Educ Psychol Rev (2014) 26:543–560
DOI 10.1007/s10648-014-9280-4

REVIEW ARTICLE

Assessment of Critical-Analytic Thinking

Nathaniel J. S. Brown · Peter P. Afflerbach · Robert G. Croninger

Published online: 18 October 2014

Abstract National policy and standards documents, including the National Assessment of Educational Progress frameworks, the *Common Core State Standards*, and the *Next Generation Science Standards*, assert the need to assess critical-analytic thinking (CAT) across subject areas. However, assessment of CAT poses several challenges for developers of both large-scale and classroom assessments: Current CAT assessments often suffer from questionable item contexts, subjective rubrics, and underdeveloped construct formulations. Attention to these aspects of assessment would improve understanding of the development of students' CAT and provide tools for helping teachers teach and students learn. We discuss these challenges within the context of several content areas and highlight the importance of developing formative assessments that capture the development of CAT in different domains of learning.

Keywords Critical-analytic thinking · Formative assessment · Large-scale assessment · Classroom assessment · National Assessment of Educational Progress · Common Core State Standards · Next Generation Science Standards

Promoting students' deeper engagement with content and critical thinking has long been a component of education reform (Kennedy et al. 1991; Knapp, M. and Associates 1995; Willingham 2007). However, policy and standards documents currently promoted by the federal government and state departments of education, including the National Assessment of Educational Progress (NAEP) frameworks (National Assessment Governing Board 2010), the *Common Core State Standards* (Common Core State Standards Initiative 2010), and the *Next Generation Science Standards* (NGSS Lead States 2013), place a renewed emphasis on what can be referred to as critical-analytic thinking (CAT), especially the capacity to evaluate multiple streams of information in different representational formats in fundamental content areas, such as English language arts, mathematics, and science. According to

N. J. S. Brown (✉)
Educational Research, Measurement, and Evaluation, Lynch School of Education, Boston College, 140 Commonwealth Avenue, Chestnut Hill, MA 02467, USA
e-mail: nathaniel.js.brown@bc.edu

P. P. Afflerbach · R. G. Croninger
Teaching and Learning, Policy and Leadership, College of Education, University of Maryland, College Park, MD, USA

reform advocates, these cognitive skills are essential for students' success in college and in the workplace (Common Core State Standards Initiative 2010). The challenges posed by such an ambitious reform are substantial, particularly in constructing developmentally appropriate assessments that help guide instruction and serve as benchmarks by which to gauge the effects of reform.

In this article, we discuss the research on CAT and why it is thought to be an important instructional goal in schools. Next, we examine examples of CAT in the subject areas as they appear in influential policy and standards documents, specifically the NAEP framework for reading, the *Common Core State Standards* for English language arts, and the *Next Generation Science Standards*. While we focus on these frameworks for the purpose of illustration, we believe that they represent the general expectation that students and teachers must attend to CAT, as espoused in similar frameworks across other subject areas (see, for example, Cobb and Jackson 2011, for a discussion of the *Common Core State Standards* for mathematics). Next, we note that the valid, reliable, fair, and useful assessment of CAT poses challenges for the developers of both large-scale and classroom assessments and that existing CAT assessments aligned with the above frameworks struggle to meet these challenges. Finally, we describe the need for assessments with formative value that help teachers make instructional decisions and that provide students with meaningful feedback.

CAT in Standards Reform

While educators have long promoted developing students' capacity to think critically and analytically (Dewey 1933; Bloom 1956), there is less agreement about what it means to do so (Kennedy et al. 1991; Willingham 2007). Our own conceptualization of CAT is that it is both a disposition and a cognitive process (Facione 2000; Halpern 1998), each of which may be influenced by interactions with other individuals and specific content material (Halpern 2001; Knapp, M. and Associates 1995; Murphy et al. 2009; Wentzel 2009). Critical-analytic thinking requires that a student accept some level of uncertainty about a task and be open to multiple solutions; it implies a certain level of inquisitiveness and motivation to examine complex content deeply, well beyond simply recalling facts or restating answers (Guthrie et al. 2013; Schraw and Robinson 2012).

This process of CAT is how one links and evaluates information to weigh the evidence for alternative solutions. Although it can be represented as a series of analytic steps or global strategies (Billing 2007; Halpern 1998), CAT is also highly contextualized, combining specific content knowledge, normative standards, and analytic skills to assess the extent to which information supports or fails to support a particular proposition (Bailin 2002; Willingham 2007). The extent to which analytic strategies or processes can be transferred from one content area or domain is contested (Bailin 2002; Halpern 1998; Willingham 2007). Generally speaking, the more distant or dissimilar the domain, the more difficult it is to transfer capacity to engage in CAT (Halpern 1998). As a result, it is quite possible for a student to demonstrate the capacity to engage in CAT in one content area, say reading, but fail to be able to do so in another, say science (Bailin 2002).

Helping students develop the disposition and evaluative skills to engage in CAT is complex and requires scaffolding over time, building on students' social and cognitive development, acquisition of relevant knowledge, and experience with prior analytic mistakes and solutions (Schraw and Robinson 2012). Children have the capacity to develop regulatory skills useful in the promotion of CAT (Kennedy et al. 1991; Willingham 2007). Over time, students' social and cognitive development can support greater attention, sustained engagement in tasks, or more nuanced judgments about problems and tasks (see Byrnes and Dunbar 2014). However, many students do not receive explicit instruction in CAT or are not provided tasks suitable to developing either the disposition or analytic skills that it requires, especially in schools that serve students from historically disadvantaged backgrounds (Farah 2010; Knapp, M. and Associates 1995).

Critical-analytic thinking has broad value, which is one reason why it has been a persistent component of curriculum reforms (Kennedy et al. 1991; Knapp, M. and Associates 1995), including current standards-based reform (Porter et al. 2011). It is a purposeful, reflective approach to living and learning, thought by many educators to be essential to civic participation and the development of professional expertise (Noddings 2006). However, these broader benefits of CAT have not been realized by prior reforms that have sought to promote and strengthen students' capacity to reason, analyze, and critique (Arum and Roksa 2011; Schraw and Robinson 2012; Willingham 2007). One of the motivations for the development of the *Common Core State Standards* is to promote CAT as a way to prepare students to be more successful in college and in the workforce, the underlying assumption being that students are not prepared currently to do so upon graduation from high school (Common Core State Standards Initiative 2010; Porter et al. 2011). Likewise, one of the motivations for the development of the *Next Generation Science Standards* is to better prepare students for the actual work of scientists and engineers, responding to the belief that current science instruction fails to meet this challenge (NGSS Lead States 2013).

To realize these broader values, we need to develop a stronger understanding of how to assess students' capacity to engage in this form of thinking and how to scaffold their ability to do so from instructional tasks to the authentic and increasingly complex challenges that students will face in life (Willingham 2007). Although studies suggest that most students, including students from historically disadvantaged backgrounds (e.g., English language learners, low-income students, African-American students, and students with learning disabilities), can develop CAT (Knapp, M. and Associates 1995), these students and their teachers face the challenge of doing so with far fewer resources (Tienken and Zhao 2013). Without developing policies and formative tools to help students achieve higher standards, especially for students in schools that serve historically disadvantaged students, policies that promote CAT are likely to only widen the achievement gap between historically advantaged and disadvantaged students (Tienken and Zhao 2013).

Before we discuss the challenges inherent in the assessment of CAT and the need to develop CAT assessments with formative value, we highlight the high standards expected of students in the area of critical-analytic thinking as expressed in three current and influential policy and standards documents: the NAEP framework for reading, the *Common Core State Standards* for English language arts, and the *Next Generation Science Standards*.

CAT in the NAEP Framework for Reading

The NAEP Reading Assessment is guided by a framework and definition of reading. The current NAEP Framework (National Assessment Governing Board 2010) derives from expert consensus and defines reading in the following manner:

Reading is an active and complex process that involves

- Understanding written text,
- Developing and interpreting meaning, and
- Using meaning as appropriate to type of text, purpose, and situation.

The NAEP Reading Assessment develops and uses test items based on three cognitive targets, or the type of thinking that is required of students who answer items correctly. The cognitive targets are as follows:

- Locate and Recall: When locating or recalling information from what they have read, students may identify explicitly stated main ideas or may focus on specific elements of a story.
- Integrate and Interpret: When integrating and interpreting what they have read, students may make comparisons, explain character motivation, or examine relations of ideas across the text.
- Critique and Evaluate: When critiquing or evaluating what they have read, students view the text critically by examining it from numerous perspectives or may evaluate overall text quality or the effectiveness of particular aspects of the text. (National Assessment Governing Board 2010)

The cognitive targets of Integrate and Interpret, and Critique and Evaluate, are components of CAT. However, the full NAEP Reading definition reflects a major shift in the conceptualization of reading, a shift that has considerable implications for assessment. According to the definition, reading is now conceptualized to include *using the meaning that is constructed through reading* and not "just" the construction of meaning. No longer is the construction of meaning—the comprehension of text—an end point. Rather, it is a midpoint, with the acknowledgment that we read both to comprehend *and* to use that which we comprehend. Following from this characterization of reading is the expectation that students will apply the information that they learn from text, as when they analyze and critique how an author uses language to create mood in a short story, synthesize their understandings across three different articles on the causes of the Civil War, or use multimedia presentations to describe their understanding of science texts.

Each of the given examples assumes that CAT is an integral part of reading. Related, such reading situated in the content areas creates the need for assessment that effectively samples reader behavior. An immediate need is for reading assessment to inform us of the nature of students' development across a spectrum of reading strategies, skills, and mind-sets that include CAT. That we read to understand *and* then use that understanding has not been a common focus of our assessments. We give attention to particular aspects of reading, including phonemic awareness, phonics, and fluency, as children develop the ability to read. We assess students' literal and inferential comprehension as they increasingly learn from content area texts. However, many assessments do not venture into the area of how students use what they learn from reading or how they think critically and analytically in relation to the texts that they read.

NAEP Reading results describe a persistent achievement gap in the USA between historically disadvantaged students (e.g., English language learners, low-income students, African-American students, and students with learning disabilities) and historically advantaged students (primarily White, middle and upper income students). Students who perform below "basic" on NAEP proficiency levels face a serious challenge in attaining higher proficiency levels that require CAT. It is especially worrisome that new literacy demands and assessments (including those related to the *Common Core State Standards*) demand a "basic" reading achievement level as an entry point to more complex and demanding performance assessments. In effect, more complex assessments and higher standards may widen the achievement gap (Tienken and Zhao 2013). Classroom instruction that narrows this gap must focus on both what is traditionally considered "basic" for reading competency (including mechanics of reading like phonics) and critical-analytic thinking (Knapp, M. and Associates 1995).

CAT in the Common Core State Standards for English Language Arts

The *Common Core State Standards* (CCSS; Common Core State Standards Initiative 2010) reflect a consistent focus on CAT and include the following English language arts standard for informational text/integration of knowledge and ideas:

Explain how specific images (e.g., a diagram showing how a machine works) contribute to and clarify a text.

Our task analysis, which is a standard component of any successful assessment development process, indicates that students who successfully meet the stated standard will do the following:

- Apply all necessary reading strategies and skills including phonics, fluency, vocabulary, and comprehension.
- Access and use appropriate prior knowledge.
- Construct meaning from the text.
- Comprehend a related image.
- Compare the two (text and image) related understandings.
- Analyze the text and image for their separate and joint contributions to understanding.
- Explain (through writing or speaking) how the two comprehended parts relate to one another.
- Describe how the image helps comprehension.

In addition, we believe that the student who is successful vis-à-vis these competencies will be metacognitive throughout the performance assessment. We note that this is a second grade standard.

This standard will be measured through performance assessments developed by the Smarter Balanced Assessment Consortium (SBAC) and the Partnership for Assessment of Readiness for College and Careers (PARCC). These performance assessments will be summative, gauging a student's ability to successfully meet the complex standard. However, students' progress toward meeting complex standards will depend on an appropriate and extended sequence of instructional activities supported by formative assessments that can chart student progress along this path (what second grade student is currently capable of meeting the above standard?). The student must be accomplished not only in the individual aspects of the performance but also in the coordination and orchestration of the different components. Students should not be expected to meet the standards without targeted, ongoing instruction that is guided by formative assessment.

We note that this second grade standard assumes that a student has already effectively constructed meaning for the text at hand. Based on this assumption, and on a student's actual realization of meaning making, entering the realm of CAT is possible. An important point here is that for many of the students who will be taking the performance assessments affiliated with the CCSS, constructing meaning or comprehending text serves as the midpoint suggested in the most recent NAEP framework for reading. Yet, as noted, many students do not reach the "basic" level of achievement on the NAEP Reading Assessment. These students, many from historically disadvantaged households, typically struggle with the establishment of a literal understanding of text, and establishing an inferential understanding is often difficult for these students. These students, unable to establish the literal comprehension baseline that is assumed of each and every CCSS standard in English language arts, face daunting challenges in both reaching school goals and in receiving equitable opportunities to do so; without a greater focus on building the capacity of these students' teachers and schools to reach higher standards of student performance, a call for promoting students' CAT is only likely to lead to greater educational and social inequalities (Knapp, M. and Associates 1995; Tienken and Zhao 2013).

CAT in the Next Generation Science Standards

The *Next Generation Science Standards* (NGSS Lead States 2013) represent a major shift in expectations of how science and engineering should be taught and learned. The NGSS recognize that CAT is a central aspect of scientific practice, as illustrated in Fig. 1, which comes from the National Research Council (2012) *Framework for Science Education*, the document on which the NGSS are based. The results from scientific investigations, and the models and theories that scientists propose to explain those results, must be constantly evaluated. Arguing, critiquing, and analyzing—all hallmarks of CAT (see Shared Claims in Alexander 2014)—are central to the job of a scientist, and the NGSS likewise position them as central in science education.

In a dramatic shift from previous science standards documents, practices such as CAT are not treated as independent or isolated topics, in which "inquiry" standards (and assessment items) are separate from "conceptual" standards (and items). Instead, the NGSS are written as performance expectations (PEs), which are meant to be "assessable statements of what students should know and be able to do" (NGSS Lead States 2013) and which incorporate three components: one or more disciplinary core ideas, a cross-cutting concept, and a scientific and engineering practice. CAT represents a third of the practices in the NGSS, and fully a third of the standards include an expectation of CAT.

Notably, these performance expectations are modeled on authentic activities engaged in by scientists, as well as socially important activities engaged in by an educated, scientifically literate populace. An example of a high school performance expectation is as follows:

> Evaluate the evidence supporting claims that changes in environmental conditions may result in: (1) increases in the number of individuals of some species, (2) the emergence of new species over time, and (3) the extinction of other species (HS-LS4-5; NGSS Lead States 2013).

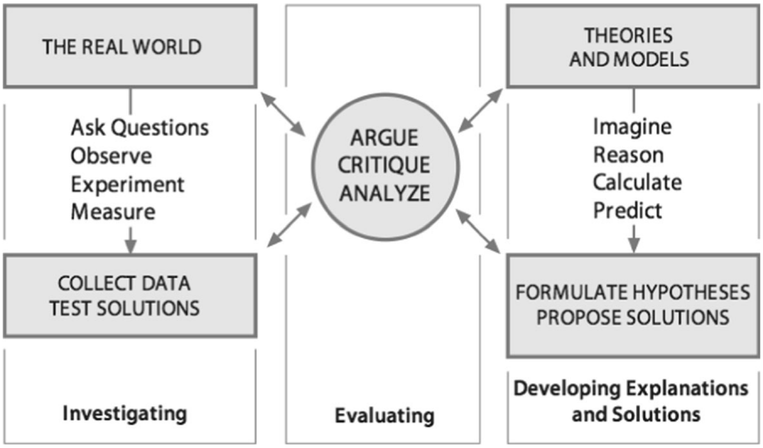

Fig. 1 Aspects of scientific and engineering practice, illustrating the centrality of CAT skills—arguing, critiquing, and analyzing—in science. Source: National Research Council (2012). A Framework for K-12 Science Education: Practices, Crosscutting Concepts, and Core Ideas. Washington, DC: National Academy of Sciences

This standard expects students to engage in CAT with respect to claims that changes in the environment, such as global climate change, will impact plant and animal life. This is both an active and important area of study in ecological and climate science and a significant requirement of everyday life as people face a steady barrage of information and misinformation about climate change. Throughout their lifetime, students will need to be in a position to evaluate scientific and pseudoscientific claims that they encounter in the media, deciding based on the evidence presented whether skepticism is warranted. And, for those students who go on to become scientists, they will need to be able to evaluate new scientific claims made by colleagues based on new data and evidence. The NGSS assert that students at all grade levels should be expected to engage in CAT.

The NAEP frameworks, the *Common Core State Standards*, and the *Next Generation Science Standards* share an expectation that students will be able to critically analyze complex material and craft arguments supported by evidence, in authentic contexts, in preparation for a life as informed citizens, workers, and practitioners. This represents enormous opportunity and challenge for teachers and students, as it does for assessment developers. As discussed in the following section, we view these CAT frameworks as opportunities largely yet to be realized due, in part, to the challenges involved in developing assessments capable of measuring a fuller array of critical-analytic thinking.

Challenges in the Assessment of CAT

Any educational assessment is a complex system with several parts, requiring a principled design approach to ensure a coherent, functioning system that produces valid and fair inferences, reliable scores, and useful information to stakeholders. When the target construct is as complex as CAT, a principled approach to assessment design is even more critical.

Principled assessment design refers to the development of an assessment with an explicit argument for how evidence will be gathered and interpreted that bears on the underlying knowledge, skills, and processes—for example, CAT—that the assessment is intended to address (National Research Council 2001). Such an argument should include (a) a model of cognition—a formulation of the construct to be assessed, (b) tasks that elicit observations of student performance, and (c) methods of interpretation that connect the performance outcomes to the model of cognition (National Research Council 2001). Expanding upon this model, the practice of assessment can be characterized as a cyclical process involving four steps (Brown and Wilson 2011). As illustrated in Fig. 2, any cycle of assessment starts with a *question* about how much a student possesses one or more latent variables, such as proficiency at CAT. To answer this question, four steps are necessary: (a) *observing*, eliciting performances assumed to depend upon the latent variable(s), leading to a set of *observations*, (b) *scoring*, categorizing different observed performances and assigning them a relative value, or *scores*, (c) *summarizing*, combining the values of the individual performances to yield *measures* of each latent variable, and (d) *interpreting*, using the measures of the latent variable(s) to answer the question and characterize how much of the CAT construct a student possesses.

Each of these four steps is mediated by one of the four component parts of an assessment: the items, the rubrics, the measurement model, and the construct formulation. As shown in Fig. 2, the assessment items mediate the process of observing, the rubrics mediate the process of scoring, the measurement model mediates the process of summarizing, and the construct formulation mediates the process of interpreting. A principled approach to assessment design focuses on all four components, bringing them into coherence and establishing the necessary

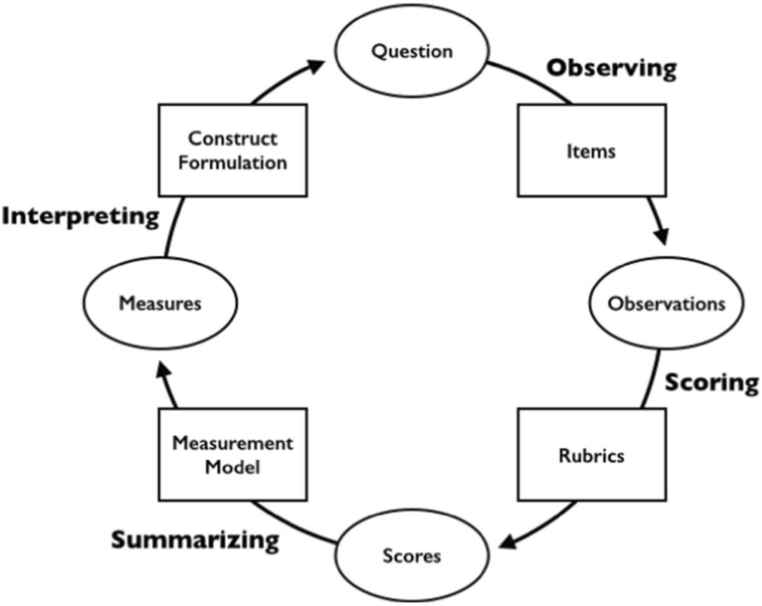

Fig. 2 The four components of an assessment—the items, the rubrics, the measurement model, and the construct formulation—each mediating one of the four steps in the process of assessment—observing, scoring, summarizing, and interpreting

evidentiary argument connecting a model of cognition, tasks, and methods of interpretation (Brown and Wilson 2011; Wilson 2005).

Developers of large-scale educational assessments, including those targeting CAT such as the NAEP Reading Assessment and the assessments aligned with the CCSS developed by PARCC and SBAC, generally excel at developing and implementing the measurement model component of an assessment, relying on the expertise of psychometricians to conduct the technical analyses necessary to turn item response data into student scale scores. The assessment of CAT, however, poses particular challenges for developers of both large-scale and classroom assessments when it comes to the other three assessment components—the items, the rubrics, and the construct formulation—as discussed in the following sections.

CAT Assessment Items

Items create and define the contexts in which students will perform. As discussed, there is evidence that CAT skills are contextual. It does not seem plausible that students will develop their capacity for CAT regardless of the problems to which they are exposed or that the learned skills will be transferrable to a wide range of settings. Instead, teaching and assessing these skills requires the selection of meaningful, relevant contexts in which students can apply the form of CAT that is authentic to that domain (Bailin 2002; Willingham 2007). For example, "thinking like a historian" involves constructing accounts and understanding historic events, characters, and places. Students who are learning to think like historians may read several texts with conflicting accounts of an historic event, and related CAT tasks include identifying the sources of information, determining the relative trustworthiness and reliability of the sources of historical information, and creating an account of history through the synthesis of varied source materials (De La Paz et al. 2014).

In the domain of history, complex assessment contexts and authentic CAT tasks have been long-standing characteristics of the document-based question (DBQ) component of the Advanced Placement United States History exam. DBQs present the student with several primary source documents and pose legitimate historical questions that require students to think like a historian (VanSledright 2013). Assessments aligned with the CCSS have shown some movement in this direction. PARCC, for example, has expressed a commitment to *texts worth reading*, "authentic texts worthy of study instead of artificially produced or commissioned passages," and *questions worth answering*, "sequences of questions that draw students into deeper encounters with texts… rather than sets of random questions of varying quality" (PARCC, 2013).

Along these lines, PARCC is developing research simulation tasks (RSTs) focused on the CCSS for literacy in history/social studies that bear some resemblance to DBQs, in which students read two or three primary or secondary sources and, following a sequence of items targeting reading comprehension, write a prose constructed response (PCR) essay analyzing the sources as a capstone item. To date, PARCC has released some sample RSTs that, on the one hand, do appear to incorporate "texts worth reading": primary and secondary sources of historical import. However, although these meaningful contexts could, in theory, support authentic historical CAT—"questions worth answering"—we note that the actual PCR capstone items vary in the extent to which they rise to this challenge.

For example, the sources included in the example grade 7 RST focus on the historical question of the fate of Amelia Earhart's final flight, which is of legitimate debate. However, the PCR prompt focuses instead on Amelia Earhart's bravery, which hardly seems like a matter of serious historical inquiry:

> You have read two texts and watched a video describing Amelia Earhart. All three include information that supports the claim that Earhart was a daring, courageous person. The three texts are: "The Biography of Amelia Earhart," "Earhart's Final Resting Place Believed Found," and "Amelia Earhart's Life and Disappearance" (video). Consider the argument each author uses to demonstrate Earhart's bravery. Write an essay that analyzes the strength of the arguments related to Earhart's bravery in at least two of the texts. Remember to use textual evidence to support your ideas (PARCC, 2013).

As a contrasting example, the sources included in the example grade 11 RST are letters between Abigail and John Adams reflecting on events leading up to the Declaration of Independence, such as the Siege of Boston, and their thoughts on the nature and potential consequences of the impending American independence. Encouragingly, in contrast to the grade 7 RST, this PCR prompt focuses on a matter of legitimate historical inquiry, namely, developing an account of the causes of the American Revolution and its impact on the American people:

> Both John and Abigail Adams believed strongly in freedom and independence. However, their letters suggest that each of them understood these terms differently based on their experiences. Write an essay that explains their contrasting views on the concepts of freedom and independence. In your essay, make a claim about the idea of freedom and independence and how John and Abigail Adams add to that understanding and/or illustrate a misunderstanding of freedom and independence. Support your response with textual evidence and inferences drawn from all three sources (PARCC, 2013).

Importantly, this item presents two sides of an authentic debate as reflected in meaningful, relevant primary sources and asks students to think critically and analytically about the topic and to stake out and support a position. This item supports a form of CAT that is closer to the practices of actual historians.

In the domain of science, although assessments aligned with the NGSS have not yet been developed, the NGSS are designed with meaningful contexts and authentic CAT tasks in mind. The standards are written as PEs, which are meant to be “assessable statements of what students should know and be able to do” (NGSS Lead States 2013). While PEs are not assessment items per se, they have been written with the idea that they can directly guide assessment development and are intended to be clear statements of the performances for which students should be held accountable.

To this end, most PEs are accompanied by clarification statements or assessment boundaries, explicit statements about appropriate and inappropriate assessment item design. For example, the PE just mentioned, concerning the impact of environmental change on plant and animal species, is accompanied by a clarification statement that suggests specific changes in environmental conditions that could be designed into an item:

> Emphasis is on determining cause and effect relationships for how changes to the environment such as deforestation, fishing, application of fertilizers, drought, flood, and the rate of change of the environment affect distribution or disappearance of traits in species (NGSS Lead States 2013).

Together, the clarification statements and assessment boundaries provide information for assessment developers to help craft specific assessment tasks based on the performance expectations.

Because of the complexity and contextuality of the performances to be assessed, CAT tasks in any domain will often need to take the form of performance items: extended constructed response items situated in authentic contexts. Developers of large-scale assessments have traditionally resisted incorporating performance items, presumably because of the difficulty involved in developing, scoring, and administering such items in a large-scale, standardized setting. However, despite these challenges, the developers of more and more large-scale assessments are deciding that the drawbacks of performance items are outweighed by their benefits; both PARCC and SBAC have committed to including them in their assessments aligned with the CCSS, and in science, there has been a recent push to include performance items making use of computer-based simulations on state science assessments (Quellmalz et al. 2012a, 2012b; Quellmalz and Haertel 2004). Such efforts join long-standing large-scale standardized assessments with substantial performance components, including academic assessments like the Advanced Placement exams and professional licensure assessments like the bar exams.

In the classroom, where standardization is not as much of an issue, item contexts can be chosen to match meaningful instructional contexts. However, issues remain with respect to the time required to administer and score such items, and the need for more objective rubrics. Examples of promising performance assessment include those of the Concept-Oriented Reading Instruction program (CORI; Guthrie et al. 2004) that provide measures that include student engagement, searching for information, narrative and expository text comprehension, and application of knowledge gained through reading. As performance assessments become more common and frequently used, we hope that their measurement characteristics and their ability to represent CAT will evolve accordingly in both large-scale and classroom settings.

CAT Assessment Rubrics

Rubrics, or scoring guides, are the assessment component used to assign scores to students’ responses. For CAT assessments, like many assessments that rely on performance items, a common model for rubrics is to take the text of the targeted standard as the description of the highest point value, choose a number of possible intermediate point values, and create descriptions of these lower point values by modifying the text of the standard with a succession of subjective, increasingly derogatory adjectives. As an example, consider the condensed scoring rubric for prose constructed response items that PARCC has proposed for items that assess reading comprehension of key ideas and details (Table 1). This rubric is intended to serve as a template for item-specific scoring guides dealing with the CCSS College and Career Readiness anchor standard number 1 for reading:

> Read closely to determine what the text says explicitly and to make logical inferences from it; cite specific textual evidence when writing or speaking to support conclusions drawn from the text. (Common Core State Standards Initiative 2010)

In the rubric proposed by PARCC, there are five score levels (0 to 4). The descriptions of the score levels are strongly parallel. The primary distinguishing features are subjective adjectives (e.g., "accurate analysis" vs. "mostly accurate analysis" vs. "generally accurate analysis" vs. "minimally accurate analysis" vs. "inaccurate or no analysis"). Rubrics like these for constructed response items are notoriously hard to implement, requiring large investments in time and money dedicated to rater training (Shavelson et al. 1992). This is largely due to the fact that the meanings of these subjective adjectives have to be operationalized through extended discussion and consensus building and taught to new raters using a large number of example responses illustrating the different score levels.

The difficulty in deriving consistent meaning from these rubrics is not just a problem for raters, however, as it is nearly impossible to use these rubrics to give students feedback, help teachers plan instruction, or develop curricular materials. This is because, even when raters can apply the rubric and reliably assign scores, the score levels have no construct-level counterpart. Consider the futility of telling a student or teacher that an essay has a "minimally accurate analysis" and that what the student should be focusing on for improvement is to develop a "generally accurate analysis." What are needed instead are descriptions of score levels that

Table 1 Score level descriptions from the PARCC grades 6–11 condensed scoring rubric for prose constructed response items assessing comprehension of key ideas and details in reading

Score level	Description
4	The student response demonstrates **full comprehension** of ideas stated explicitly and inferentially by providing an **accurate** analysis and supporting the analysis with **effective and convincing** textual evidence.
3	The student response demonstrates **comprehension** of ideas stated explicitly and/or inferentially by providing a **mostly accurate** analysis and supporting the analysis with **adequate** textual evidence.
2	The student response demonstrates **basic comprehension** of ideas stated explicitly and/or inferentially by providing a **generally accurate** analysis and supporting the analysis with **basic** textual evidence.
1	The student response demonstrates **limited comprehension** of ideas stated explicitly and/or inferentially by providing a **minimally accurate** analysis and supporting the analysis with **limited** textual evidence.
0	The student response demonstrates **no comprehension** of ideas by providing inaccurate or no analysis and **little to no** textual evidence.

Emphasis in original

Source: Partnership for Assessment of Readiness for College and Careers (2014). Grades 6–11 Condensed Scoring Rubric for Prose Constructed Response Items. Retrieved from www.parcconline.org/samples/english-language-artsliteracy/grades-6-11-generic-rubrics have inherent meaning, rather than those that rely on subjective, normative comparisons with adjacent score levels.

As an example of a rubric with more objective score levels, consider Table 2. This rubric comes from the San Diego Striving Readers project (Dray et al. 2011) which developed assessments of content area literacy consistent with the secondary school curriculum Strategies of Literacy Independence Across the Curriculum (SLIC; McDonald et al. 2009), which focuses on both basic aspects of reading comprehension, such as identifying the main idea of a text, and more sophisticated aspects of CAT, such as identifying and critiquing authorial intent. The score levels specify the meaningful ways in which students' responses differ, presuming a cognitive model of reading comprehension in which students approach the analysis of a text with a toolbox of tactics that they could apply. Student responses differ in the number of successful tactics used and, if multiple tactics are used, whether the resulting analyses are kept separate or integrated.

Like the PARCC rubric above, this rubric is a general version intended to provide a framework for the development of many different item-specific rubrics. Although the structure of the levels would remain constant, the rubric for each item would identify and draw on its own list of tactics. For example, an item asking the student to identify the main idea of a paragraph might draw on tactics such as:

- Cross-check with text features.
- Look for signal words and phrases (e.g., "and" and "that is why").
- Identify and eliminate supporting sentences (e.g., "for example," quotes, and facts).
- Identify and eliminate supporting details.
- Look for general as opposed to specific language.
- Look for repetition or reiteration of ideas or terms.
- Scan for bolded terms.

while an item asking for a more sophisticated CAT task, such as critiquing authorial intent, might draw on tactics such as the following:

- Identify author's key ideas.
- Identify the tools (e.g., language, evidence, logic, or comparisons) that the author uses to convey information.
- Consider the relations among the form, ideas, and tools used.
- Identify how the author used form, ideas, and tools to convey a particular view.
- Generate hypotheses/evaluate intent.

Although rater training is still a necessity, our experience implementing rubrics like these leads us to believe that raters find them to be quicker to internalize and easier to apply, as the distinctions between levels are more memorable and tied to specific, operationalized characteristics of student responses. Examples of similar rubrics can be found in the Evidence-Based Reasoning Assessment System (Brown et al. 2010), which assesses scientific argumentation, one of the foundational CAT skills emphasized in the NGSS.

CAT Assessment Construct Formulations

The construct formulation, representing a model of the cognitive processes involved in responding to the assessment, is critically important as it anchors the evidentiary argument required of an assessment and provides the foundation for development, interpretation, and

Table 2 Score level descriptions from a rubric for constructed response items assessing content area literacy

Score level	Title	Description
5	Complete	Response is complete in relation to the information provided by the text.
4	Cross-checked	Response includes multiple items of information from tactic-based sources and cross-checks or combines them.
3	Multiple	Student responds with at least two items of information from tactic-based sources.
2	Single	Student responds with one item of information from tactic-based sources.
1	Incorrect	Student gives an incorrect response.
0	Blank	No response.

validation (Brown and Wilson 2011). In particular, the construct formulation is essential to interpreting the results of an assessment. Without an explicit construct formulation, assessment results are meaningless numbers, providing no information about what a student's strengths or weaknesses are and providing no guidance for improving instruction.

Typical performance rubrics such as the one illustrated in Table 1 can be characterized as top-down. That is, they first define the top score level by adopting the text of the assessed standard, then define the lower score levels by adding increasingly derogatory modifying adjectives. In a similar manner, typical formulations of the construct of CAT, as reflected in successive grade-level standards, could also be characterized as top-down. In effect, they first describe a challenging CAT performance that the uppermost grade band is expected to demonstrate and then define lower grade-level standards as ways of attempting but failing to demonstrate that same challenging performance.

For example, the CCSS for English language arts rely heavily on "anchor standards," general statements of expected performance covering a large grade band (e.g., grades 6–12). Grade-level-specific standards are then written to look like less and less proficient responses to the sorts of tasks implied by the anchor standard. For example, the Reading Standards for Informational Text 6–12, Key Ideas and Details, Standard Number 1, based on the CCSS anchor standard described in the previous section, are shown in Table 3. The distinctions between grade levels read like a rubric, describing more and more sophisticated responses (e.g., "cite textual evidence" vs. "cite several pieces of textual evidence" vs. "cite strong and thorough textual evidence" and "evidence to support analysis" vs. "evidence that most strongly supports an analysis").

Top-down construct formulations like these, which identify a single challenging skill and differentiate grade-level standards based on better and worse attempts at demonstrating that skill, stand in contrast to bottom-up formulations that specify a progression of increasingly sophisticated skills that students are expected to learn. Although a student with weak CAT skills, when presented with a difficult task, might cite textual evidence that does not strongly support an analysis, this does not represent an intermediate and grade-appropriate skill. No teacher would approach the learning of CAT by first instructing students to look for evidence that does not most strongly support analysis before moving on to looking for evidence that does. In a similar manner, looking for two pieces of evidence instead of one represents a trivial amount of instruction that belies the difficulty of learning complex CAT skills.

Knowing the different ways that a student can fail at a challenging task may be useful for developing a reliable rubric for that task, but it does not provide useful feedback to students to

Table 3 Common core state standards grade-level reading standards for informational text 6–12

Grade level	Standard
6	Cite textual evidence to support an analysis of what the text says explicitly as well as inferences drawn from the text.
7	Cite several pieces of textual evidence to support an analysis of what the text says explicitly as well as inferences drawn from the text.
8	Cite the textual evidence that most strongly supports an analysis of what the text says explicitly as well as inferences drawn from the text.
9–10	Cite strong and thorough textual evidence to support an analysis of what the text says explicitly as well as inferences drawn from the text.
11–12	Cite strong and thorough textual evidence to support an analysis of what the text says explicitly as well as inferences drawn from the text, including determining where the text leaves matters uncertain.

Key ideas and details, standard number 1

Source: Common Core State Standards Initiative (2010). Common Core State Standards for English Language Arts and Literacy in History/Social Studies, Science, and Technical Subjects. Retrieved from www.corestandards.org/ELA-Literacy

support learning or a useful roadmap to support instruction and curriculum development. We believe that what are desperately needed are descriptions of intermediate, qualitatively distinct steps on students' paths from intuition to expertise. In order for CAT assessments to have strong formative value, their construct formulation has to expand beyond what expert CAT looks like, to include a description of how novices are expected to engage in CAT, what are the first important things to learn about CAT, and what are the more difficult or more peripheral aspects to learn.

For example, with respect to the standards illustrated in Table 3, there are many different kinds of textual evidence that could be used to support an analysis. Which kinds of evidence are the easiest to identify and incorporate into an argument, should be taught first, and should be the focus of feedback for the weakest students? Which kinds of evidence are more difficult to identify or require synthesizing multiple sources of information? Likewise, there are many different kinds of inferences that could be drawn from a text, for example, those that join pieces of information within a paragraph and those that identify an author's agenda and purpose. Which are more fundamental, and which are more advanced?

There are hints of these kinds of meaningful differences in the CAT construct formulations in the CCSS. For example, in Table 3, an inference in which "the text leaves matters uncertain" seems like a meaningful, specific type of inference that would be particularly difficult to draw. Curricula, instruction, and formative feedback could be developed that target learning how to draw such inferences. Such hints unfortunately appear to be the exception rather than the rule.

Having a cognitive model that explicitly lays out the instructionally relevant, intermediate steps leading to expertise is the key to meaningful and useful alignment between the standards, curriculum, instruction, and assessment. In science education, this idea has gained particular traction and is called a learning progression (National Research Council 2012). Recognizing that different aspects of authentic scientific practice are appropriate and achievable at different grade levels, the NGSS make the logic of learning progressions a central organizing feature, articulating a sequence of grade band endpoints that describe the levels of CAT that are expected by the end of grades 2, 5, 8, and 12 (NGSS Lead States 2013). These progressions show how each science and engineering practice, including the CAT practices of arguing, critiquing, and analyzing, can be introduced in a meaningful way at the K-2 level and develop in sophistication over the years through high school.

As an example, Table 4 illustrates the learning progression associated with Science and Engineering Practice 8: Obtaining, Evaluating, and Communicating Information. As can be seen, the different grade-level descriptions do not describe better and better responses to a singular and challenging task. Instead, they identify qualitatively different aspects of CAT and articulate which aspects are more basic and fundamental and should be learned first (e.g., "obtain scientific and/or technical information"), which are intermediate and depend on previously learned aspects (e.g., "obtain scientific and technical ideas and describe how they are supported by evidence"), and which are more sophisticated and should be learned last (e.g., "obtain scientific and/or technical information to summarize complex evidence, concepts, processes, or information…by paraphrasing them in simpler but still accurate terms").

Progression-based, bottom-up construct formulations like this provide an explicit, detailed roadmap to support CAT curriculum development and instruction. They also provide a foundation for the principled development of CAT assessments with strong formative value. In concert with assessment items that provide authentic contexts in which to apply CAT, and rubrics that provide meaningful, constructive feedback, a CAT assessment building on such a construct formulation has the potential to support student learning, not just measure existing student performance.

The Need for Formative Value in Assessments of CAT

Developing CAT assessments with formative value, not just summative value, is fundamental for realizing the high expectations set by the standards and for ensuring that existing achievement gaps are not further widened. In the context of reading, the need for formative assessments of CAT is particularly obvious. First, teachers must be able to determine if indeed a student is able to establish a literal understanding of the texts that are included in an assessment. When constructing meaning from text is but a midpoint, the nature of reading assessment must evolve. Consequences of not doing so include student failure and alienation, and the lack of knowledge on the teacher's part of what exactly a student can do or needs in order to progress. Complex, time-consuming, and consequential assessments should not be

Table 4 Excerpts from the next generation science standards grade-level science and engineering practice standards for obtaining, evaluating, and communicating information

Grade level	Excerpt from the standard
K-2	Read grade-appropriate texts and/or use media to obtain scientific and/or technical information to determine patterns in and/or evidence about the natural and designed world(s).
3–5	Read and comprehend grade-appropriate complex texts and/or other reliable media to summarize and obtain scientific and technical ideas and describe how they are supported by evidence.
6–8	Critically read scientific texts adapted for classroom use to determine the central ideas and/or to obtain scientific and/or technical information to describe patterns in and/or evidence about the natural and designed world(s).
9–12	Critically read scientific literature adapted for classroom use to determine the central ideas or conclusions and/or to obtain scientific and/or technical information to summarize complex evidence, concepts, processes, or information presented in a text by paraphrasing them in simpler but still accurate terms.

SOURCE: NGSS Lead States (2013). Next Generation Science Standards: For States, By States. Retrieved from: http://www.nextgenscience.org/next-generation-science-standards given to students who will falter in the first phase, in this case, demonstrating a literal understanding of text.

Second, formative assessment systems must be capable of helping teachers understand students' needs and strengths in all of the components of what is ultimately a grand performance (and its assessment). In the case of the second grade informational text standard described, these include analyzing related texts and images to determine their joint and individual contributions to understanding, conducting a comparison of the text and the image, and writing or speaking to explain one's work and thinking. We suspect that many second graders will require developmental work in each of these areas and that formative assessment represents the "critical middle" in between a student learning important aspects of a complex performance and eventually achieving and demonstrating success at that performance.

From our perspective, formative assessment maps well onto the notion of *zones of proximal development* of Vygotsky (1978). Accomplished teachers identify students' current levels of achievement and attainment, and they craft instruction that is best suited to build on students' current capability while scaffolding to new areas of learning. If we revisit the bulleted list that describes all that a student must be able to do to meet a second grade informational text standard, we may be concerned with the difficult challenge that it represents for some students. We may also be encouraged, recognizing the potential of formative assessment as a means to continually identify student strengths and needs within zones of proximal development and help students build toward new learning and new performances. Just as readers continually update their mental models of text as it is read, accomplished teachers continually update their understandings of individual student's strengths and needs. Formative assessment is at the center of this process.

Developing an assessment with strong formative value, however, is a difficult process, requiring a comprehensive rethinking of assessment design. Existing assessments of CAT largely do not live up to the promise of providing meaningful, constructive formative feedback. In addition, such formative assessments demand teacher expertise that is rarely supported through certification and professional development programs. Consequently, many teachers and students lack useful information about how to progress toward the types of critical-analytic thinking that are demanded by the standards and that students must demonstrate on summative assessments.

Conclusions

While the NAEP assessments and the Race to the Top assessments aligned with the *Common Core State Standards* have been extensively examined technically for their psychometric properties, less work has been done on developing the value of these assessments as formative evaluations—that is, as tools for helping teachers and students determine *how* to develop the type of CAT that the assessments purportedly measure and that the standards expect. We believe that the complex nature of CAT intensifies the importance of the question of how it develops, which, in turn, highlights the need for formative assessment that helps chart and describe this development. We argue that if CAT is to be more than an innate ability or a set of test-taking strategies, assessments must provide teachers and students with feedback that can be used to hone these skills and transfer them to non-testing environments. Such feedback is especially important in schools where teachers and students face the greatest challenges in meeting these higher expectations for performance. In other words, CAT assessments must have strong formative value, perhaps even more so than the summative value that is currently emphasized. If standards-based, high-stakes summative tests are to be used to certify student learning and teacher accountability, then formative assessment must also be available to guide both students and teachers (Afflerbach 2012); otherwise, the current wave of standards-based reform is only likely to reinforce the current achievement gap.

To meet this requirement, several shifts in assessment design will need to occur. First, learning progressions will need to be developed that describe the intermediate steps toward expertise in CAT, to serve as a blueprint for curriculum, instruction, and assessment. These progressions will help teachers and students set markers on the path to complex performances that require CAT. Second, items will need to be designed to provide meaningful contexts for observing authentic forms of CAT that will be useful for students in their everyday lives and professional careers. Third, rubrics will need to be written with formative assessment purposes in mind, to diagnose students' strengths and weaknesses and provide meaningful, actionable feedback. These aspects of principled assessment design will require considerable effort to implement, and we do not wish to understate the magnitude of the task (for an overview of the many issues involved in using principled assessment design to develop progression-based assessments, in the context of large-scale physical science assessment, see the chapter by Brown, Maderer, and Wood, in the forthcoming volume *Meeting the challenges to measurement in an era of accountability*, edited by Braun for the NCME Edited Book Series). However, despite the difficulty involved, we believe that this approach is necessary to ensure the valid, reliable, fair, and useful assessment of critical-analytic thinking.

References

Afflerbach, P. (2012). *Understanding and using reading assessment, K-12* (2nd edn.). Newark, DE: International Reading Association.

Alexander, P. A. (2014). Thinking critically-analytically about critical-analytic thinking: an introduction. *Educational Psychology Review.* doi:10.1007/s10648-014-9283-1.

Arum, R., & Roksa, J. (2011). *Academically adrift: limited learning on college campuses*. Chicago: University of Chicago Press.

Bailin, S. (2002). Critical thinking and science education. *Science & Education, 11*(4), 361–375.

Billing, D. (2007). Teaching for transfer of core/key skills in higher education: cognitive skills. *Higher Education, 53*(4), 483–516.

Bloom, B. (Ed.). (1956). *Taxonomy of educational objectives, the classification of educational goals–handbook I: cognitive domain*. New York: McKay.

Brown, N. J. S., & Wilson, M. (2011). A model of cognition: the missing cornerstone of assessment. *Educational Psychology Review, 23*, 221–234.

Brown, N. J. S., Nagashima, S. O., Fu, A., Timms, M., & Wilson, M. (2010). A framework for analyzing scientific reasoning in assessments. *Educational Assessment, 15*, 142–174.

Byrnes, J. P., & Dunbar, K. (2014). The nature and development of critical-analytic thinking. *Educational Psychology Review.* doi:10.1007/s10648-014-9284-0.

Cobb, P., & Jackson, K. (2011). Assessing the quality of the common core state standards for mathematics. *Educational Researcher, 40*(40), 183–185.

Common Core State Standards Initiative. (2010). *Common Core State Standards for English language arts and literacy in history/social studies, science, and technical subjects*. Retrieved from www.corestandards.org/ELA-Literacy. Accessed 1 Oct 2014.

De La Paz, S., Felton, M., Monte-Sano, C., Croninger, R., Jackson, C., Deogracias, J., & Hoffman, B. (2014). Developing historical reading and writing with adolescent readers: effects on student learning. *Theory and Research in Social Education, 42*(2), 228–274.

Dewey, J. (1933). *How we think, a restatement of the relation of reflective thinking to the educative process.* Boston: D. C. Heath.

Dray, A. J., Brown, N. J. S., Lee, Y., Diakow, R., & Wilson, M. (2011). *The assessment of reading comprehension in adolescents: the San Diego striving readers project (final report to the Institute of Education Sciences)*. Berkeley, CA: University of California, Berkeley Evaluation and Assessment Research Center.

Facione, P. (2000). The disposition toward critical thinking: its character, measurement, and relation to critical thinking skill. *Informal Logic, 20*(1), 61–84.

Farah, M. (2010). Mind, brain and education in socioeconomic context. In M. Ferarri & L. Vuletic (Eds.), *The developmental relations of mind, brain and education* (pp. 243–256). New York: Springer.

Guthrie, J. T., Wigfield, A., & Perencevich, K. C. (Eds.). (2004). *Motivating reading comprehension: concept-oriented reading instruction*. Mahwah, NJ: Lawrence Erlbaum Associates.

Guthrie, J., Klauda, S., & Ho, A. (2013). Modeling the relationships among reading instruction, motivation, engagement, and achievement for adolescents. *Reading Research Quarterly, 48*, 9–26.

Halpern, D. (1998). Teaching critical thinking for transfer across domains: dispositions, skills, structure training and metacognitive monitoring. *American Psychologist, 53*(4), 449–455.

Halpern, D. (2001). Assessing the effectiveness of critical thinking instruction. *The Journal of General Education, 50*(4), 270–286.

Kennedy, M., Fisher, M., & Ennis, R. (1991). Critical thinking: literature review and needed research. In L. Idol & B. Jones (Eds.), *Educational values and cognitive instruction: implications for reform* (pp. 11–40). Hillsdale, NJ: Lawrence Erlbaum Associates.

Knapp, M. & Associates. (1995). *Teaching for meaning in high-poverty classrooms*. New York: Teachers College Press.

McDonald, T., Thornley, C., Staley, R., & Moore, D. W. (2009). The San Diego striving readers' project: building academic success for adolescent readers. *Journal of Adolescent and Adult Literacy, 52*(8), 720–722.

Murphy, K., Wilkinson, I., Soter, A., Hennessey, M., & Alexander, J. (2009). Examining the effects of classroom discussion on students' comprehension of text: a meta-analysis. *Journal of Educational Psychology, 101*, 740–764.

National Assessment Governing Board. (2010). *Reading Framework for the 2011 National Assessment of Educational Progress*. Retrieved from: www.nagb.org/publications/frameworks.html. Accessed 1 Oct 2014.

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

National Research Council. (2012). *A framework for K-12 science education: practices, crosscutting concepts, and core ideas*. Washington, DC: The National Academies.

NGSS Lead States. (2013). *Next Generation Science Standards: For states, by states*. Retrieved from: http://www.nextgenscience.org/next-generation-science-standards. Accessed 1 Oct 2014.

Noddings, N. (2006). *Critical lessons: What our schools should teach*. Cambridge, UK: Cambridge University Press.

Partnership for Assessment of Readiness for College and Careers. (2013). *Advances in the PARCC ELA/Literacy summative assessment*. Retrieved from www.parcconline.org/samples/ELA. Accessed 1 Oct 2014.

Partnership for Assessment of Readiness for College and Careers. (2014). *Grades 6–11 condensed scoring rubric for prose constructed response items*. Retrieved from www.parcconline.org/samples/english-language-artsliteracy/grades-6-11-generic-rubrics. Accessed 1 Oct 2014.

Porter, A., McMaken, J., Hwang, J., & Yang, R. (2011). Common core standards: the new U.S. intended curriculum. *Educational Researcher, 40*(3), 103–116.

Quellmalz, E. S., & Haertel, G. D. (2004). *Use of technology-supported tools for large-scale science assessment: Implications for assessment practice and policy at the state level*. Washington, DC: National Research Council.

Quellmalz, E. S., Timms, M. J., Silberglitt, M. D., & Buckley, B. C. (2012a). Science assessments for all: integrating science simulations into balanced state science assessment systems. *Journal of Research in Science Teaching, 49*(3), 363–393.

Quellmalz, E., Davenport, J., & Timms, M. (2012b). *21st century science assessments*. Washington, DC: American Association for the Advancement of Science.

Schraw, G., & Robinson, D. (2012). *Assessment of higher order thinking skills*. Charlotte, NC: Information Age Publishers.

Shavelson, R. J., Baxter, G. P., & Pine, J. (1992). Performance assessments: political rhetoric and measurement reality. *Educational Researcher, 21*(4), 22–27.

Tienken, C., & Zhao, Y. (2013). How common standards and standardized testing widen the opportunity gap. In P. Carter & K. Welner (Eds.), *Closing the opportunity gap. What American must do to give every child an even chance* (pp. 111–122). New York: Oxford University Press.

VanSledright, B. A. (2013). *Assessing historical thinking and understanding: Innovative designs for new standards*. New York: Routledge.

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

Wentzel, K. (2009). Students' relationships with teachers as motivational contexts. In K. Wentzel & A. Wigfield (Eds.), *Handbook of motivation at school* (pp. 301–322). Mahwah, NJ: Lawrence Erlbaum Associates.

Willingham, D. (2007 summer). Critical thinking. Why is it so hard to teach? *American Educator*, 8–19.

Wilson, M. (2005). *Constructing measures: an item response modeling approach*. Mahwah, NJ: Lawrence Erlbaum Associates.