, nesting items within a shared topic can provide some additional support for students, either in being able to contextualize the topic or using material from one question to help answer another in the set. Both a short constructed-response item and the thematic block items are assumed, in generic terms, to be less demanding than an extended constructed-response item. Extended constructed-response items are essays (free-response and document-based), and they routinely require more from a student both in product and thinking than a short constructed-response item.

(2) Number and type of historical sources

The second dimension in this category is the type of materials that students encounter and work with in the test item. Source work is essential to doing history (e.g., Bain, 2005; Holt, 1995; Levstik and Barton, 2005; Stearns, Seixas, and Wineburg, 2000; VanSledright, 2002) and also enables creating an item that allows for disciplinary skills to be demonstrated (e.g., Ercikan and Seixas, 2015; Monte-Sano, 2011). If the item uses a single historical source, it is credited with one point; if it uses multiple historical sources, it earns another point as it is necessarily more cognitively demanding for a student; and if it uses multiple sources of varied formats and genres, it earns yet another point. In history, one reads different genres of sources differently (e.g., primary, secondary, diary, political cartoon) and different genres mean that a student must be able to make sense of each kind of genre (see http://teachinghistory.org/best-practices/using-primary-sources; Pope, 2003).

(3) Integration of historical knowledge and skills

An item earns one point for cognitive complexity if it requires that students use both historical knowledge and skills in an integrated way. This approach is not only reflective of the discipline being tested, but it also means that students are required to integrate different aspects of the discipline, thus increasing the item's cognitive complexity.

Aspect 2: Disciplinary demands

The disciplinary demands we evaluated include 1) the kind of historical reading required by the item, 2) the kind of historical writing and argumentation required by the item, and 3) the kind of historical knowledge students must use to complete the item.

(1) Required historical reading/analysis

The first dimension of this category addresses the kinds of historical reading/analysis students have to do to successfully complete the item. While this dimension is related to the Design Features: Use of Materials dimension explained above, this dimension differs from that one in its focus on what students have to do with those materials, not just whether particular materials are included in an item. Reading in the domain of history has specific features and requires sourcing, contextualizing, and corroborating (Hynd, Holschuh, and Hubbard, 2004; Reisman, 2015; Wineburg, 1991, 2001; Wineburg, Martin, and Monte-Sano, 2013). While researchers and assessment developers frame historical reading in other ways as well (e.g., AP U.S. History frames historical reading in four ways), all agree that historical reading is important to historical understanding and requires more than just generic comprehension skills. This tool captures the importance of historical reading and analysis partly by distinguishing between reading a historical source as fact (i.e., an immutable report) and reading it as a "trace of the past" (Ercikan, Seixas, Lyons-Thomas and Gibson, 2015). For example, a student who reads a primary source as fact can simply pull information from it without question or regard to the author's purpose or point of view; whereas a student who reads a primary source as a "trace of the past" must interrogate it to consider it as a voice from the past with accompanying motives and perspective.[2]

...all agree that historical reading is important to historical understanding and requires more than just generic comprehension skills.

Reading historical sources as traces of the past necessarily increases the level of cognitive

[2] This is a relatively blunt difference. To distinguish and assess particular types of historical reading, more distinctions would be necessary (Ercikan et al., 2015). However, this overarching distinction was sufficient in this instance to capture an item's demands.

complexity when compared to reading solely for comprehension and an item that requires this earns one point. Similarly, if students must synthesize or corroborate multiple sources to successfully answer the test item, the cognitive complexity of the item is higher than if analyzing a single source—earning another point. In addition, if these multiple sources represent different perspectives on the same historical phenomenon or present contrasting information that the student must navigate, this also raises the cognitive complexity of the item. Each of these three ways of analyzing sources or reading historically earns a point on our tool, so a single test item could earn three points for cognitive complexity regarding historical reading.

(2) Required historical writing/argumentation

The second dimension of disciplinary demands relates to constructing a historical argument and, in these sample items, is most commonly connected to the task of historical writing. A point is awarded for each of the following four characteristics, as each of these demands increases the level of cognitive complexity of an item: 1) establishing an evidence-based claim; 2) explaining and integrating evidence into the argument; 3) including a counterclaim; and 4) requiring complexity. The characteristic "requiring complexity" is a broadly conceived feature that is meant to capture requirements that get at qualitative distinctions between the types of arguments that students construct. For example, do students have to qualify their argument in the face of contrary or limited evidence? Do they have to extend the argument beyond the immediate topic? (Schneider and Zakai, 2016; College Board, 2015) If an item requires students to go deeper or broader in their argument, the item earns a complexity point. Constructing a plausible argument from available evidence is a key historical task, and each of these four demands can be embedded within such a task. When an item includes more than one of these four demands, it likely poses a question with multiple plausible answers and invites student interpretation.

We identify three categories of historical knowledge that, unlike the existing focus on decontextualized and discrete facts in too many large-scale tests, require students to make connections to a larger framework and/or other historical specifics. This contextualized knowledge more accurately reflects how specifics are used in history...

(3) Required use of historical knowledge

The third dimension of disciplinary demands concerns the types of historical knowledge that students have to use to complete the item. We identify three categories of historical knowledge that, unlike the existing focus on decontextualized and discrete facts in too many large-scale tests, require students to make connections to a larger framework and/or other historical specifics. This contextualized knowledge more accurately reflects how specifics are used in history (Immerwahr, 2008; National Research Council, 2005) and also increases the cognitive complexity of an item. These three categories of historical knowledge are as follows: 1) contextualized factual knowledge, 2) disciplinary concepts, and 3) knowledge external to the item.

An item that demands that students use "contextualized factual knowledge" requires that a student make connections to a larger narrative or argument. Items that demand contextualized factual knowledge might require that students employ background knowledge of a time or historical phenomenon, use specific historical examples, or make connections between historical specifics. Another type of historical knowledge is the use of a disciplinary concept. If the student must apply a disciplinary concept—such as periodization, change and continuity, or causation—to complete the test item, this also increases cognitive complexity. Sometimes these two types of knowledge are available to the student in the item, for example, in the form of a timeline or a brief orienting passage. Other times a student must use knowledge they have independent of the test or its materials. If the item demands that students use external knowledge, this also increases the complexity of the item and merits the item another point.

Summary of Findings and Implications for Large-Scale History Assessment Design

There are some persistent issues in assessing complex competencies in history, including reducing construct-irrelevant factors such as high reading demands or identifying the appropriate necessary background knowledge (Seixas and Ercikan, 2015; Reisman, 2015). Developing items that represent authentic disciplinary tasks that students will complete and that fit within specific testing constraints can also be a design puzzle. These factors may partly explain the relatively narrow range of items currently being used on large-scale standardized history tests. The persistence and ubiquity of multiple-choice items on these tests is also likely due, in part, to their relative affordability and a conceptualization of this discipline as focused on learning important names, dates, and places.

However, the possibilities of other item types exist—items that aim to measure more complete or complex historical understanding. In the next section of this paper, we share our analysis of each of the six promising history items that we selected. While these items are not necessarily ideal, each illustrates design features that support cognitively complex assessment in history. We also urge state and district assessment directors and test developers to think beyond the limited set of item formats and measurement targets presented here to imagine new possibilities. Examining examples of validated history/social studies items not currently being used on standardized tests that provide clear information about students' achievement can help inform those possibilities[3], as can consulting international examples and the C3 Framework for Social Studies.

Looking forward, we believe the following set of questions can help test developers and other stakeholders aim for more cognitively complex assessments that will support the teaching and learning of the historical discipline in all of its complexity.

1. Does the item prioritize domain-specific skills and knowledge? If the item requires reading and writing, does it require historical reading and/or writing? Does the scoring system

[3] This is a relatively blunt difference. To distinguish and assess particular types of historical reading, more distinctions would be necessary (Ercikan et al., 2015). However, this overarching distinction was sufficient in this instance to capture an item's demands.

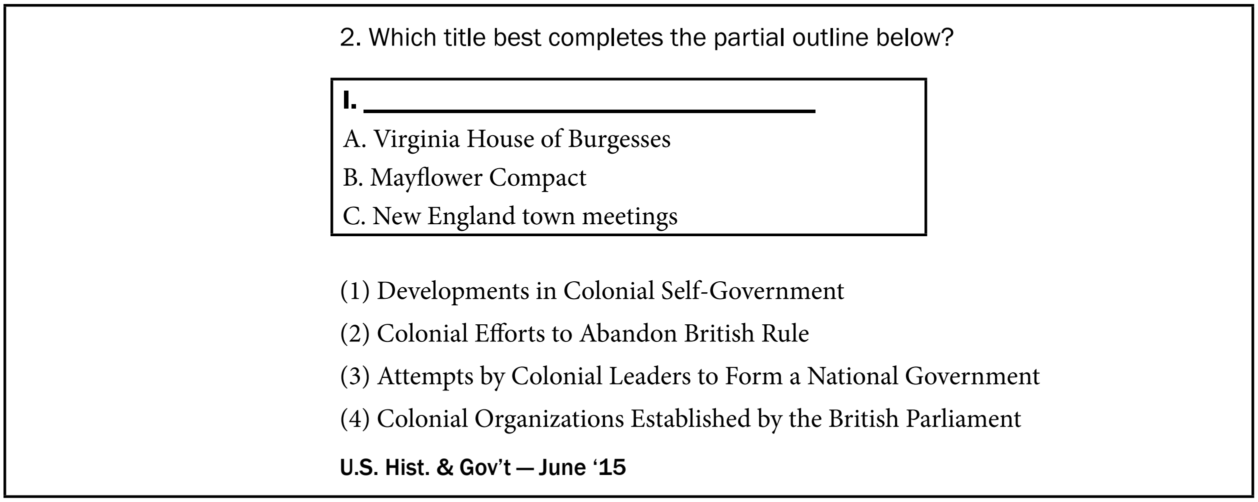
focus on disciplinary competencies?

2. Does the item represent a facsimile of disciplinary work, both in the materials used and what students are asked to do for the item? When an item requires students use historical specifics, does it demand they use knowledge that is connected or contextualized?
3. Are the materials used in the item carefully selected and prepared to minimize confounding factors such as reading ability and necessary background knowledge?
4. Are there multiple pathways to a correct answer or multiple correct responses, or does the item assess knowledge of one specific?
5. Is the item "balanced" in construction? Is the item designed to maximize the elicitation of evidence regarding the disciplinary measurement targets? Does it simplify or scaffold aspects of the item that demand skills or knowledge that are not being measured?

Review of Selected Assessment Items

In this section, we present six sample items from large-scale assessments that were selected to illustrate item design features that support more complete measurement of complex historical understandings and skills. Two are selected-response items, one is a short constructed-response item, one is an extended constructed-response item, one is a thematic block, and one is a performance task DBQ (document-based question essay). We describe what each item measures and particular design features that support cognitive complexity.

Item Example 1



2. Which title best completes the partial outline below?

I. ______________________
A. Virginia House of Burgesses
B. Mayflower Compact
C. New England town meetings

(1) Developments in Colonial Self-Government
(2) Colonial Efforts to Abandon British Rule
(3) Attempts by Colonial Leaders to Form a National Government
(4) Colonial Organizations Established by the British Parliament

U.S. Hist. & Gov't — June '15

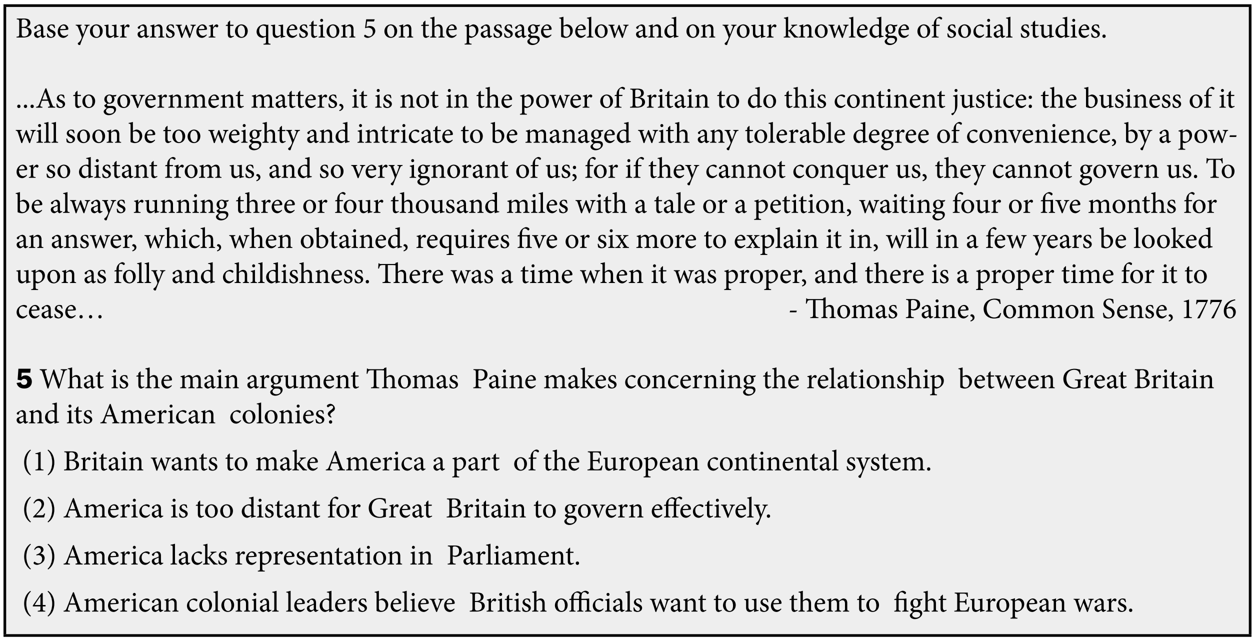
Figure 1. Selected-response item from the New York State Regents Exam in U.S. History and Government, June 2015. Item reproduced with permission in accordance with Terms of Use granted by NYSED.

Item Example 1 – Item Profile	
Source: From the New York State Education Department. *Regents Exam in United States History and Government*, p. 2. Internet. Available from http://www.nysedregents.org/USHistoryGov/615/ushg62015-examw.pdf; Accessed 1, February, 2016.	
Item Type	Selected Response (Multiple Choice)
Grade Level	High School
Cognitive Complexity Score	2

This selected-response item is focused on assessing students' historical knowledge. Students could use knowledge of the Virginia House of Burgesses, Mayflower Compact, New England town meetings, or more generally, Colonial America, to select the right answer (i.e., (1) Developments in Colonial Self-Government). While this item is similar to many other selected response items in that it focuses on historical knowledge, its strength is that ignorance of one discrete specific will not mean that students fail the question. There are multiple things a student could know that would help answer the question, from being able to explain any of the specifics listed in A through C to a broader understanding of Colonial American governments or the antecedents of the American Republic. This variety of paths towards the correct answer, including specific historical knowledge and more general knowledge of a trend or era, is a strength of the question as it allows students to know different, albeit related, facts to succeed on the question.

A student would also need to know how an outline works and that the answer must be a title that categorizes all three specifics. This fill-in-the-outline approach, essentially creating a hierarchy of ideas, allows the required knowledge (general or specific) needed to answer the question to be more contextualized and connected. And while readers may be concerned that the item is dominated by academic vocabulary, this vocabulary is core to understanding the past.

Item Example 2



Base your answer to question 5 on the passage below and on your knowledge of social studies.

...As to government matters, it is not in the power of Britain to do this continent justice: the business of it will soon be too weighty and intricate to be managed with any tolerable degree of convenience, by a power so distant from us, and so very ignorant of us; for if they cannot conquer us, they cannot govern us. To be always running three or four thousand miles with a tale or a petition, waiting four or five months for an answer, which, when obtained, requires five or six more to explain it in, will in a few years be looked upon as folly and childishness. There was a time when it was proper, and there is a proper time for it to cease… - Thomas Paine, Common Sense, 1776

5 What is the main argument Thomas Paine makes concerning the relationship between Great Britain and its American colonies?

(1) Britain wants to make America a part of the European continental system.

(2) America is too distant for Great Britain to govern effectively.

(3) America lacks representation in Parliament.

(4) American colonial leaders believe British officials want to use them to fight European wars.

Figure 2. Selected-response item from New York State Regents Exam in U.S. History and Government, June 2015. Item reproduced with permission in accordance with Terms of Use granted by NYSED.

Item Example 2 - Item Profile	
Source: From the New York State Education Department. *Regents Exam in United States History and Government*, p.3. Internet. Available from http://www.nysedregents.org/USHistoryGov/615/ushg62015-examw.pdf; accessed 1, February, 2016.	
Item Type	Selected Response (Multiple Choice)
Grade Level	High School
Cognitive Complexity Score	3

This selected response question is focused on measuring students' ability to understand and identify an argument in a primary source. Students read a short excerpt from Thomas Paine's Common Sense to then select the argument that Paine makes in that excerpt. One strength of this item is that it engages students in reading an important historical source. Another is that students must use and understand the passage to select the right answer (choice #2), and cannot just use the stem of the question as there is more than one plausible answer from which to choose. For example, option 3 (America lacks representation in Parliament) is factually correct, but not what Paine argues in this particular passage. This means that the item is measuring students' ability to identify Paine's main argument in this passage, as it is designed to do, rather than more general knowledge of Paine's treatise. General background knowledge of Paine, Common Sense, or the year 1776, will likely also help a student make sense of the excerpt by being able to more quickly identify who "this continent" and "us" refers to. Because this item focuses on measuring a student's ability to identify an author's argument in a primary source, it has a disciplinary lens and requires historical reading. We can imagine a more extended use of this source in a test with the addition of questions that ask students to contextualize or use background knowledge to further analyze Paine's purpose or impact in writing Common Sense.

Item Example 3

Source H: This is a quotation taken from an interview with Mike Royko, who became a journalist in Chicago.

> *I was nine years old when the war started. It was a typical Chicago working-class neighborhood. It was predominantly Slavic, Polish…In those days they put out extras. I remember the night the newsboys came through the neighborhood…Germany had invaded Poland: '39. It was the middle of the night, my mother and father waking. People going out in the streets in their bathrobes to buy the papers. In our neighborhood with a lot of Poles, it was a tremendous story.*
>
> *Suddenly you had a flagpole. And a marker. Name went on the marker, guys from the neighborhood who were killed. Our neighborhood was decimated. There were only kids, older guys, and women. Suddenly I saw something I hadn't seen before. My sister became Rosie the Riveter. She put a bandanna on her head every day and went down to this organ company that had been converted to war work. There was my sister in slacks. It became more than work. There was a sense of mission about it. Her husband was Over There…*
>
> *There was the constant idea that you had to be doing something to help. It did filter down to the neighborhood: home-front mobilization. We had a block captain…*
>
> *The world was very simple. I saw Hitler and Mussolini and Tojo: those were the villains. We were the good guys…*

Using information from the quotation in Source H, describe two important ways the Second World War influenced the actions and beliefs of people at home.

Figure 3. Short constructed-response item from the National Assessment of Educational Progress (NAEP), 2010 United States history. Item in the public domain, reproduced from the National Assessment of Educational Progress (NAEP), 2010 U.S. History, with fair-use permission from NCES.

Item Example 3 – Item Profile	
Source: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2010 U.S. History [Item# 2010-12H11 #14]. Accessed on February 18, 2016 from NAEP Questions Tool http://nces.ed.gov/nationsreportcard/itmrlsx/	
Item Type	Short Constructed Response *[categorized as an Extended Constructed response by NCES]*
Grade Level	High School, Grade 12
Cognitive Complexity Score	7

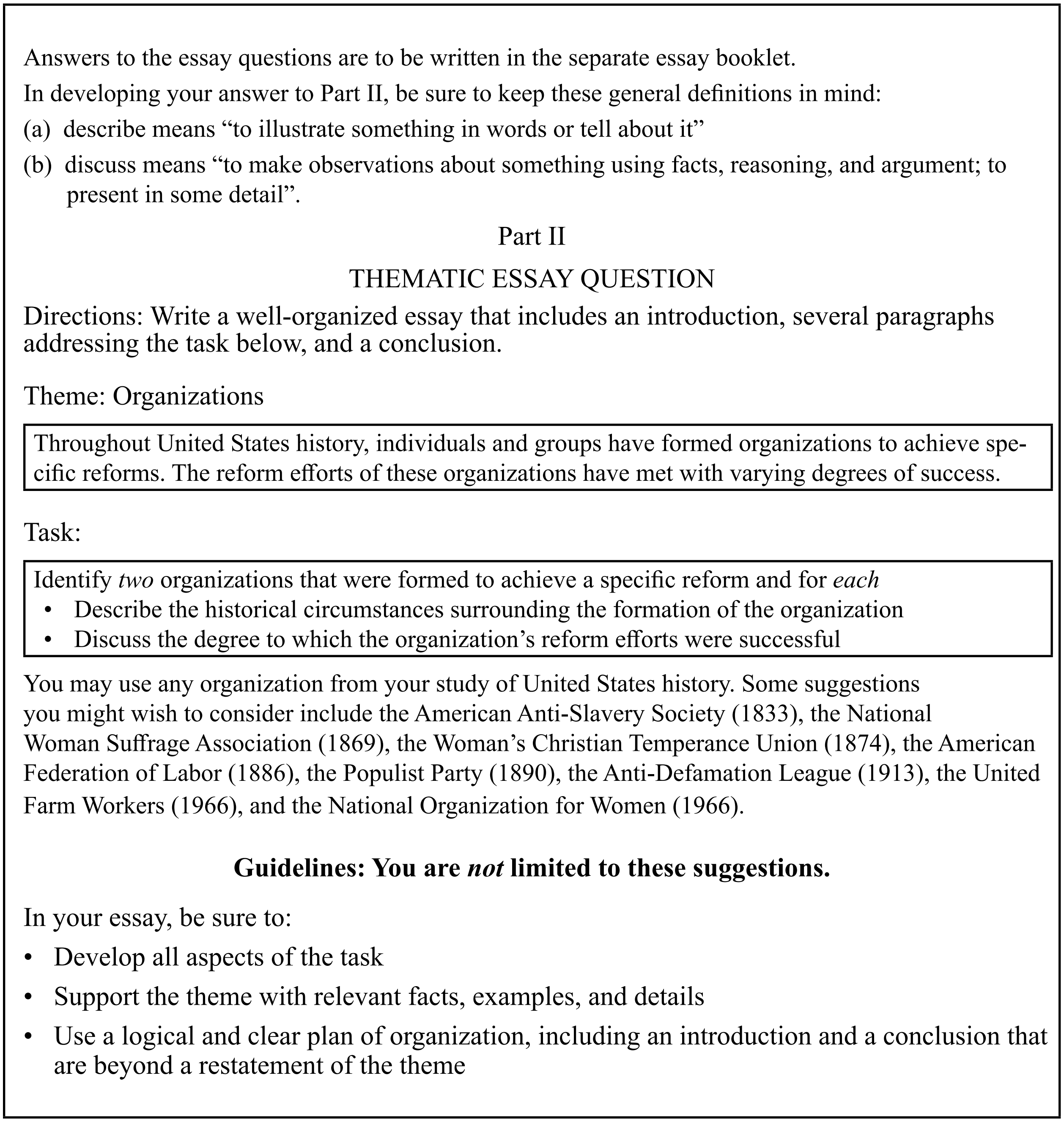
This short constructed-response item is focused on measuring students' reading and writing skills in response to a historical source. Students must use information from that source to identify and describe two impacts of World War II on people at home. The item primarily measures students' ability to write a complete response to a prompt and employ textual evidence to support an historical explanation. It also measures whether students understand the source and the demands of the prompt. While background knowledge of the period is not absolutely necessary for the students to succeed on this item, it will help students understand the primary source more quickly, accurately, and deeply (i.e., it will help students to understand specifics and phrases like "Rosie the Riveter," "Over There," "Hitler, Mussolini and Tojo").

...the prompt is a text-based question—one that demands students return to the text and use it even if they can answer the general question using just background knowledge. This reflects the evidentiary nature of history and is an important aspect of historical writing.

This item has several strengths, including that students encounter and read a primary source and use it to craft an explanation and learn more history—important disciplinary skills. Additionally, the prompt is a text-based question—one that demands students return to the text and use it even if they can answer the general question using just background knowledge. This reflects the evidentiary nature of history and is an important aspect of historical writing. Another strength of the item is the alignment between scoring criteria (shown below) and the demands of the prompt. The scoring criteria clearly focus on whether the student has completely answered the prompt and used textual evidence to support the description. "The response describes two key ways in which the Second World War affected the homefront AND supports each with clear (explicit or implicit) evidence from the quotation."

There are two features of this item that we recommend considering. First, the prompt asks for "important" impacts on both "actions and beliefs." The word "important" may be confusing as all of the impacts mentioned by Royko could count as such and this signifier does not matter to a student's score. The phrase "actions and beliefs" could help a student understand that impacts can be of various types, but it could also be confusing to students, leading them to think they must address both actions and beliefs, which is not required by the scoring criteria. Second, this item focuses on disciplinary writing rather than reading as students are expected to read this account as fact and not expected to interrogate it. (Notably, the student does not have access to complete information about when the text was produced or for whom.) This emphasis on writing is appropriate given the item's measurement focus, but consideration should be given to coupling the item with one or more additional questions that measure more disciplinary thinking and reading. For example, a student could be asked about the limitations of this single source for answering this question, a question that would recognize that generalizing from a single account to a more general explanation of two important impacts of WWII demands caution.

Item Example 4



Answers to the essay questions are to be written in the separate essay booklet.

In developing your answer to Part II, be sure to keep these general definitions in mind:

(a) describe means "to illustrate something in words or tell about it"

(b) discuss means "to make observations about something using facts, reasoning, and argument; to present in some detail".

Part II

THEMATIC ESSAY QUESTION

Directions: Write a well-organized essay that includes an introduction, several paragraphs addressing the task below, and a conclusion.

Theme: Organizations

Throughout United States history, individuals and groups have formed organizations to achieve specific reforms. The reform efforts of these organizations have met with varying degrees of success.

Task:

Identify *two* organizations that were formed to achieve a specific reform and for *each*

- Describe the historical circumstances surrounding the formation of the organization
- Discuss the degree to which the organization's reform efforts were successful

You may use any organization from your study of United States history. Some suggestions you might wish to consider include the American Anti-Slavery Society (1833), the National Woman Suffrage Association (1869), the Woman's Christian Temperance Union (1874), the American Federation of Labor (1886), the Populist Party (1890), the Anti-Defamation League (1913), the United Farm Workers (1966), and the National Organization for Women (1966).

Guidelines: You are *not* limited to these suggestions.

In your essay, be sure to:

- Develop all aspects of the task
- Support the theme with relevant facts, examples, and details
- Use a logical and clear plan of organization, including an introduction and a conclusion that are beyond a restatement of the theme

Figure 4. Extended constructed-response item from the New York State Regents Exam in U.S. History and Government, June 2015. Item reproduced with permission in accordance with Terms of Use granted by NYSED.

Example Item 4 – Item Profile	
Source: From the New York State Education Department. *Regents Exam in U.S. History and Government,* p.13. Internet. Available from http://www.nysedregents.org/USHistoryGov/615/ushg62015-examw.pdf; Accessed 1, February, 2016. The scoring key and rating guide may be found at http://www.nysedregents.org/USHistoryGov/615/ushg62015-rg1.pdf p.3-5.	
Item Type	Extended Constructed Response
Grade Level	High School, Grade 12
Cognitive Complexity Score	10

This thematic essay item assesses students' historical knowledge and skills in an integrated way. Students choose two reform-minded organizations to describe and discuss in writing, drawing from their knowledge of U.S. history. The writing task asks students to both describe the "historical circumstances" of each organization's creation and make an argument about the degree of success for that organization's reform efforts. To do this successfully, students must know the origins of their two chosen organizations and specific conditions, events, or actors that prompted people to organize and come together to fight for a particular cause or set of causes. They must also know about specific results (or lack thereof) of the organization's efforts. In this way, this item measures multiple kinds of historical knowledge, including broad knowledge of at least one historical period, related historical facts and specifics, and the application of the disciplinary concepts of contextualization and causation. This item measures whether students can integrate all of this outside knowledge in an argumentative essay, thereby also measuring key disciplinary writing skills such as making claims, identifying and using evidence (i.e., information, details) to support claims, and describing historical conditions clearly and accurately.

Contextualizing history is about working to understand historical phenomena (e.g., events, people, sources) as they existed in their original worlds in order to understand them on their own terms and not through a modern lens. It is so essential to understanding history that one scholar noted, "For the historian, context is all."

A major strength of this item is that it demands that students integrate and demonstrate historical knowledge and skill in their response. In addition to demonstrating the historical knowledge described above, students must contextualize each reform organization, a skill that is central to understanding the past. If contextualization were not required by the item, a student could represent the genesis of these organizations ahistorically or as time neutral. Contextualizing history is about working to understand historical phenomena (e.g., events, people, sources) as they existed in their original worlds in order to understand them on their own terms and not through a modern lens. It is so essential to understanding history that one scholar noted, "For the historian, context is all" (Berlin, 2004, p. 1263).

Additional strengths of the item include the opportunity for students to choose the organizations they analyze. Writing a successful response to the item requires substantial and complex historical knowledge and skill, so student choice helps the item measure what students know rather than what they don't know. This opportunity for student choice combined with the construction of the prompt also helps balance the complexity of the item. The item prompt provides an overarching argument for the essay so students do not generate that on their own. ("The reform efforts of these organizations have met with varying degrees of success.") Another strength is the prompt's use of bullet points to describe the two mandatory topics to discuss for each selected reform organization and specific requirements for the essay. The use of bullets likely helps students more easily understand what is required of them than a non-bulleted prompt would.

This item is scored holistically on a 0-5 scale, with four key score criteria described at each level. The first and fourth criteria focus on the thoroughness and completeness of the answer and its logical organization, while the second and third criteria focus on employing disciplinary knowledge and skills. We recommend weighting the history specific knowledge, thinking, and writing skills more heavily than this rubric currently does. And while the holistic scoring approach makes some sense for a large-scale standardized test, given the variability with which students master specific skills and knowledge, a more analytic rubric may be useful for these essays. This would more accurately capture students' competence (and struggle) with specific aspects of their response.

Item Example 5*

*a thematic block of 17 items,

To view this item go to http://nces.ed.gov/nationsreportcard/nqt/ and Select "Search NAEP Questions." Select "U.S. History" and grade 12. Then select year 2010. See the 17 questions that make up this thematic block.

Item Example 5 – Item Profile	
Source: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2010 U.S. History. [Items 2010-12H11 #1-17] Accessed on February 18, 2016 from NAEP Questions Tool http://nces.ed.gov/nationsreportcard/itmrlsx/	
Item Type	Thematic block (selected-response and short and extended constructed-response items that are based on the same theme)
Grade Level	High School, Grade 12
Cognitive Complexity Score	13

This "item" is actually a block of seventeen separate items about the U.S. and WWII that measure both historical knowledge and skill, sometimes in tandem, sometimes independently. The thematic block includes nine selected-response items and eight constructed-response items.[4]

[4] One of those constructed-response items is analyzed in prior sections of this paper (Item Example 3).

These items are grouped into three related parts: the U.S. Entry into WWII; the Impact of the War on the U.S. Economy; and the Home Front. In each of these parts, the items generally become more challenging and complex as the sequence of items progresses. For example, the Part 1 sequence includes, in order, a constructed response item that requires students to read an excerpt written by Charles Lindberg to explain his perspective on war; a multiple-choice question that requires students to use their knowledge of the swastika in a propaganda poster to identify one foe as Germany; and a constructed response item that demands that they identify the point of view in both these sources and describe the differences between them. In sequences like this, initial items ask students to read, understand and analyze one primary source before cross-checking or synthesizing it with another. The final constructed response item in the thematic block asks students to make a claim about the ways wars impact the home front, and to use evidence from more than one of the 14 primary sources and five graphs used in the prior 16 questions to support that claim.

Most standardized tests consist of items that jump around in time period, topic, and skill so the student must continually pull up different frames of analysis and context to respond to each item. The thematic block approach eliminates this need, allowing for assessment of deeper disciplinary competencies.

Key historical skills that are measured by this thematic block include: identifying the message and purpose of primary sources (in this case, propaganda posters); understanding graphs and identifying relevant information they do not include; corroborating and synthesizing historical sources; evaluating a claim-evidence connection; and making claims and selecting and using evidence to support those claims.

This type of thematic block allows students to access knowledge about a particular historical topic and time and then stay within that topic for the duration of the block. This is more representative of how history is studied and taught, where a focus, provided by a historical question or topic, frames varied sources and information. And indeed, this block of items includes multiple and varied historical sources that differ in format (i.e., text, posters, graphs) and perspective. The block also allows the measurement of multiple kinds of historical knowledge and skill. Notably, there are items that measure the skill of historical reading, as students must identify the purpose of sources or contextualize a point of view. Most standardized tests consist of items that jump around in time period, topic, and skill so the student must continually pull up different frames of analysis and context to respond to each item. The thematic block approach eliminates this need, allowing for assessment of deeper disciplinary competencies. Indeed, the thematic block was designed to assess students' "ability to work in focused areas" rather than their breadth of knowledge and ability to work "across the domain" (Lazer, 2015, p. 148).

One concern about this thematic block's final question that requires students to pull together sources and information to explain important effects of wars at home, is the logistical difficulty for students to flip back and forth between pages that include more than 15 sources. The solution to this difficulty may be the use of computer-based testing that, with a thoughtful design, could allow students to revisit and reference sources more easily.

Item Example 6

To view this item, go to pages 6-10 of the pdf at https://secure-media.collegeboard.org/digitalServices/pdf/ap/ap15_frq_us_history.pdf

The associated rubric may be found on pages 3-5 at https://secure-media.collegeboard.org/digitalServices/pdf/ap/rubrics-ap-histories-historical-thinking-skills.pdf

Item Example 6 – Item Profile	
Source: The College Board, AP® United States History 2015 Free-Response Questions, pp 6-10. Accessed on February 1, 2016 from https://secure-media.collegeboard.org/digitalServices/pdf/ap/ap15_frq_us_history.pdf	
Item Type	Performance Task (Document-Based Question)
Grade Level	High School
Cognitive Complexity Score	16

This Document-Based Question (DBQ) measures both disciplinary skills and knowledge in an integrated way, using a format that is a facsimile of a disciplinary task. Students must analyze and synthesize six primary sources (i.e., excerpts from books, a letter, and a political platform) to write an argument that explains a historical trend—the rise of a new conservatism between 1960 and 1989. Within a persuasive argument, students must state a thesis, use supporting evidence from available sources and relevant outside examples, and contextualize sources or events. This item is a challenging one because students must demonstrate historical reading, writing, and thinking skills while employing their historical knowledge—all in a 55-minute task that is embedded in a three hour and fifteen minute test.

To demonstrate their reading abilities, students need to understand the prompt and comprehend, analyze, and use the sources appropriately in their argument. This requires that they read the sources historically and consider the origins of each source and its historical context. For example, students could characterize Milton Friedman's words in Capitalism and Freedom as a response to New Deal policies or the excerpt from a citizen's letter to Governor Rockefeller as an example of constituent pressure.

Students demonstrate their historical writing skills by writing a historical argument that includes a thesis and supporting evidence from documentary analysis. This requires key skills like introducing and integrating evidence into a paragraph as well as accurately explaining how that evidence supports a claim. While both the reading and writing demands connect to students' historical knowledge, this DBQ also explicitly demands that students further show their knowledge by using examples from outside the documents as evidence, and by including a synthesis that extends the argument or accounts for contradictory evidence.

The item's requirement that students integrate multiple facets of historical understanding to complete it is an important strength. Additionally, the item includes other key desirable design features. Students work with a varied set of primary source excerpts that have been carefully prepared for student access. Each excerpt is short and focused and the relevant origins of each are clearly and easily identified. This set of sources allows students who may struggle with historical background and knowledge on this topic to still construct an answer to the prompt, although, minus outside examples and knowledge, they would not receive full credit for their response. Other important design features of the item are that the prompt allows for multiple legitimate answers and the item reflects the core disciplinary understanding that explaining the past routinely requires making an evidence-based argument about the past.

...the prompt allows for multiple legitimate answers and the item reflects the core disciplinary understanding that explaining the past routinely requires making an evidence-based argument about the past.

The item's scoring rubric reflects the item's focus on domain-specific skills and knowledge and also has additional strengths. For example, the analytic rubric is strong in several generic ways, including the limited number of domains (4), and an easy to use 0-1 point scale for three of those domains. These domains are also designed so writing a logical, source-based, argument is not a sufficient response—students must show evidence of historical thinking to receive full credit (Monte-Sano, 2010). The specific historical skill of "contextualization" is one domain on the rubric. The defining of contextualization so as to be assessable on a 0-1 point scale and the fact that it is always assessed on these DBQs reflects its key and privileged place in the study of history. Overall, this item (the prompt, task materials, and scoring system) requires a high degree of cognitive complexity and prioritizes domain-specific skills and knowledge.

This complexity is appropriate for an item that assesses students' competence after taking an AP course that is intended to offer high school students a college-level experience with history. Its complexity may need to be moderated for assessing students' historical understanding in other high school courses. For example, the New York State Regents test also includes a document-based question, but students answer questions about individual or paired documents for a score before responding to the entire prompt (which is segmented into multiple parts). Additionally, the requirement that students use outside knowledge to provide additional examples and contextualize events or phenomena also assumes that students have studied particular historical periods and themes. To include similar demands on a DBQ for a large-scale state test would require a careful consideration of what, if any, curriculum and content students have necessarily studied.

Appendix A: Evaluating Cognitive Complexity in History Items

This tool is designed to provide a quantitative rating of the cognitive complexity of a history assessment item. The higher the number of points, the higher the cognitive complexity and cognitive demand of the item.

Increasing Cognitive Complexity

Aspect 1: DESIGN FEATURES

1. Item Format – Select One

1 point
Multiple choice

2 points
Constructed-response
OR
Thematic block

3 points
Extended constructed-response

Points

____

2. Number and Types of Historical Source(s) – Select One

1 point
Single historical source

2 points
Multiple historical sources

3 points
Historical sources of varied formats/ genres

Points

____

3. Integration of Historical Knowledge and Skills – Select One

0 point
No

1 point
Yes

Points

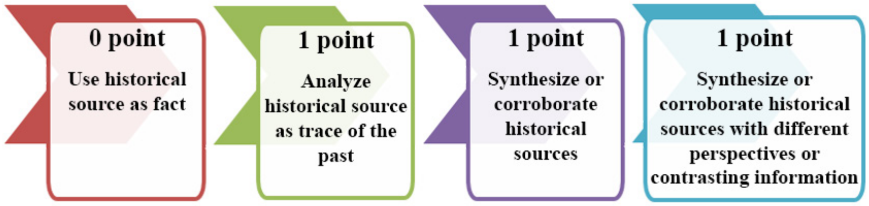
____

This tool is designed to provide a quantitative rating of the cognitive complexity of a history assessment item. The higher the number of points, the higher the cognitive complexity and cognitive demand of the item.

Increasing Cognitive Complexity

Aspect 2: DISCIPLINARY DEMANDS

4. Demand: Kind of Historical Reading/Analysis Required - Select All That Apply



0 point
Use historical source as fact

1 point
Analyze historical source as trace of the past

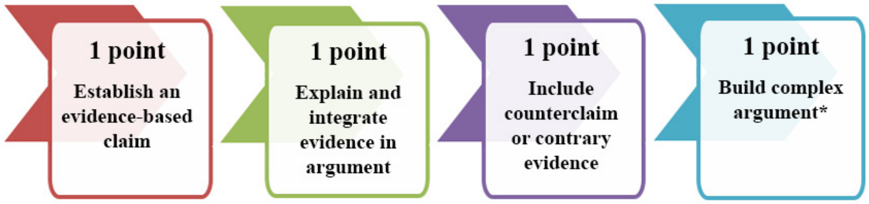
1 point
Synthesize or corroborate historical sources

1 point
Synthesize or corroborate historical sources with different perspectives or contrasting information

Points

____

5. Demand: Kind of Historical Writing (Argumentation) Required - Select All That Apply



1 point
Establish an evidence-based claim

1 point
Explain and integrate evidence in argument

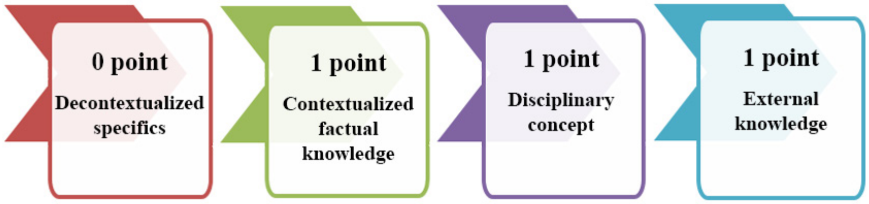
1 point
Include counterclaim or contrary evidence

1 point
Build complex argument*

Points

____

6. Demand: Kind of Historical Knowledge Assessed - Select All That Apply



0 point
Decontextualized specifics

1 point
Contextualized factual knowledge

1 point
Disciplinary concept

1 point
External knowledge

Points

____

Total Points

____

APPENDIX B

Table 1.
Scores of Selected Items on *“Evaluating Cognitive Complexity in History Items” Tool*

		Dimensions	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6
Aspect 1:	Design Features	1 Item format	1	1	2	3	2	3
		2 No. and type of sources	0	1	1	0	3	3
		3 Integration of knowledge & skill	0	0	1	1	1	1
Aspect 2:	Disciplinary Demands	4 Historical reading demands	0	1	0	0	3	3
		5 Historical writing demands	0	0	2	3	2	3
		6 Historical knowledge demands	1	0	1	3	2	3
		Total	**2**	**3**	**7**	**10**	**13**	**16**

References

American Historical Association (1997). *Criteria for standards in history/social studies/social sciences.* http://www.historians.org/teaching-and-learning/classroom-content/resources-on-k-16-teaching/criteria-for-standards-in-historysocial-studiessocial-sciences

Bain, R. B. (2005). *"They thought the world was flat?": Applying the principles of how people learn in teaching high school history.* Retrieved February 1, 2016 from http://www.nap.edu/read/10126/chapter/5

Berlin, I. (2004). American slavery in history and memory and the search for social justice. The *Journal of American History, 90*, 1251-1268. doi: 10.2307/3660347

College Board. (2015). *AP United States History: Including the curriculum framework* updated Fall 2015. New York, NY: Author. https://secure-media.collegeboard.org/digitalServices/pdf/ap/ap-us-history-course-and-exam-description.pdf

Ercikan, K., & Seixas, P. (Eds.). (2015). *New directions in assessing historical thinking.* New York, NY: Routledge.

Ercikan, K., Seixas, P., Lyons-Thomas, J., and Gibson, L. (2015). Cognitive validity evidence for validating assessments of historical thinking. In K. Ercikan & P. Seixas (Eds.), *New directions in assessing historical thinking* (pp. 206-220). New York, NY: Routledge.

Holt, T. (1990). *Thinking historically: Narrative, imagination, and understanding.* New York, NY: College Entrance Examination Board.

Hynd, C., Holschuh, J. P., & Hubbard, B. P. (2004). Thinking like a historian: College students' reading of multiple historical documents. *Journal of Literacy Research, 36*(2), 141–176. doi:10.1207/s15548430jlr3602_2

Immerwahr, D. (2008, February 1). *The fact/narrative distinction and student examinations in history.* Retrieved February 24, 2016, from History Teacher, http://eric.ed.gov/?id=EJ791705

Lazer, S. (2015). A large-scale assessment of historical knowledge and reasoning: NAEP U.S. History Assessment. In K. Ercikan & P. Seixas (Eds.), *New directions in assessing historical thinking* (pp. 145-158). New York, NY: Routledge.

Levstik, L., & Barton, K. (2005). *Doing history: Investigating with children in elementary and middle schools.* Mahwah, New Jersey: Lawrence Erlbaum Associates.

Martin, D., Maldonado, S., Schneider, J., & Smith, M. (2011). *A report on the state of history education: State policies and national programs.* National History Education Clearinghouse. Retrieved from http://teachinghistory.org/system/files/teachinghistory_special_report_2011.pdf

Mink, L. (1987). *Historical understanding*. Ithaca, NY: Cornell University Press.

Moje, E. B., Ciechanowski, K. M., Kramer, K., Ellis, L., Carrillo, R., & Collazo, T. (2004). Working toward third space in content area literacy: An examination of everyday funds of knowledge and discourse. *Reading Research Quarterly, 39*(1), 38–70. doi:10.1598/rrq.39.1.4

Monte-Sano, C. (2010). Disciplinary literacy in history: An exploration of the historical nature of adolescents' writing. *The Journal of the Learning Sciences, 19*(4), 539–568. doi:10.1080/10508406.2010.481014

Monte-Sano, C. (2011). Beyond reading comprehension and summary: Learning to read and write in history by focusing on evidence, perspective, and interpretation. *Curriculum Inquiry, 41*(2), 212–249. doi:10.1111/j.1467-873x.2011.00547.x

National Council for the Social Studies (NCSS) (2013). *The College, Career, and Civic Life (C3) Framework for Social Studies State Standards: Guidance for Enhancing the Rigor of K-12 Civics, Economics, Geography, and History* (Silver Spring, MD: NCSS). http://www.socialstudies.org/c3

National Governors Association Center for Best Practices, Council of Chief State School Officers (2010). *Common Core State Standards for Literacy in History/Social Studies, Science, and Technical Subjects.* National Governors Association Center for Best Practices, Council of Chief State School Officers, Washington D.C.

National Research Council. (2005). *How students learn: History in the classroom.* Committee on How People Learn, A Targeted Report for Teachers, M.S. Donovan and J.D. Bransford, Editors. Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

Pope, D. (2003, June). "Making sense of advertisements," *History matters: The U.S. survey course on the web*, http://historymatters.gmu.edu/mse/Ads/. Accessed February 1, 2016.

Reich, G. A. (2009). Testing historical knowledge: Standards, multiple-choice questions and student reasoning. *Theory & Research in Social Education, 37*(3), 325–360. doi:10.1080/00933104.2009.10473401

Reisman, A. (2015) The difficulty of assessing disciplinary historical reading. In K. Ercikan & P. Seixas (Eds.), *New directions in assessing historical thinking* (pp. 29-39). New York, NY: Routledge.

Roy Rosenzweig Center for History and New Media, (n.d.). *Using primary sources.* http://teachinghistory.org/best-practices/using-primary-sources. Accessed February 2, 2016.

Schneider, J., & Zakai, S. (2016). A rigorous dialectic: Writing and thinking in history. *Teachers College Record, 118*(1), 1-36.

Seixas, P., & Ercikan, K., (2015). Introduction: The new shape of history assessments. In K. Ercikan & P. Seixas (Eds.), *New directions in assessing historical thinking* (pp. 1-13). New York, NY: Routledge.

Seixas, P., Gibson L., & Ercikan, K. (2015). A design process for assessing historical thinking: The case of a one-hour test. In K. Ercikan & P. Seixas (Eds.), *New directions in assessing historical thinking* (pp. 102-115). New York, NY: Routledge.

Shanahan, T., & Shanahan, C. (2008). Teaching disciplinary literacy to adolescents: Rethinking content-area literacy. *Harvard Educational Review, 78*(1), 40–59. doi:10.17763/haer.78.1.v62444321p602101

Stearns, P. N., Seixas, P., & Wineburg, S. (Eds.) (2000). *Knowing, teaching, and learning history: National and international perspectives.* New York, NY: New York University Press.

U.S. Department of Education (2009). *Race to the Top Program Executive Summary.* Washington, DC.

VanSledright, B. (2002). *In search of America's past: Learning to read history in elementary school.* New York, NY: Teachers College Press.

Wineburg, S. (1991). On the reading of historical texts: Notes on the breach between school and academy. *American Educational Research Journal, 28* (3) 495-519. doi: 10.3102/00028312028003495

Wineburg, S. (2001). *Historical thinking and other unnatural acts: Charting the future of teaching the past.* Philadelphia, PA: Temple University Press.

Wineburg, S. (2004). Crazy for history. *The Journal of American History, 90*(4), 1401-1414. doi:10.2307/3660360

Wineburg, S., & Martin, D. (2004). Reading and rewriting history. *Educational Leadership 62*(1), 42-45.

Wineburg, S., Martin, D., & Monte-Sano, C. (2013). *Reading like a historian: Teaching literacy in middle and high school history classrooms*. New York, NY: Teachers College Press.