K-12 Center
at ETS

Invitational Research Symposium on
Technology Enhanced Assessments

Using Technology to Assess Hard-to-Measure Constructs in the Common Core State Standards and to Expand Accessibility: English Language Arts

Karen Barton and Gretchen Schultz

May 7–8, 2012

K-12 Center at ETS

Invitational Research Symposium on
Technology Enhanced Assessments

Using Technology to Assess Hard-to-Measure Constructs in the Common Core State Standards and to Expand Accessibility: English Language Arts

Karen Barton and Gretchen Schultz

CTB/McGraw-Hill

Executive Summary

There is clearly an increased attention to and demand for greater use of new technologies and features in assessments, especially when one considers the requirements set forth by the Race to the Top initiatives. Technology innovations open wide possibilities in flexibility and creativity in developing and administering assessments to and building learning environments for students with various abilities and access needs, and through varied administration devices. With advances in assessment development and administration, psychometric applications follow, such as in the areas of automated scoring, psychometric modeling, and validation research. These advancements ideally lead to time and cost savings through efficiencies in development, administration, and scoring of assessments on small and large scales, and improvements in student motivation, assessment validity, and instructional relevance as desirable targets. To make successful and actualized progress towards such advancements, the educational assessment communities will need to consider practical shifts in how assessments are designed, developed, and deployed.

As part of the Invitational Research Symposium on Technology Enhanced Assessments, we were asked to develop a technology enhanced English language arts assessment item that can measure a “hard-to-measure” construct or standard that is typically avoided in traditional assessments. As part of the development we were asked to attend to issues of measurement, including reliability and validity, student engagement and motivation, instructional relevance, and feasibility in development and scoring. This document serves to provide the considerations, challenges, and development process for a set of five technology enhanced items aligned to Common Core State Standards (CCSS) for English language arts. The technology enhancements include such features as video and audio, student choice and response flexibilities, a constrained online search environment, pop-up glossaries, online accommodations, automated and rule-based scoring for text and oral responses, and instructionally helpful performance reports. We describe the development process first in terms of the standards,

claims, targets and evidence, followed by considerations in technology enhancements, accessibility, scoring methodologies and psychometrics. We provide a full description of each item in the set in the appendix and conclude by summarizing the challenges and our recommendations for assessment development processes important to future technology enhanced assessments.

Challenges specific to innovative, technology enhanced assessment development are further described, including authoring environments that support interactivity, new issues in alignment, and iterative review cycles. We conclude with development-specific recommendations and challenges to current approaches around the following topics:

1. *Evidence as the driver* – minimizing technology-driven development, maximizing enhanced evidence
2. *Interactive capabilities* – building item authoring environments capable of demonstrating to authors examinee-assessment interactions, and ensuring such items are do not introduce new learning situations
3. *Alignment considerations and quantifications* – including alignment variables such as interactivity, scoring expectations, and technology features in alignment practices; and the use of inter-alignment reliability statistics
4. *Pool versus form development* – implications on development planning and budgeting when building and considering scoring expectations for pools of items
5. *Iterative content and technology review cycle* – inclusive of content experts, technologists, psychometricians, and scoring experts
6. *Comparability versus enhancement* – with technology that is truly enhancing, re-considerations for comparability in terms of evidence and implications for item development
7. *Accessibility* – evidence-based considerations and research on use statistics to assure fidelity in administrations with accommodation and accessible item types
8. *Metadata* – benefits to development, scoring, and reporting when capturing and leveraging item metadata

Introduction

As part of the Invitational Research Symposium on Technology Enhanced Assessments, we were asked to develop a technology enhanced item that can measure a “hard-to-measure” construct, or a standard that is typically avoided in traditional assessments. As part of the development we were asked to attend to issues of measurement, including reliability and validity, student engagement and motivation, instructional relevance, and feasibility in development and scoring. This document serves to provide the considerations, challenges, and development process for a set of five technology enhanced items aligned to Common Core State Standards (CCSS) for English language arts. The technology enhancements include such features as video, audio, student choice and response flexibility, a constrained online search environment, pop-up glossaries, online accommodations, oral response considerations, automated essay and rule-based scoring, and instructionally helpful performance reports. We describe the development process in terms of the standards, claims, targets and evidence, and considerations in technology enhancements, accessibility, scoring methodologies and psychometrics. We provide a full description of each item in the set in the appendix and conclude by summarizing the challenges and our recommendations for assessment development processes for technology enhanced assessments.

Standards, Claims, Targets, and Evidence

We first identified a single hard to measure standard. We discussed how the standard might be relevant in careers or in college coursework. We then thought about and identified how the standard might be relevant in a classroom. We found that the single standard, particularly in authentic contexts of careers, coursework, or classrooms, readily lead to the identification of additional, related standards. In particular, the standards we identified present evidence and reflect the steps in a research process. The resulting item set represents a succession of the standards that are interrelated, and build or scaffold the student experience in an attempt to parallel and simulate a more extensive research activity. The set as a whole underscores the need for students to make connections throughout a task and integrate diverse but related skills and concepts.

For each standard selected, we identified the master claim the standard supported. Then we identified the subclaims for the individual standard(s), along with instructional targets (what the teacher might want to see and what the assessment intends to measure). The evidence and actual knowledge, skills, and abilities (KSAs), one might observe in careers, coursework, and classrooms relative to the targets lead to documentation of the assessment-related KSAs. How much evidence and the particular KSAs expected for various levels of performance then resulted in scoring expectations. An example of the item template we used to document the process for developing each item is provided in Table A1 in Appendix A.

Once the observable evidence was clearly described and documented, we were able to consider evidence that the student does or does not have the KSAs expected for each target. Knowing what a student does not know is essential to accurately providing suggested instructional next steps to improve and extend a student’s ability and a teacher’s instructional decisions. For each target and related evidence, we identified possible performance limitations of the students (i.e., the specific reasons they do not know how to respond or lack the KSAs required) and/or of the limitations of the assessment, such as barriers introduced by the technology that may prevent some students from accessing and responding to the task(s). (We provide further discussion of the potential barriers in the section on accessibility considerations.)

Only after articulating the KSAs as evidence against the targets and claims related to the standards identified were we able to consider possible item designs to validly capture the evidence; the specific and appropriate technology features; and ways to minimize accessibility issues within each item. In this way, we identify evidence first, followed by item design. For this set of items, students will provide evidence of their reading, writing, research, and technical proficiencies. While each item may give focus to the measurement of a primary construct, how the constructs of the whole set interrelate is equally important. The set is, thus, reflective of tasks in which students will engage beyond the classroom. When the seventh graders for whom this task has been written achieve college and career status, the expectation will be that they have internalized a research process; however, at this point in their education, the scaffolding and modeling of a process will facilitate students' ability to make connections throughout a task or process.

Technology Enhancements

Technology in computer-based test administrations provides a wide range of potential innovations in assessment. For example, technology can enhance item presentation and response capture capabilities, as well as the design of items from the simple, multiple-choice items with limited presentation, selection, and response possibilities to more fully open and even simulation-based assessment environments. The use of technology in assessments is very attractive for many reasons. The use of technology has been shown to have a positive effect on student motivation, student engagement, and increases in higher order thinking and problem-solving skills, as summarized across studies in Cradler, McNabb, Freeman, and Burchett (2002). Technology also provides flexibility in administration conditions and increases the potential for collecting stores of interactive and cognitively relevant data. It has been said to "bridg(e) the often-perceived distance between learning and assessment" (Bechard et al., 2010, p. 8). Furthermore, the cost benefits in administration, scoring, and reporting and timeliness thereof are readily apparent with technology-based assessment design and administration, when the infrastructure supports it. A distinction must be made between technology-based and technology-enhanced. The infusion of technology in assessments, or even in instruction and learning environments, particularly when the goal is to enhance the measurement, must still be subjected to validity analyses and result in acceptable levels of validity evidence, at a minimum. In other words, the interest in and attractiveness of technology features as "enhancements" should not supersede validity (Bennett, 1999; Scalise & Gifford, 2006; van der Linden, 2002).

In developing the technology enhanced items for this project, we considered how any assessment (with or without technology) might capture the same or very similar evidence to the KSAs expected. Then, we identified how technology for an online administration would enhance our ability to validly capture the evidence, as well as how technology would enhance the way in which students are able to provide an evidentiary response. This was a critical step in ensuring that the technology tools considered, whether or not they could be readily deployed, properly reflected the evidence the assessment is intended to gather. If the technology is considered first and in terms of what technology is available, the assessment can become severely limited by the kinds of evidence and how much evidence can be captured. If that is the case, the technology is not really an enhancement and the validity of the evidence is diminished. For example, if we had first started with the technology features readily available, we would have been limited to the kinds of KSAs we were able to consider with the technology and that students can perform in the assessment, thereby limiting the evidence needed to make claims.

The technology-first approach also severely constrains item authoring. The ideas that are generated when considering how the evidence might be observed in authentic settings will inevitably be less innovative and could minimize the validity of the measures if technology is the driver. The item development process, therefore, requires the content and assessment experts to consider the technology that would be enhancing, not only in terms of motivation but especially in terms of validity. The content and assessment experts can then work with technology experts to determine available features in light of the evidence needed.

One way to organize, document, and communicate across experts the potential and available technology features is through a taxonomy, such as the one provided in Appendix B. We use the taxonomy to guide consideration of features desired and then selection of features actually available. Other similar technology based item design taxonomies have been developed and can be helpful in this process. For example, Scalise and Gifford (2006), in their comprehensive paper on technology-enhanced assessments, described an "intermediate constraint taxonomy for e-learning assessment questions and tasks" (p. 8) and provided relevant research associated with each type of item ("question or task").[1]

In our development process, given the claims, targets, and evidence in each item, the desired technology features desired were discussed with our technology experts. As the items began to take shape, the available features of the technology within each item were refined through a storyboarding and iterative review cycle. Again, the focus of each feature was on how the feature enhanced the representation of and method(s) for collecting the evidence against the targets and claims. It is important to note that the technology features available at the time of development did not drive the item design or the evidence that was captured. The evidence was specified first. When the technology features were not available to capture the evidence desired, the items required some redesign in light of feature availability.

Accessibility Considerations

Technology can enhance the validity of assessments by opening the options for assessing a construct, but only when the utilization of the technology is considered in light of the evidence about the construct. Utilization is not limited to how item developers use technology to enhance validity. Utilization also includes the way in which examinees are able to navigate, interact with, and demonstrate the expected KSAs. When either the item or the examinee is limited by the technology, it becomes a barrier, rather than an enhancement, to validly assessing the construct. This is a case for accessibility considerations with technology features and enhancements. Validity, which is really at the heart of accessibility, includes an interaction between examinees and the items to which they respond (Messick, 1995). Accessibility is therefore not limited to how students will or will not be able to access test materials, items, and tasks but is also an interaction of examinee and assessment access, both of which impact validity (Barton, 2007). Student navigation even through paper-based test forms is an important factor in evaluating accessibility; however, typically, the attention to universal design of assessments has focused on the item level. When assessment environments incorporate technology, the

[1] The items are ordered first by categories of increasing levels of the constraint associated with student responses for each item and are further described by level of item complexity. They present seven categories of item types ordered by level of constraint, from least to greatest: multiple choice, selection/identification, reordering/ rearranging, substitution/correction, completion, construction, and presentation. Within each category of item type are examples of items that increase in complexity. For example, one of the most constrained and least complex items is the true/false multiple choice type, whereas the least constrained and most complex items are Diagnosis and Teaching items.

demand on student interactivity is likely to increase. The attention to accessibility and universal design will need to extend beyond items, and even navigation, on to the entry into and interactivity within the enhanced items.

As with giving forethought to universal design, knowing in advance what evidence is to be captured, particularly within interactive items, is important in assuring that the assessment's items, environment, and technology do not become barriers to the targeted construct and that the environment is supportive of students' needs. For example, examinees should be able to equitably access the test directions and item stimuli, be able to navigate the test, understand what is expected of them in terms of their responses, and be able to provide a response in flexible ways (e.g., written or text entry, verbally, keyboard tabs, assistive technologies). At the same time, the assessment should be able to accurately and consistently assess examinees' KSAs relevant to the evidence and related target, without inadvertently assessing *unrelated* KSAs. When the level of accessibility is high, that is, when students can access and respond within the assessment and the assessment can accurately tap into students' relevant KSAs, measurement errors (both systematic and random) are reduced, resulting in increased reliability and validity.

It is also important to keep in mind that accessibility does not mean tests and delivery environments will be created such that students will no longer need accommodations. Given the evidence and claims of the assessment, that is, the validity and end use of results, it is highly probable that assessments will continue to need to be amenable to various accommodations and assistive technologies, as well as maximize accessibility through variations in item designs. To increase the accessibility of items through accommodations, certain information needs to accompany the item. For example, when a student needs to access an item using a keyboard or tab functions, or to use a screen reader online, the item must be coded in such a way that the keyboards can be used and screen reader technologies can "read" what is on the screen. The coding can be invoked in most any programming language, such as xml, html, html5, and Flash. This coding of metadata about the item is also important for the application of automated scoring methods, as well as automated form assembly, computer adaptive testing, and interoperability across platforms.

As a note on interoperability, items will need to be built so that they can be interoperable with multiple platforms and devices, while maintaining their validity. The assessment and learning technology fields are working to define interoperability standards so that items can be comparably deployed across devices and platforms, accommodations, assistive technologies, and assessment developers. A major effort, the Accessible Portable Item Profile (APIP) standards, is part of earlier and ongoing work by the IMS Global Learning Consortium (IMS GLC), a nonprofit international standards-making body for digital assessment and learning technology. It extends their work on the Question and Test Interoperability (QTI), the Access for All specifications, and the National Instructional Materials Accessibility Standards (NIMAS). APIP development was initiated by eight states and IMS GLC, with input from a broad spectrum of education officials, education industry organizations, and test providers, including CTB/McGraw-Hill.

The APIP draft standards guide the development of interoperable formats of digital assessment items, metadata describing the features of the item to support accommodations, and metadata describing students' needs. Use of these standards for item tagging structure(s) will enable assessment items and accommodations to be used across a wide variety of item-authoring tools, item pools, test forms, and delivery systems. Furthermore, the APIP draft standards describe the use of a profile of student needs intended to guide accommodation availability during online testing. Eventually, through computer adaptive testing and the collection of the student interactivity data, one could expect the profile to serve as merely a starting point, and one that needs validation particularly given the decades of research that suggest accommodation decisions are often inconsistent and can be inappropriate for individual students and assessments.

Building an Accessible Item Set

Just as evidence is the critical driver of technology enhancements, it is also the driver in evaluating accessibility for the development of this item set. To ensure the items designed for this effort validly capture intended evidence, once the claims and evidences were designated, we identified the KSAs and performance expectations for each item. Then we considered possible reasons why students may not perform as expected. For example, a student who is not able to read and understand the passage or who has difficulty identifying supporting details in a passage would receive a lower score on the items. Knowing what the student does not know is helpful, but why does he or she not know? Is it because of an inability to pick out supporting details in general, or because he or she cannot read the text? If the student cannot read the text, is that due to an obvious physical barrier, such as a student with a visual impairment or physical disability, because the student has a reading disability, or perhaps because the student is an English language learner? For each item, we considered what might keep the student from responding as expected, and in each case, we identified potential accessibility issues or barriers, and relevant solutions. If a student could potentially receive a lower score due to a visual barrier, certain item design characteristics, such as increased and clear fonts, attention to contrast, and identified accommodations, can enhance visual access, as a start.

In addition to the item design itself, the online platform, CTB's Online Assessment System (OAS), provides a read-aloud accommodation through screen reader technology, which can help students who have difficulty reading and who are permitted such an accommodation (guided by accommodation policies, or the designated accommodation profile, of course). The seamless integration for screen reading on the OAS was carefully chosen to support high-stakes tests and ensure that students have a comparable experience regardless of where they are testing. The delivery client works with screen enlargement and other visual features of software such as MAGic and ZoomText. Text Enlargement, image or art zooming, and accessible color settings are also available within the online delivery software itself. OAS incorporates its own text reader functionality based on TextHelp software hosted on CTB's servers. Available to any student as a test accommodation, this provides screen reading (with a volume and speed-adjustable, synthetic voice) of all test content for accommodated students on any workstation, without separate software installation or licensing required. The OAS platform also supports the ability to increase font size and change font and background color, which enables visually-impaired students to understand and navigate through an assessment. It includes accommodations for large print, including 18 point font, scaling of graphics, and magnification, scratch pad, auditory calming, masking bar, option eliminator, calculators, and various online manipulatives. All accommodations in OAS are configurable to a specific student's accommodations needs.

Such accommodations and item design are only part of the solution. Technology features that accompany the items also need attention in terms of accessibility. As example, consider a student with a hearing impairment accessing an item that contains video and audio features. Screen readers or read-alouds are clearly not going to meet the hearing impaired student's needs. Providing transcriptions of the video and audio elements can help. Signing avatars as on-screen translators may also be useful, as well as translations of the video and audio transcriptions that may be helpful for students with difficulties comprehending English. In addition, when words in the passage potentially pose a challenge to students, maybe because the words are of a slightly higher grade level or not typical in the English language (e.g., *relic*), pop-up definitions and word helps can be provided, thus supporting students with reading disabilities, English language learners, or any student unfamiliar with the terms in context. Being able to provide such accessibility features requires item-specific considerations and proactive development such that the features (transcriptions, translations, pop-up text, etc.) accompany the item via metadata, as the draft APIP standards require. For this item set, the OAS platform provides online accommodation tools, and the items carry metadata relevant to the pop-up word and even pictorial helps, and video and audio transcriptions.

To increase accessibility by way of student response flexibility, we provide two key innovations. First, we allow students to hear what they have written in the open-ended response areas. This provides an opportunity for students to listen to what they have written and correct any portions of text prior to submitting their responses for automated scoring. We also provide students the opportunity to make audio recordings of their responses. This flexibility could be used in three distinct ways: 1) students' oral responses could be considered as an accommodation, such as a scribe might perform, 2) the oral response functionality could be used to capture evidence with respect to claims within speaking standards (not measured here), and 3) the oral responses could be subjected to automated scoring. For example, the student might be asked to write speaking notes for a presentation and then actually read his or her notes as if presenting to an audience. In this case, we can actually score the student's oral presentation against his or her written script. This is a functionality CTB has in place now and is working to bring to a larger scale. Eventually, as student oral responses are captured, they could be shared within student groups for collaborative work and shared spaces.

We have documented the accessibility considerations and solutions in an accessibility matrix for each item, provided in Appendix A. The process required a task-analytic approach to evaluating the student experience and interactivity within each item, including any technologies, where both focal and nonfocal KSAs were considered in terms of potential barriers to student access and assessment access to what students really know and can do.

Instructional Relevance

The importance of connecting student results to instruction that can be targeted and successfully impact student progress cannot be overstated. This is true in formative or summative assessments. The item development process should therefore reflect best instructional practices and provide useful examples of and results to inform instruction. To do so, each item should consider the instructional implications given the evidence acquired and student performance, instructional targets, and claims to be made. For example, as the student meets the expectations of the task by performing well on an item or set of items in an activity, instructional extensions or enrichment activities might be considered next; or, given a lower performance, areas of remediation can be targeted by the instruction. We considered what a student's performance means in terms of instruction for each item and documented possible areas of enrichment or remediation. Once this kind of instructionally relevant information is identified and documented with the item, as with the other item metadata, the information can be readily and automatically accessed for use in student reports. For example, our reports generate instructional recommendations and links to online free and open resources for each item based on the level of student performance, as well as provide the performance level descriptors overall. We have documented samples of the detailed, instructionally relevant considerations in the score reports available within the demonstration's online application and in the item descriptions provided in Appendix A.

Scoring Methodologies

Innovations in item administration and response bring new ways to score. Items might be scored through any of a number of automated methods, each of which requires some tagging of item attributes as metadata so that the responses can be automatically captured and scored. Example methods include:

1. Simple automated scoring such as with multiple choice items or items that have a set of correct responses
2. Rule-based automated scoring with rubrics or scoring rules such as sequencing items or items with prespecified and limited acceptable responses
3. Rule and logic-based automated scoring with scoring logic programs (i.e., Boolean logic), where a range of response(s) might be acceptable, decision trees for extended responses, or any number of response variations and combinations are acceptable and need not be prespecified
4. Artificial intelligence scoring for extended open responses such as writing prompts, short constructed response items, and oral responses

Scoring is the operationalization of the level of expectation relative to the evidence. A scoring rubric simply translates what is expected of a student into a given score level. To ensure that the scoring accurately reflects the evidence, the rubric should be developed during—that is, at the same times as—item development and in line with the evidence desired. In many traditional assessments, generic rubrics are applied across items. Given the emphasis on evidence, when generic rubrics are applied, to be sure they remain consistent with the evidence targeted, some evaluation of the expectations against evidence (the scoring rubric) in each item should be considered and identified.

For the items in this set, we used a mix of generic and item-specific scoring rubrics along with automated scoring algorithms – automated scoring engines and rule-based models. Each unique rubric was developed to reflect the expectations of the response(s) in terms of evidence commensurate with the KSAs of novice, master, and expert students. For fairly interactive items that do not require free text entry, such as described by Scalise and Gifford (2006) as moderately constrained and complex item types, the online system captures specific student responses automatically and applies scoring rules that result in a summary score based on expectations set forth in a holistic scoring rubric. For example, one of the items requires students to organize text and video clips into a Venn diagram, table, or presentation slides. Each clip of text or video is tagged in the online system as being unique to the video, the text, or shared by both the video and text. Based on a set of rules, the online system tabulates the student-placed correct and incorrect responses, and translates the results into a holistic score.

For open-ended, free text entry items, more generic scoring rubrics are applied through an automated scoring engine that accounts for multiple features, the most extensive being for the extended writing essay, including the following:

- Organization
- Development
- Language Use/Grammar and Usage
- Range of Vocabulary
- Appropriateness of Details

- Syntactic Variety
- Sentence Variety
- Mechanics

CTB's Bookette automated scoring engines, used for this item set demonstration, operate on approximately 90 features of student-produced text. These features may be classified as structural, syntactic, semantic, and mechanics based. Most commonly, the features are used to model trait-level scores, which may be reported separately and/or combined to produce a total writing score. The analytic scoring guide that underpins the CTB system produces integer scores (ranging from 1 to 6 points) based upon well-recognized traits of effective writing, all of which are shared in exact or similar form to the traits for the items in this set (listed previously):

- Organization
- Development
- Word Choice/Grammar Usage
- Sentence Structure
- Mechanics

CTB's Bookette engines are able to model trait-level and/or holistic-level scores for essays with a similar degree of reliability to an expert human rater. The engines use a natural language processing system with a neural network to model expert human scores. The engines are trained using essays that have been scored by expert raters and validated against a separate set of papers also scored by expert raters during the model-building phase. The engines are then monitored using human-to-engine comparisons during the implementation phase for uses in which students are scored "live" for accountability purposes.

Note that the first two items in the set—a short constructed response and an extended writing essay—were previously field tested in a paper administration and scored by human raters. Because we have actual student responses, we were able to develop and apply the automated essay scoring (AES) engines. The papers from the field tests were originally given on paper and scored by trained raters, where 5% of the papers were subjected to second scores, from which inter-rater reliability statistics were calculated. Then a randomly selected sample of 500 papers was transcribed into digital format. All digital papers were then rescored for a census second read. Approximately two-thirds of the papers were used to train the AES engines, and the remaining one-third were used for testing various AES engines and feature sets, and for final engine validation and selection. Future engine building should be sure to include samples of students from various ability groups and with various access needs to ensure their response patterns are adequately represented and validly scored.

For each item, regardless of scoring method, the rubric is provided directly to the students. The reasons for exposing the rubrics to the students are (at least) threefold:

1. Students need to understand what is expected of them during any assessment. We tell them what to do and how their responses will be evaluated.
2. Teachers should understand the meaning of student scores for each item, particularly when automated scoring is applied and whenever teachers will use the information to inform instruction in short order (formative contexts) or in the future (summative contexts; Scalise & Wilson, 2006).

3. Reporting student performance diagnostically and/or in instructionally relevant terms requires clear indication of actual performance against expectations.

Psychometric Considerations

With initial administration of most any item, the more technologically enhanced and interactive items should be subjected to interactive data analyses and cognitive labs with various student groups, particularly students with various access needs, and larger field trials. Then, as the student interactivity data for such items accumulates, scoring rules and AES engines should be verified and adjusted as necessary and as appropriately relevant to the evidence and claims targeted by each item. Review of student responses and manual adjustment of automated scoring rules should be conducted, which is very common, such as with items of the gridded-response type. At some point in the future, the intelligence of the scoring rules, with the support of analyses of validity data, may provide dynamic scoring adjustments or Boolean logic application, in place of some of the hard-coded, rule-based scoring algorithms, leading to increased scalability and efficiencies.

Given the items in this set are presented as a scaffolded set, it would be prudent to evaluate the items independently as well as collectively. This will require the separate administration of each item, apart from the set. Such evaluation will inform future test assembly requirements of the items and in terms of computer adaptive testing algorithms. In case the items are found to be more appropriately administered as a set rather than independently, the form administration and assembly requirements for keeping the items as a set can become critically important, not only to the assembly, but also to the draw of the item pool (and requirements for additional items in the pool). Consider further the utilization of automated test assembly, be it dynamic or static, and computer adaptive testing. In such situations, if the items must be administered in a set, there will need to be metadata that is associated with the items to identify them as part of the set and the requirements, either for manual or algorithm form selection and administration. There are solutions, such as van der Linden (2005) described through the use of shadow tests that leverage the constraints of the set for administration in computer adaptive testing, a functionality readily applied in CTB’s adaptive test assembly and computer adaptive processes.

In terms of psychometric models, there are multiple item and form design possibilities for which the decisions surrounding the application of psychometric models to the items depends on the intended uses of the items and results. If the desire is to maintain a pool of items that are on the same psychometric scale, the items could be calibrated via item response theory (IRT) modeling with the rest of the items in the pool. This requires the majority of the items to be administered at the same time or at least linked if they are in separate administrations. With the common use of and assumptions of independence associated with the unidimensional IRT model, future research may reveal that the discrete item scoring, that is, each item independent of the set, even at the domain level, may introduce dependency in items. Given that the items in this example set often share common stimuli and represent a progression of research activities around a common theme, the testlet response theory (TRT; Wainer, Bradlow, & Wang, 2007) may be a more appropriate modeling method than the unidimensional IRT. When there is multidimensionality in the data, the multidimensional IRT model (Reckase, 2009) might be more appropriate. Preliminary findings from ongoing research by Skorupski, Barton, and Springhorn (2012) based on simulations of multidimensional performance events (not very different from the item set described here) indicate that even simpler models, such as the unidimensional IRT model, remain comparable to more complex, TRT and multidimensional IRT models.

Further research is needed to validate the application of the appropriate psychometric model(s) with live data across items distinct from and within a given set.

If the desire is to report domain-level scores, such as a score for Geometry or Reading of Informational Texts, the items might be scaled by domain rather than by form, whereby all items that align to the same domain are scaled together. Given the desire to gain diagnostic and/or instructionally sensitive information, or as in traditional summative reports, to provide domain-specific performance information, domain scaling offers a solution resulting in more comparable domain scores across administrations, and over time. For example, typical domain-specific performance is reported in either raw score, percentage correct, or with the use of Bayesian applications within IRT such as the objective performance index. In these instances, the domain scores are not strictly comparable across time or administrations, even when the items are typically on the same scale as the total score. If domain scores are scaled independently, the domain scores can be compared across time and administrations, where the total score is likely to be a composite of domain scores.

The challenge to domain scaling is ensuring that enough items aligned to the domain are available during scaling. Though item pools can and likely will continue to be proportionally representative of a blueprint, adhering to a pre-specified and form-specific blueprint would further limit the utility of domain scaling, not to mention of computer adaptive testing, which relies on a fairly large item pool at the level of scaling and reporting. One way to increase the size of item pools is by embedding items into adaptive or static administrations. The items could be embedded directly into operational and ongoing administrations (such as with formative assessments), and distributed to specific subgroups, or students with various needs. The use of online administrations to embed items could also be used to conduct experimentally-designed studies with treatment/control groups to validate evidence, instructional programs, or even accessibility profiles and accommodation use and needs. For example, students could be administered nonoperational items or even practice items with and without access features to evaluate the use of the features and the validity of the pre-determined accessibility profile (as with APIP); or to evaluate a learning progression and variations thereof across students; or data to validate the impact of or claims made by an instructional program on improving student scores could be analyzed (if the student-level data are available regarding program participation). In the latter example, students participating (treatment) or not (control) could receive items aligned to the instructional targets of the program. Similar such assignments of tests and instruction have been extended through algorithms developed at CTB and are being analyzed as part of evaluations of a school-based learning program.

A unique challenge in this particular item set is that one of the items allows students to choose one of three response methods for organizing information: a Venn diagram, a table, and three presentation slides. It is likely that students will choose different options, which will decrease the number of students responding to the item and choice combination and complicate the algorithms for adaptive testing. A smaller sample size for each option could become a challenge to adequately evaluate the performance of the item, especially when IRT models are desired. This can be overcome, however, as data are accumulated. Additional analyses would also need to be conducted on the characteristics of students relative to the choices that they make and implications for evidence and level of student performance.

Assessment technology presents a unique opportunity to evaluate how students navigate through an assessment and engage in each item. Even with static and less interactive items, multiple data points can be captured, such as response time, response changes, tool usage, search logs, item navigations and items skipped. For more interactive items, various item-specific interaction data, such as frequency of drags and drops, size or length of text selected for dragging into more open response areas, number of variations for grouping or sequencing, and extent of graphing response variations can be made available for additional analyses. These data and more represent a digital history of how students access the assessment, which cannot be captured on paper, especially with large samples and only rarely through intense cognitive labs with small samples of examinees. While many of the data variables exist in paper-based assessments, such as response time and student navigation or item skipping, it is only with technology that those data can be systematically observed and captured. Future measurement methods will need to consider how to handle and leverage the newly observable data and its relevance to and enhancement for measuring evidence.

Summary and Recommendations

There is clearly an increased attention to and demand for greater use of new technologies and features in assessments, even in large-scale summative assessments including those being developed under the Race to the Top initiatives. Technology innovations open wide possibilities in flexibility and creativity in developing and administering assessments to and building learning environments for a variety of applications and us, across students various abilities and access needs, and through a range of administration devices. With advances in technology-based assessment development and administration, psychometric applications follow, such as in the areas of automated scoring, modeling, and validation research. These advancements ideally lead to time and cost savings through efficiencies in development, administration, and scoring of assessments on small and large scales, and with improvements in student motivation, assessment validity, and instructional relevance.

To realize the benefits of these advancements, critical attention must be made to the assessment processes. For the current symposium initiative, we have only begun to describe areas to attend to and specific steps to take to push towards valid innovations in technology-based assessment. We have outlined a process for developing a small set of technology enhanced assessment items. Any one of the considerations could be extensively described, far beyond what is suitable for this context. We described the attention to evidence—the KSAs that must be observed to validly make a standards-related claim. We provided an evidence-first context around the valid selection of available technology features, where technology truly enhances yet does not drive development, and where the evidence and technology are evaluated in terms of accessibility and validity. We described scoring methodologies that serve to enhance the validity of item development, where rubrics are considered at the time of item authoring, along with scoring approaches that enhance the efficiencies of scoring and reporting through automated scoring engines and rule-based algorithms. Psychometric considerations have been provided that server merely as precursors to the extensive discussions and research that should take place in the near future, particularly given the assessments may result in high-stakes use, including teacher evaluation.

While additional challenges and considerations should be addressed in the development process and subsequent research, we conclude with a focus on issues in the development process of most immediate consideration, from authoring and alignment to review, and recommendations to the field across development, research, and use of data.

Authoring

As a start, the development of truly evidence-based assessments requires both deep thought and documentation around the specifications to guide development. We believe that it means much more than considering the number of items desired to populate an item pool. It means much more than specifying item types to be developed to specific standards or types of items with certain technology features. The item authoring environment should support clearly evidence-based item design and use of technology.

While this may seem obvious to some, specific attention must be paid to the creation of the item-authoring tools in light of evidence, accessibility, and scoring, and where technology features are enhancements and not drivers of development. Otherwise, the resulting items may have the “coolness” factor, but will drive up costs, and will not be as accessible, useful, or desirable, especially for building high-stakes assessments with stringent validity requirements. Part of the authoring process and tools should include functionality for content experts to evaluate assessment items or tasks as they would be presented to examinees during administration. Features of items or tasks should be “live” so that the item authors can evaluate the validity of the item interactively, prior to submission of the item for quality and accessibility reviews. The reviews should also help authors ensure that each item does not itself become a new learning scenario for examinees. The item response expectations should be clear, the assessments navigationally intuitive, and for newer item types, accompanied by practice item sets.

The technologists and content experts must also work collaboratively prior to item authoring to ensure that the authoring environment and tools capture the important metadata for interoperability. Furthermore, the specifications and training for authoring will need to address the definitions of metadata to ensure the metadata are reliably documented. The item authoring environment should also automatically present accessibility considerations for technology features, and constrain item formatting choices to ensure items are developed in universally designed formats. Such constraints support authors who may not understand universal design issues, as well as minimizing rework and related costs to reformatting items to ready them for administration. Such rework could also inadvertently change the author’s intent with and validity of the item. Even with accessibly supportive authoring environments and tools, content experts and item authors, if not well versed in accessibility issues, should at minimum be required to complete accessibility training and document accessibility issues in matrices such as those provided in Appendix A.

Alignment

Prior to item authoring, content experts should be heavily involved in operationalizing the standards through to the actual evidence desired, which requires clear documentation and evaluation of alignment. Alignment of items to the grade, standard, claims, and targets will continue to be needed. It is clear, however, that alignment of items will extend well above and beyond what is typically considered “alignment,” particularly when the enhancements include highly interactive situations, a various KSAs, and a range of evidence-based performance level expectations. As an example, technology-enabled performance events will likely be designed to measure multiple standards and cognitive skills at various levels of proficiency, where valid evidence that can be reported for instruction is included in the item or task designs. As such, multiple types of alignments (grade, standard, proficiency level expectations, construct, error constructs) and validity considerations will be required. As instructionally informative and diagnostic data are increasingly desired from assessments, and as demands grow for student performance profiles, progress against items or tasks within nodes along learning progressions, growth measures, scaffolding and feedback in learning environments, and the like, extensive alignments will need to be made. Such levels of alignment will likely shift the way in which items are developed and then subjected to quality reviews.

Review

In addition to reviews of items for correctness and alignment, a review of the technology enhanced items and tasks in light of the level of interactivity, the interaction between the examinee and the item, and how each access the other will be important to assuring that each item continues to provide valid evidence. This includes a review of the scoring rubric or logic and the level of alignment of each score expectation to the claim(s) targeted, as well as the impact of interactivity on the validity of the measure of those claims. A key focus and subsequent alignment consideration with technology should be whether or not the features are actually enhancing.

Research should be part of the review process. More extensive reviews through usability pilots should also be implemented. How students will interact in these new technologically enhanced environments has yet to be deeply researched. Therefore the items and tasks should be subjected to a range of studies, including cognitive labs or think-alouds, the results of which must inform further item development. One might even consider including video observations in small scale trials to provide data not only for validity analyses to researchers, but also to inform and train content experts and item authors on just how examinees respond to and interact with the items and tasks. Where these studies are not yet in place, we recommend embedded training item sets into the authoring environment and quantifying the authoring process. For example:

1. Alignment training: Similar to check-sets in hand scoring, previously created technology enhanced items can be provided to authors electronically to check their judgments on alignment (grade, standards, claims, KSAs related and unrelated, error strategies, instructional implications), interactivity, and scoring expectations. Reviewers and trainers could use the information to provide feedback to authors to improve item development, which should decrease editing efforts, time, and costs after items have been submitted for review.
2. Inter-alignment reliability: Check-set alignment results could also be analyzed to capture inter-alignment reliability statistics, similar to inter-rater reliability statistics.
3. Finally, the reviews should be cyclical and informative, where results of the initial reviews, research studies, and development and performance data collected feed back into item development processes.

Recommendations

The business model(s) and processes for developing innovative, technology-based assessments, as well as technology-based learning environments can initially be expensive and time consuming. Once the authoring process and technologies are in place, the savings can be readily realized. It often requires, however, a shift in assessment development paradigms and related systems.

1. *Evidence as the driver* – First and foremost, the validity of the technology enhanced items, when considered initially through the lens of evidence, results in items that are not driven by what is available, but by what KSAs are targeted and observable. That means, while the authoring environment should encourage the use of available technology features, those features should not drive the item design. Items should be designed and documented with the most valid and innovative thinking in mind, regardless of feature availability. When the

technology catches up, the best items will have already been developed. Until then, the evidence should continue to drive the design, tempered by available technologies.

2. *Interactive* capabilities – Building an item authoring environment that is capable of demonstrating to authors the interactions required by an examinee to respond is an important functionality in assuring the KSAs targeted are clearly evidenced. It requires some degree of paradigmatic shift in how an author considers evidence and builds an item with technology to capture that evidence, barrier-free. The reviews should also help authors ensure that each item does not itself become a new learning scenario for examinees. The item response expectations should be clear, the assessments navigationally intuitive, and for newer item types, accompanied by practice item sets.

3. *Alignment considerations and quantifications* – What is considered "alignment" will evolve to include additional variables, beyond standards and to interactivity, scoring expectations, and technology features as evidence enhancements (not detriments). Assessment developers will benefit from conducting alignment reviews with highly interactive items and capturing inter-alignment reliability statistics to identify mis-alignments and confirm key alignment variables. This is fertile ground for new research.

4. *Pool versus form development* – Traditionally, item development planning and budgeting starts with form design, blueprints, numbers of items by type, total score, and so forth. When items will be developed for an item pool from which many forms can be created, including those administered computer-adaptively, the focus of item development may shift to support the accumulation of evidence to make claims rather than a strict blueprint and set number of item types. For example, we have found that item developers must shift their thinking around scoring expectations – from intrinsically "weighting" the importance of a given item and type in a probable "test form and blueprint" scenario to quantifying the actions and expectations for each unique item, along with the expectations and scores relative to evidence. In addition, ample numbers of items across various performance levels (e.g., novice, master, expert) will be needed and available for increasing the amount of statistical information across such expectations and student ability, especially with CAT, as well as overages to support reliable domain or objective-level scores. As instructionally relevant data demands grow, so too will the demand for items of various complexities and diagnostically rich data.

5. *Iterative content and technology review cycle* – The item-authoring system should also support and require multiple iterations of review between item authoring and technology-based item deployment stages, where content experts, technologists, psychometricians, and scoring experts are involved. There should also be a mechanism that feeds research results back into the item development system to continually improve item development.

6. *Comparability versus enhancement* – As technology provides enhanced opportunities to capture authentic evidence, issues of comparability between technology based and paper-based assessments should be reconsidered. No longer can items be authored as if on paper and then transferred into an online space requiring that the paper and computer based versions remain "comparable." If the technology is truly enhancing, it should offer something that is not available on paper. That means items will shift to online development, literally and paradigmatically, where paper-based forms (as opposed to the current focus on

simply transfer to computer-based forms) will need to be reviewed in terms of the comparability of measures of construct, such as might be considered across multiple, parallel, and unique test forms. In other words, given the targeted claims, alternative methods for capturing evidence on paper should result in comparable claims to the technology enhanced methods.

7. *Accessibility* – The focus on accessibility at the time of item-authoring, especially when technology is included, is a clear cost saver. As items are developed in light of evidence and specifically in terms of barriers to evidence, theoretically, fewer special forms will need to be developed, administered, scored, and processed separately. Retrofitting items, systems, technology, and evidence is costly, not only in time and money, but also in validity. We strongly encourage research on use statistics to validate the assigned accommodations, and to assure fidelity in administrations with accommodations and accessible item types. Additionally, response flexibility, such as through oral response, should be considered in item and scoring development. Finally, when automated scoring is utilized, analyses and engine building should include responses from students with various abilities, accommodations, and access needs to ensure their response patterns are validly scored.
8. *Metadata* – With item metadata in hand, there can be benefits to interoperability, development, scoring, and reporting. For example, automated form selection or adaptive administration algorithms can readily draw on pools of items rich with metadata to present highly efficient (psychometrically and administratively), accessible, and secure test administration, and form assembly. Reports can further benefit from the item and scoring expectation metadata, by using metadata with item scores to auto-assemble performance profiles and relevant instructional feedback. To improve assessment and student accessibility, accessibility metadata should be captured and documented, including text tagging and accommodations information in item profiles (such as described by APIP). Thus, upfront and detailed attention to and documentation of the desired metadata is critical.

In the end, regardless of the technology features desired, assessments should provide valid, trustworthy results. If the results are not valid and trustworthy, they do nothing to help educators and students. When technology provides opportunities for innovations in assessment, active attention to and tangible support for models and processes in developing, administrating, scoring, and reporting assessments will be necessary, along with validity research, innovations in scoring and psychometrics, and shifts in business approaches and existing paradigms.

Acknowledgements

We would like to thank Pat Forgione and Nancy Doorey of the K-12 Center at ETS for providing us the privilege to participate with our fellow presenters from ETS and SRI in this symposium. We thank our esteemed moderators, Sue Rigney and Kathleen Scalise, for their encouragement, guidance, and constructive feedback on our ideas and efforts for this work. We appreciate the CTB/McGraw-Hill executive leadership and support by Richard Patz, and especially the TEA team for their dedication, innovativeness, thoughtfulness, and collegiality, including: Nate Cohen (Technology Lead), Premratan Kalani (Automated Scoring), Jaivardhan Mehrotra (Technologist), Robynn Upton (Content Editor), and Bryan Waters (Automated Scoring).

References

Barton, K. (2007). Validity and accommodations: The journey toward accessible assessments. In C. Cahalan Laitusis & L. Cook (Eds), *Large-scale assessment and accommodations: What work*s? Arlington, VA: Council for Exceptional Children.

Bechard, S., Sheinker, J., Abell, R., Barton, K., Burling, K., Camacho, C., Cameto, …Tucker, B. (2010). *Measuring cognition of students with disabilities using technology-enabled assessments: Recommendations for a research agenda*. Dover, NH: Measured Progress, and Menlo Park, CA: SRI International.

Bennett, R. E. (1999). Using new technology to improve assessment. *Educational Measurement: Issues and Practice, 18*, 5–12.

Cradler, J., McNabb, M., Freeman, M., & Burchett, R. (2002). How does technology influence student learning? *Learning and Leading With Technology*, *29*(8), 46–49. Retrieved from http://dixiesd.marin.k12.ca.us/dixieschool/Dixie%20Tech%20Plan/ResearchCradler.pdf

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*, 741-749.

Reckase, M. (2009). *Multidimensional item response theory.* New York, NY: Springer.

Scalise, K., & Gifford, B. (2006). Computer-based assessment in e-learning: A framework for constructing "intermediate constraint" questions and tasks for technology platforms. *Journal of Technology, Learning, and Assessment, 4*(6). Retrieved from http://www.jtla.org/

Scalise, K., & Wilson, M. (2006). Analysis and comparison of automated scoring approaches: Addressing evidence-based assessment principles. In D. M. Williamson, I. J. Bejar, & R. J. Mislevy (Eds.), *Automated scoring of complex tasks in computer based testing* (pp. 373-401). Mahwah, NJ: Erlbaum.

Skorupski, W., Barton, K., & Springhorn, A. (2012). *A comparison of approaches for scaling multi-step performance events.* Paper presented at the annual conference of the National Council on Measurement in Education. Vancouver, BC.

Wainer, H., Bradlow, E. T., & Wang, X. (2007). *Testlet response theory and its applications.* New York, NY: Cambridge University Press.

van der Linden, W. (2002). On complexity in CBT. In C. Mills, M. Potenza, J. Fremer, & W. Ward (Eds.), *Computer-based testing* (pp. 89–102). Mahwah, NJ: Erlbaum.

van der Linden, W. J. (2005). *Linear models for optimal test design*. New York, NY: Springer.

Appendix A: Item Set

For each item in the set, we have identified CCSS, claims, targets, and evidence. We describe the technology and accessibility considerations for each item, designed as a result of the evidence required. We end by discussing the scoring methodologies for the item set. An example of the template we used to document the process is included as Table A1.

For each item, we provide the accessibility matrices in this appendix. In each matrix, we describe the steps required for a student to navigate through the item and provide a response. The focal KSAs and nonfocal KSAs are documented, and for each, accessibility concerns are noted. Certain subpopulations were targeted for this exercise and are noted as E (English language learners), V (visually impaired), H (hearing impaired), M (motor impaired), and one blank for any additional group identified.

The resulting item set represents a succession of the standards that are interrelated and build or scaffold the student experience in an attempt to parallel and simulate a more extensive research activity. The set as a whole underscores the need for students to make connections throughout a task and integrate diverse but related skills and concepts. For this set of items, students will provide evidence of their reading, writing, research, and technical proficiencies. While each item may give focus to the measurement of a primary construct, how the constructs of the whole set interrelate is equally important. The set is, thus, reflective of tasks in which students will engage beyond the classroom. When the seventh graders for whom this task has been written achieve college and career status, the expectation will be that they have internalized a research process; however, at this point in their education, the scaffolding and modeling of a process will facilitate students' ability to make connections throughout a task or process.

Similarly, the CCSS for ELA and Literacy illustrate the importance of integrating knowledge and skills, moving away from an atomistic approach that focuses on small, discrete sub-skills to a holistic approach. While knowledge of specific content pieces is important, much more vital in the 21st century is how one connects and applies that knowledge when reading, writing, and thinking critically.

The items in the set share a common informational passage on the topic. During the development, we were able to draw on and learn from a pool of items that have already been field tested using the same passage and on our recent cognitive labs around performance assessments, considering the use of existing items and developing new items for this symposium. We describe each item in the resulting set next.

Table A1. Example Item Template

Item form	A
Item no.	4
Targeted construct/overall claim	W.7.6. Use technology, including the Internet, to produce and publish writing and link to and cite sources RI.7.7. Compare and contrast a text to an audio, video, or multimedia version of the text, analyzing each medium's portrayal of the subject (e.g., how the delivery of a speech affects the impact of the words) W.7.9. Draw evidence from literary or informational texts to support analysis, reflection, and research W.7.7. Conduct short research projects to answer a question, drawing on several sources and generating additional related, focused questions for further research and investigation
Master claim	Students can engage appropriately in independent inquiry to investigate/research topics, pose questions, and gather and present information
Subclaim 1	Students identify similarities and differences between and among the sources
Evidence	Venn diagram, table chart, graphic organizer, outline pros/cons, build a presentation for/against, etc.—create on own, or complete one provided, and pull from existing list of statements/citations For any type of response, must provide minimum number of similarities, plus minimum number of differences; must come from actual sources (not opinions); level of specificity important Cannot confuse similar with different
Scoring	0 = no response, incorrectly identify similarities as differences, or vice versa 1 = correctly identifies minimum number of similarities and minimum number of differences but does not draw from sources—uses opinions, adds to, or detracts from source or not specific 2 = correctly uses sources and correctly identifies similarities/differences
Evidence of what they do not know	Does not use the sources, only provides opinions—does not understand how to use sources effectively to inform Does not understand similarities/differences Unfamiliarity with one or more of the mapping/response tools (Venn, table, etc.)
Barriers	Inability to use one of the response modes due to physical disability (though the point of providing alternatives is to minimize this barrier)
Access	Providing multiple modes for responding; oral response option?
Links to instruction	Introduce variety of note-taking tools and presentation tools for identifying information in sources and comparing similarities/differences. Recognizing source-provided information as distinct from one's own opinion and not adding to/detracting from the source.
Instructional extensions	Practice taking multiple sources and written text about the source to evaluate accuracy of text to source. Conduct taped interviews and provide report of similarities/differences, distinguishing source-provided evidence from own opinion/details.
Technology being requested for item	Three response areas as optional—student clicks graphic of tool they plan to use, send to page with that tool and source information ready to drag/drop, highlight, area for open response?
How tech will help student show response	Multiple options for response mode; technology helps to provide the options and capture the information/response of each
How item will look when enhanced	Venn diagram Two-column table (similarities/differences) Presentation slides—2 (1 similarity, 1 difference)

Overarching Standards

We began by selecting a standard that is not typically measured in traditional assessments. Specifically, we selected the following Grade 7 Standard from Writing:

> W.7.6. Use technology, including the Internet, to produce and publish writing and link to and cite sources as well as to interact and collaborate with others, including linking to and citing sources.

The standard requires the use of technology to publish well-supported writing, something that is not typically leveraged even on computer-based assessments for large-scale use. The standard readily signals instruction in preparing students for future writing situations. Consider writing for a college class or in a career in journalism or research, where one is likely to take steps in preparation for the writing piece: searching the Internet or journal database, reading and conducting research on multiple sources, and then evaluating each source in light of the writing topic and eventually making comparisons across sources—quite akin to a research activity.

Thus certain related Reading and Writing standards readily follow:

> RI.7.7. Compare and contrast a text to an audio, video, or multimedia version of the text, analyzing each medium’s portrayal of the subject (e.g., how the delivery of a speech affects the impact of the words).

> W.7.7. Conduct short research projects to answer a question, drawing on several sources and generating additional related, focused questions for further research and investigation.

Because these three standards blend nicely, we considered developing a set of items that leverage each of the standards to simulate a typical research process within an assessment environment. Before developing the actual research activity and related technology to support the research within an assessment environment, we first identified the actual claims to be made as a result of the student’s behavior in the activities. For example, at the level of a journalist, what would the editor expect from the journalist in terms of conducting research, or what would a professor expect of college students, to what would a teacher expect of students working on a research project, and finally, what would be expected of an examinee on an assessment?

Item 1: Short Constructed Response

The first item is aligned to the following Reading CCSS and requires students to consider and become familiar with the passage provided to them. As the students progress through the item set, they are provided space for taking notes on sources as they go. This will help them throughout the items, particularly for the last two items in the set.

> RI.7.1. Cite several pieces of textual evidence to support analysis of what the text says explicitly as well as inferences drawn from the text.

The Master Claim and related Subclaim for the standard are related to reading literary and informational texts and are provided here along with the instructional and assessment targets and related evidence:

> **Master Claim:** Students read and comprehend a range of sufficiently complex literary and informational texts.

Subclaim: Students demonstrate comprehension and draw evidence from reading of grade-level, complex informational texts.

Target 1**:** Demonstrate understanding of informational text and cite evidence to support that understanding.

Evidence 1**:** Students can analyze the text and cite additional textual evidence to support the identified idea. The evidence should be congruent with the central idea. This can be in the form of their own words, summarizing multiple supporting ideas from the text, or by repeating the author-provided phrasing of the supporting ideas.

Real-world connections. As a whole, the set of items is reflective of the process in which one would engage when conducting research on or exploring a topic to gain information. The evidence students provide in response to the items/task mirrors the evidence one might provide in the real world. For example, the first item calls for students to provide evidence that they understand a key idea of an informational text. For the assessment item, students write (or speak) their response to show their understanding. Outside of assessment, one might share information that has been learned from a magazine article, newspaper, or other informational text in a conversation with or message to a friend or work colleague. With certainty, citing key details from text is ongoing throughout one's professional and personal lives.

To address the evidence for this target, we selected a previously field tested, short constructed response item in which the student must read the passage and then respond to a prompt with details from the article.

Technology enhancements. The authentic passage is an informational text that includes some terms that do not serve the measure of the target. They are slightly above grade level and may pose challenging for students, including students with reading difficulties and who are English language learners. Therefore we include word helps and graphics for select terms, accessed via hyperlinks, which cause the help, definition, or graphic to pop up.

Accessibility matrix. In addition to the readily available online accommodations in CTB's OAS platform and the use of the pop-ups for challenging words, the response area provides an option for students to hear what they have written, a common accommodation for students with learning disabilities. Students can also record their oral responses to the item. The oral response is a working technology enhancement; however, the use of the oral response is not yet functional. When it becomes functional, it will minimize the need for students to use scribes, and as oral response data accumulate, AES engines can be trained and validated to score the responses.

Table A2. Identifying Accessibility Barriers: *Item 1_SCR*

	Description of steps	Focal KSAs			Nonfocal KSAs		
Step 1	Read or listen to a passage about the Baghdad Battery	Demonstrate comprehension and draw evidence from reading of grade-level, complex informational texts			Perceive and utilize key word pop-ups [✓ E] [✓ V] H [✓ M]	Perceive and utilize read-aloud feature [✓ E] V [✓ H] M	Perceive text E [✓ V] H M
Step 2	Provide text-based, open-ended response to short prompt	Identify and provide accurate evidence from passage in response to prompt	Cite explicitly stated textual evidence to support analysis of what text says	Demonstrate comprehension of prompt and text	Write/keystroke entry online E [✓ V] H [✓ M]	E V H M	E V H M

Table A3. Identifying Ways to Enhance Accessibility: ***Item 1_SCR***

For each step: top row—identify relevant populations (1 per column); middle—specify the difficulty; bottom—suggest solution.

Step 1	Population →	Visually impaired	Motor impaired	ELL	
	Difficulty →	Screen size, pop-up feature	Scrolling, mouse clicks	Passage terminology	
	Solution→	Read-aloud with key word and button tags, magnification	Keyboard short-cut accessible	For culturally challenging terms, add pop-up definitions and graphics	
Step 2	Population →	Visually impaired	Motor impaired		
	Difficulty →	Text entry of response	Text entry of response		
	Solution→	Assistive technology, scribe, oral response	Assistive technology, scribe, oral response		

Scoring and Instructional Links. The first item is a short constructed response, scored with a two-point holistic, item-specific rubric. Recall that the papers from the paper-based field testing were scored by trained raters, where 5% of the papers were subjected to second scores, from which inter-rater reliability statistics were analyzed. Then a randomly selected sample of 500 papers was transcribed into digital format. All electronic versions of the papers were rescored for a census sample second read. Approximately two-thirds of the papers were used to train CTB's Bookette engines, and the remaining one-third was used for building the engines and feature sets and for final engine validation and selection. Note that this type of item presents unique challenges for the development of AES engines. The item is undergoing AES engine training and validation and will be available at the time of the demonstration and final version of this report.

Item 1 scoring rubric. The scoring rubric for Item 1 follows.

2 points: Response is a thorough explanation of how the artifact discovered in 1936 came to be called the Baghdad Battery. It includes relevant details from the text to provide evidence of complete understanding of the task and the text.

1 point: Response is a partial explanation of how the artifact discovered in 1936 came to be called the Baghdad Battery. It includes minimal details from the text to provide evidence of some understanding of the task and the text.

0 points: Response is an inaccurate or irrelevant explanation of how the artifact came to be called the Baghdad Battery. It includes no relevant details from the text to provide evidence of any understanding of the task and the text—or no response was provided.

Item 1 instructional links. Samples of suggested instructional links for each score point in Item 1 follow.

2 points Expert/Master: The student shows readiness for

- reading more complex text, including non-print text, that requires deeper inferential thinking
- making connections and comparisons within and among texts

Sample suggested resources for Expert/Master:

1. Strategies to Help Readers Make Meaning through Inferences
   Students use prior knowledge to make inferences about the text that they are reading. The strategy works well with nonfiction and fiction texts.
2. 15 strategies for reading a complex text
   Here's a range of methods for approaching a complex text. You may find different methods are helpful on different occasions.

1 point Novice: The student needs additional help

- identifying details that support an assertion or conclusion
- making connections between key ideas and details

Sample suggested resources for Novice:

1. “Attacking the Common Core Standards” Informational Texts – Part ...
   Mar 18, 2012 ... Cite textual evidence to support analysis of what the text says ... Cite the textual evidence that most strongly supports an analysis…
2. Standard—R.I.6.1 & R.L. 6.1 Cite textual evidence to support ...
   Standard—R.I.6.1 & R.L. 6.1 Cite textual evidence to support analysis of what the text says explicitly as well as inferences drawn from the text. RH6.8.8

0 points Novice or below: The student needs instruction and support in some or all of the following:

- reading grade level text
- determining unknown words
- making connections between key ideas and details
- composing a response to text

Sample suggested resources for Novice or below:

1. Lesson 3: Strategies for Determining the Meaning of a Word
   Will explain that sometimes when I am reading, I can read a word by sounding it out, but I don't understand what it means. I will explain that I can stop and think ...
2. Main Idea and Supporting Details
   The supporting details are the things that describe the main idea. ... The 363 mile canal connected Albany, New York to Lake Erie in Buffalo for the first time.

Item 2: Extended Writing Essay

Given the passage as a source on the topic, the next step is to evaluate the source for the information it does or does not provide about the topic. Instructionally, a teacher might ask students to take notes on the source, author viewpoints, and supporting statements or lack thereof. A writing product can provide the evidence a teacher needs to assess how well a student can critically evaluate a single source. The writing can also provide the teacher with evidence about how well the student writes (language conventions) and articulates his or her critique (topic development). To assess the student’s ability to write critically about the first source, a writing prompt was selected as the next item in the set. This particular item was not newly developed but was selected from a pool of items previously administered. The item was scored with a 4-point holistic rubric evaluating writing opinions or arguments, including criteria for reading, writing, and language usage. The item was scored by both human raters and through AES. As with most essays, the teacher can evaluate students across multiple standards. Similarly, the prompt selected addresses multiple standards with evidence to targets described subsequently.

W.7.1 . Write arguments to support claims with clear reasons and relevant evidence.

- Introduce claim(s), acknowledge alternate or opposing claims, and organize the reasons and evidence logically.

- Support claim(s) with logical reasoning and relevant evidence, using accurate, credible sources and demonstrating an understanding of the topic or text.
- Use words, phrases, and clauses to create cohesion and clarify the relationships among claim(s), reasons, and evidence.
- Establish and maintain a formal style.
- Provide a concluding statement or section that follows from and supports the argument presented.

W7.9. Draw evidence from informational texts to support analysis.

b. Apply Grade 7 Reading standards to literary nonfiction (e.g. “Trace and evaluate the argument and specific claims in a text, assessing whether the reasoning is sound and the evidence is relevant and sufficient to support the claims”).

W7.4. Produce clear and coherent writing in which the development, organization, and style are appropriate to task, purpose, and audience.

L.7.1. Demonstrate command of the conventions of standard English grammar and usage when writing or speaking.

L.7.2. Demonstrate command of the conventions of standard English capitalization, punctuation, and spelling when writing.

L.7.3. Use knowledge of language and its conventions when writing, speaking, reading, or listening. Choose language that expresses ideas precisely and concisely, recognizing and eliminating wordiness and redundancy.

Master Claim: Students can read informational text and write effectively when using and/or analyzing sources.

***Target 1*:** Student can introduce claim(s), acknowledge alternate or opposing claims, and organize the reasons and evidence logically.

***Evidence 1*:** Student is able to introduce claim(s), find or provide alternate claim(s), and organize his or her reasoning for choosing such claim(s).

***Target 2*:** Student is able to support claim(s) with logical reasoning and relevant evidence, using accurate, credible sources and demonstrating an understanding of the topic or text. This is directly related to W.7.9 with RI 7.8.

***Evidence 2*:** Student is able to provide evidence that he or she is able to support the claim in text as a demonstration of his or her understanding of the text and ability to draw evidence from what he or she reads.

***Target 3*:** The student can use words, phrases, and clauses to create cohesion and clarify the relationships among claim(s), reasons, and evidence.

***Evidence 3*:** Produce clear and coherent writing in which the development, organization, and style are appropriate to task, purpose, and audience. For example, the student can demonstrate control of writing throughout the piece as a whole by using transitional phrases and words to connect ideas, such as *first, second,* and *most importantly*, to show cohesion among the ideas and by ending one paragraph with a global statement (number of good theories) and following it with supporting paragraphs. This represents students' ability to demonstrate cohesion between the ideas via transitional devices and that the student has control of his or her writing and the piece as a whole.

Target 4**:** Establish and maintain a formal style of writing.

Evidence 4**:** Student uses formal language as opposed to slang or texting language.

Target 5**:** Provide a concluding statement or section that follows from and supports the argument presented.

Evidence 5**:** Student's writing provides a logical conclusion (part of cohesion with a finisher) such as a restatement of his or her argument.

Target 6**:** Demonstrate command of the conventions of standard English grammar and usage when writing.

Evidence 6**:** Student is able to use grammar and usage correctly with few errors that interfere with meaning.

Target 7**:** Demonstrate command of the conventions of standard English capitalization, punctuation, and spelling when writing.

Evidence 7**:** Student is able to write with few capitalization, punctuation, or spelling errors.

Target 8**:** Use knowledge of language and its conventions when writing, where language choice demonstrates ability to expresses ideas precisely and concisely, recognizing and eliminating wordiness and redundancy.

Evidence 8**:** Student is able to demonstrate word choice precision.

Real-world connections. Item 2 requires greater cognitive demand of the reader, for it calls for evaluation of the degree of evidence that the author provides to support a key claim in a text. The student who can write a clear, complete, and coherent essay and provide evidence that he or she has proficiency at doing so will become the critical reader who can read a political candidate's platform document and evaluate the argument, for example, or the critical thinker who can evaluate the evidence used to support or refute a claim, or the critical writer of an effective letter to the editor.

Technology enhancements. See Item 1

Accessibility matrix. See Tables A4 and A5.

Table A4. Identifying Accessibility Barriers: *Item 2_WP*

	Description of steps	Focal KSAs			Nonfocal KSAs		
Step 1	Read or listen to a passage about the Baghdad Battery	Demonstrate comprehension and draw evidence from reading of grade-level, complex informational texts			Perceive and utilize key word pop-ups [✓ E] [✓ V] H [✓ M]	Perceive and utilize read-aloud feature [✓ E] V [✓ H] M	Perceive text E [✓ V] H M
Step 2	Provide text-based, open-ended response to extended prompt	Identify and provide accurate evidence from passage in response to prompt	Introduce a claim, find or provide alternate claims, and organize reasoning	Provide evidence with accurate, credible source information to support the claim in text	Write/keystroke entry online E [✓ V] H [✓ M]	E V H M	E V H M
		Demonstrate understanding and evaluate the author's arguments by drawing evidence from what has been read or viewed	Use precise word choice	Use formal language instead of slang or texting language			

	Description of steps	Focal KSAs			Nonfocal KSAs		
		Provide a logical conclusion, such as restating the main argument	Use grammar and usage correctly, with few errors that interfere with meaning	Use standard English, capitalization, and punctuation correctly, and spell correctly, with few or no errors that interfere with meaning			
		Demonstrate control of writing throughout the piece as a whole by using transitional phrases and words to connect ideas, such as first, second, and most importantly, to show cohesion among the ideas and by ending one paragraph with a global statement (number of good theories) and following it with supporting paragraphs					

Table A5. Identifying Ways to Enhance Accessibility: *Item 2_WP*

For each step: top row—identify relevant populations (1 per column); middle—specify the difficulty; bottom—suggest solution.					
Step 1	Population →	Visually impaired	Motor impaired	ELL	
	Difficulty →	Screen size, pop-up feature	Scrolling, mouse clicks	Passage terminology	
	Solution→	Read-aloud with key word and button tags, magnification	Key board short-cut accessible	For culturally challenging terms, add pop-up definitions and graphics	
Step 2	Population →	Visually impaired	Motor impaired		
	Difficulty →	Text entry of response	Text entry of response		
	Solution→	Assistive technology, scribe, oral response	Assistive technology, scribe, oral response		

Scoring and Instructional Links. This second item, and extended writing essay, is currently scored with a 4-point rubric, provided subsequently, with the use of an artificial intelligence engine. The original exact agreement rates were lower than typically desired at about 42% exact agreement, 54% exact and adjacent, with a weighted Kappa of 65.5%. The AES engine resulted in improved agreement statics: 57% exact, 92% exact and adjacent, and a weighted Kappa of 81.1%. Each of the remaining items was scored via automated scoring logic, described within each item's description.

Item 2 scoring rubric. The scoring rubric for Item 2 follows:

4 points:

- Response is a well-developed essay that develops and supports an opinion or argument
- Effectively introduces an opinion or claim
- Uses logical, credible, and relevant reasoning and evidence to support opinion or claim
- Uses an organizational strategy to present reasons and relevant evidence
- Acknowledges and counters opposing claims, as appropriate
- Uses precise and purposeful word choice
- Uses words, phrases, and/or clauses that effectively connect and show relationships among ideas
- Uses and maintains an appropriate tone
- Provides a strong concluding statement or section that logically follows from the ideas presented
- Has no errors in usage and conventions that interfere with meaning

3 points:

- Response is a complete essay that develops and supports an opinion or argument
- Clearly introduces an opinion or claim
- Uses reasoning and evidence to support opinion or claim
- Uses an organizational structure to present reasons and relevant evidence
- Attempts to acknowledge and/or counter opposing claims, as appropriate
- Uses clear word choice
- Uses words and/or phrases to connect ideas
- Uses an appropriate tone
- Provides a concluding statement or section that follows from the ideas presented
- Has few, if any, errors in usage and conventions that interfere with meaning

2 points:

- Response is an incomplete or oversimplified essay that develops and supports an opinion or argument
- Attempts to establish an opinion or claim
- Develops, sometimes unevenly, reasons and/or evidence to support opinion or claim
- Attempts to use an organizational structure
- Makes little, if any, attempt to acknowledge or counter opposing claims
- Uses simple language, which sometimes lacks clarity
- Provides a weak concluding statement or section
- May have errors in usage and conventions that interfere with meaning

1 point:

- Response provides evidence of an attempt to write an essay that offers an opinion or argument
- Weakly states or alludes to an opinion or claim
- Has minimal support for opinion or claim
- May be too brief to demonstrate an organizational structure
- Makes no attempt to acknowledge or counter opposing claims
- Uses words that are inappropriate, overly simple, or unclear
- Provides a minimal or no concluding statement or section
- Has errors in usage and conventions that interfere with meaning

0 points:

- Response is completely irrelevant or incorrect, or there is no response

Item 2 instructional links. Additional practice in writing will support and extend understanding for writers of most any level. Suggested instructional links are provided regardless of score point for the writing prompt.

1. So What Do You Think? Writing a Review
   Writing a review of an author's work challenges students to develop their critical thinking skills. It provides an opportunity for students to speak their minds—and to enjoy being heard.
2. Improve Students' Writing Using Online Workshops
   "If tomorrow morning the sky falls...have clouds for breakfast." Cooper Eden's book *If You're Afraid of the Dark, Remember the Night Rainbow* offers unconventional responses to life's challenges, and this lesson's activities encourage students to do the same. Online writing groups provide a forum for constructive peer review.
3. How to Write a Summary - Information, Facts, and Links
   Writing a good summary demonstrates that you clearly understand a text and that you can communicate that understanding to your readers.

Item 3: Constrained Internet Research and Source Evaluation

Once students have familiarity with the text and how best to critically evaluate the initial source, they need to be able to evaluate and compare additional sources on the topic. This is in line with the following Reading CCSS:

> RI.7.7. Compare and contrast a text to an audio, video, or multimedia version of the text, analyzing each medium's portrayal of the subject (e.g., how the delivery of a speech affects the impact of the words).

Therefore, and in relation to standard W.7.6's call for the use of technology, the next item provides the student with a list of hyperlinks, mimicking a Web page of results from an Internet search conducted on the topic. The student is asked to evaluate each "link" in the constrained Internet environment, rating each source for his or her personal opinion (like–dislike) and evaluating each source in terms of its relevance to the topic (or, in Grade 7 terms, *related*) and trustworthiness (*credibility*). The item is directly measuring the following:

Master Claim: Students can engage appropriately in independent inquiry to investigate/research topics, pose questions, and gather and present information.

Subclaim: Students demonstrate comprehension and draw evidence from reading of grade-level, complex informational texts.

Target 1**:** Use technology and the Internet to link to/view sources

Evidence 1**:** Student can navigate an assimilated Internet search to read text and view video clips

Target 2**:** Draw evidence from and evaluate multiple sources, including online sources, for relevance and credibility in supporting the topic under investigation

Evidence 2**:** When provided multiple links/sources, student evaluates and documents the relevancy (relatedness) and credibility (trustworthiness) of each source relative to the given topic.

Real-world connections. The third item in the set is reflective of what many people do daily: conduct an Internet search to gain information on a topic or answer a question and then evaluate the list of offerings for relevancy and credibility. Because of the constraints of assessment, students do not actually conduct an Internet search when responding to this item, but they are provided the simulated results of a search and must access each source and evaluate it as a related, trustworthy source of information, just as every user of Internet searches must do.

Technology enhancements. Audio, video, and search features are provided as enhancements to the item, in addition to a like–dislike voting feature to engage students and mimic current iconic voting tools. In addition, student responses are captured in an interactive grid, similar to a drag-and-drop response space.

Accessibility matrix. See Tables A6 and A7.

Table A6. Identifying Accessibility Barriers: *Item 3_Eval*

	Description of steps	Focal KSAs		Nonfocal KSAs	
Step 1	Read (or hear) Web page of search results, multiple sources presented	Use technology and the Internet to link to/view sources		Perceive text and links, opinion tools, evaluation table	
				E [✓ V] H M	E V H M
Step 2	Click on and review each source through hyperlinks	Student can navigate an assimilated Internet search to read text and view video clips		Perceive textual description of each source as well as a Web page, text, and videos of sources	Navigate Web page, utilizing commonly presented icons to hear text or play video
				E [✓ V] [✓ H] M	E [✓ V] H [✓ M]
Step 3	Provide opinion (like–dislike) of each source	Student can indicate initial evaluation of resource by clicking like–dislike buttons		Perceive and click opinion tool	
				E [✓ V] H [✓ M]	E V H M
Step 4	Evaluate each source by providing mark in grid	When provided multiple possible links/sources, student evaluates the credibility (trustworthiness) and relevancy (relatedness) of each source relative to the given topic	Student completes a grid for each source to represent his or her evaluation results	Perceive and click evaluation grid	
				E [✓ V] H [✓ M]	E V H M

Table A7. Identifying Ways to Enhance Accessibility: *Item 3_Eval*

For each step: top row—identify relevant populations (1 per column); middle—specify the difficulty; bottom—suggest solution.

Step 1	Population →	Visually impaired	Motor impaired	ELL	
	Difficulty →	Screen size, text size, pop-up feature	Scrolling, mouse clicks	Passage terminology	
	Solution→	Read-aloud with key word and button tags, magnification	Key board shortcut accessible	For culturally challenging terms, add pop-up definitions and graphics (note that sources cover same topic, so pop-ups from original source are adequate)	
Step 2	Population →	Visually impaired	Hearing impaired	ELL	Motor impaired
	Difficulty →	Page navigation, linking to sources, starting and watching videos	Hearing videos	Video terminology	Page navigation, linking to sources, starting videos
	Solution→	Screen reader, read-aloud/audio only of videos, limited to 2 videos without graphically intense footage, hot keys for navigability	Script provided and/or closed captioning could be provided	Video reviewed and challenging terminology covered in passage pop-ups	Hot keys, assistive devices
Step 3	Population →	Visually impaired	Motor impaired		
	Difficulty →	See opinion tool, click on like–dislike	Click on like–dislike buttons		
	Solution→	Screen reader/magnification, hot keys, scribe	Hot keys/AT, scribe		
Step 4	Population →	Visually impaired	Motor impaired		
	Difficulty →	See evaluation grid, click evaluation within grid	Click evaluation within grid		
	Solution→	Screen reader/magnification, hot keys, scribe	Hot keys/AT, scribe		

Scoring and Instructional Links. This scoring and instructional links for Item 3 follow.

Item 3 scoring rubric. The scoring rubric for Item 3 follows:

4 points:
Student correctly determined whether **all** four sources are related and trustworthy.

3 points:
Student correctly determined whether **three** of the four sources are related and trustworthy.

2 points:
Student correctly determined whether **two** of the four sources are related and trustworthy.

1 point:
Student correctly determined whether **one** of the four sources is related and trustworthy.

0 points:
Student's response does not show that he or she can correctly evaluate whether each source is related or trustworthy.

Item 3 instructional links. Samples of suggested instructional links for each score point in Item 3 follow:

4 points Expert/Master: The student shows readiness to

- evaluate texts at a high range of complexity
- conduct Internet searches effectively
- compare multiple sources in real or simulated environments

3 points Master: The student shows readiness to

- evaluate texts of a range of complexity
- conduct frequent Internet searches in order to use evaluating skills
- compare multiple sources in real or simulated environments

Sample suggested resources for Expert/Master:

1. Source Credibility – Evaluating The Reliability of a Source
   Source Credibility – Evaluating The Reliability of a Source. Not every source is suitable for use in a formal research paper, and the ultimate guide of what is

2 points Novice: The student needs additional help

- understanding characteristics of reliable and credible sources
- evaluating sources with increasing complexity
- developing note taking skills while watching or listening to media clips
- highlighting key ideas and details when reading

Sample suggested resources for Novice:

1. Helpful Hints to Help You Evaluate the Credibility of Web Resources
   Developing a keen sense of the credibility of sources, based on such clues as ... with various types of Web resources and the reliability of the information. 1.
2. How to Judge the Reliability of Internet Information
   However, judging the reliability of sources found on the Internet is crucial ... worried that the information may lack credibility, try starting with a source you know is ...

1 point Novice and below: The student needs additional help

- understanding characteristics of reliable and credible sources
- evaluating sources with increasing complexity
- developing note taking skills while watching or listening to media clips
- highlighting key ideas and details when reading

Sample suggested resources for Novice and below:

1. Helpful Hints to Help You Evaluate the Credibility of Web Resources
   Developing a keen sense of the credibility of sources, based on such clues as ... with various types of Web resources and the reliability of the information. 1.
2. How to Judge the Reliability of Internet Information
   However, judging the reliability of sources found on the Internet is crucial ... worried that the information may lack credibility, try starting with a source you know is ...
3. How to Take Great Notes - Study skills | GreatSchools
   Aug 31, 2006 ... Taking good notes requires students to evaluate, organize and ... It's a key survival skill your child will need through high school and beyond.
4. Marzano's Instructional Strategies
   Integrating Technology into the Classroom using Instructional Strategies based on the research from: ... Summarizing and Note Taking, Setting Objectives and Providing Feedback. Reinforcing Effort and ... High School Ace · Brain Pop Movies ...
5. Note-Taking Tips
   Carlos and Cecilia were both straight-A students in middle school. ... Note-taking is a skill that can help you do well on all your schoolwork.

0 points below Novice: The student needs instruction and support in some or all of the following:

- reading , viewing, and listening to a variety of sources
- identifying how one source does or does not relate to another source
- understanding characteristics of a reliable source
- understanding what makes a source credible or trustworthy
- developing note taking skills while watching or listening to media clips
- highlighting key ideas and details when reading

Sample suggested resources for below Novice:

1. Helpful Hints to Help You Evaluate the Credibility of Web Resources
   Developing a keen sense of the credibility of sources, based on such clues as ... with various types of Web resources and the reliability of the information.
2. How to Judge the Reliability of Internet Information
   However, judging the reliability of sources found on the Internet is crucial ... worried that the information may lack credibility, try starting with a source you know is ...
3. How to Take Great Notes - Study skills | GreatSchools
   Taking good notes requires students to evaluate, organize and ... It's a key survival skill your child will need through high school and beyond.
4. Note-Taking Tips
   Carlos and Cecilia were both straight-A students in middle school. ... Note-taking is a skill that can help you do well on all your schoolwork.

Item 4: Source Comparisons

Now that the student has been given and has evaluated a related and trustworthy source (the passage), and then has been asked to evaluate additional sources, the student compares sources one to another. In keeping with the same standard RI.7.7 and claims as the previous item, an additional target and related evidence specifically draw out the student's ability to compare and contrast the sources.

> ***Target 3***: Students identify similarities and differences between and among sources on the same topic.
>
> ***Evidence 3***: Student is able to select and use an organizer, one of multiple response methods to illustrate comparisons across sources, for example, through Venn diagrams, table or chart form, or in outline form. The student provides multiple similarities and differences from sources, not including student opinion, and does not confuse similarities with differences within and across the sources.

Real-world connections. Internet users and researchers routinely compare the information found in multiple resources. For this fourth item, students compare information from a multimedia resource and a prose resource, noting similarities and differences. Furthermore, in the world outside of a classroom and assessment, people make choices regarding how they wish to record comparisons or any kind of information. Some people may record accumulated information in a PowerPoint deck, on a smart phone notepad, in spreadsheets or graphic representations, or by hand on paper. For this item, students choose from three options and thus demonstrate a similar degree of autonomy and personal choice.

Technology enhancements. For this item, we allow students to choose the method by which to represent their comparisons, as there are multiple methods taught (which may reflect students' opportunities to learn the various organizers) and typically used in college and careers. Students compare two sources: the initial informational text and a newly provided, video-based source. Providing a unique source rather than one of the earlier linked sources ensures some independence between the two items so as not to clue the previous item. The item was built to allow students to review the formats from which they can choose (Venn diagram, table, presentation slides) prior to making their selection and to view both the informational text and video within the same online space. The text and video have been segmented so that students can readily and easily make their selections and placements (drag and drop) of text and video clips into their chosen organizer.

Accessibility matrix. See Tables A8 and A9.

Table A8. Identifying Accessibility Barriers: ***Item 4_Organizers***

	Description of steps	Focal KSAs		Nonfocal KSAs		
Step 1	Read or listen to a passage about the Baghdad Battery	Read or listen to and comprehend informational text		Perceive and utilize key word pop-ups [✓ E] [✓ V] H [✓ M]	Perceive and utilize read-aloud feature [✓ E] V [✓ H] M	Perceive text E [✓ V] H M
Step 2	Watch a video about the Baghdad Battery	Watch, listen to, and comprehend video describing topic		Perceive video E [✓ V] H M	Hear video [✓ E] V [✓ H] M	E V H M
Step 3	Review and select response method	Recognize and select organizer to show comparison between text and video: Venn diagram, table chart, outline		Determine preferred method for providing similarities and differences E V H M	Perceive tools E [✓ V] H M	E V H M
Step 4	Identify similarities and differences between sources	Identify and cite minimum number of the differences and similarities of sources, with specificity, without introducing own opinion	Not confuse similarities with differences	E V H M	E V H M	E V H M
Step 5	Document similarities and differences between sources	Categorize elements of sources as similar or different	Uses technology—mouse—to drag elements to appropriate places in selected organizer	Capture and place parts of passage E [✓ V] H [✓ M]	Capture and place parts of video E [✓ V] H [✓ M]	E V H M

Table A9. Identifying Ways to Enhance Accessibility: ***Item 4_Organizers***

For each step: top row—identify relevant populations (1 per column); middle—specify the difficulty; bottom—suggest solution.

Step 1	Population →	Visually impaired	Motor impaired	ELL
	Difficulty →	Screen size, text size, pop-up feature	Scrolling, mouse clicks	Passage terminology
	Solution→	Read-aloud with key word and button tags, magnification	Key board short-cut accessible	For culturally challenging terms, add pop-up definitions and graphics
Step 2	Population →	Visually impaired	Hearing impaired	ELL
	Difficulty →	Watching the video	Hearing the video	Video terminology
	Solution→	Read-aloud/audio only; this video does not contain graphically intense footage	Closed captioning could be provided	Video reviewed and challenging terminology covered in passage pop-ups
Step 3	Population →	Visually impaired		
	Difficulty →	Seeing the Venn diagram, outline, chart		
	Solution→	Enlargement on-screen		
Step 4	Population →			
	Difficulty →			
	Solution→			
Step 5	Population →	Visually impaired	Motor impaired	
	Difficulty →	Drag and drop elements	Drag and drop elements	
	Solution→	Enlargement, assistive technology, scribe, oral response	Key board shortcut accessible, assistive technology, scribe, oral response	

Item 4 scoring rubric. The scoring rubric for Item 4 follows:

3 points:
The student has conducted a **thorough** comparison of the text and the video, correctly and completely showing which information is in both sources and which is in only one source. The student has correctly placed **all** pieces of the article and video in your organizer.

2 points:
The student has conducted a **partial** comparison of the text and the video by identifying some information that is in both sources and some that is in only one source. The student has correctly placed at least half of the pieces of the article and video in your organizer.

1 point:
The student has conducted a **minimal** comparison of the text and the video by identifying a minimal number of pieces of information that is in both sources and some that is in only one source. The student has correctly placed at least one-fourth of the pieces of the article and video in your organizer.

0 points:
The student has conducted an **incomplete** comparison of the text and the video. The information the student identified is not enough, or it is inaccurate and/or misplaced.

Item 4 instructional links. Samples of suggested instructional links for each score point in Item 4 follow:

3 points Expert: The student shows readiness to

- Compare more than two sources, and across source types
- Extend the level of detail in comparisons
- Organize the details with higher level of granularity

2 points Master: The student shows readiness to

- Compare more than two sources, and across source types
- Extend the level of detail in comparisons

Sample suggested resources for Expert and Master:

1. Venn Diagram, 3 Circles
   This interactive tool allows students to create Venn Diagrams that contain three overlapping circles, enabling them to organize their information logically.
2. Compare and Contrast Electronic Text With Traditionally Printed Text
   Students become familiar with the similarities and differences between electronic and printed text by comparing the textual aids included in a textbook with those of an educational website.

1 point Novice: The student may need assistance to improve understanding about and methods for making strong comparisons and use of organizational tools; and the student needs instruction and support in some or all of the following:

- Identifying similarities and differences between sources across types
- Using a variety of graphic organizers to represent comparisons

Sample suggested resources for Novice:

1. Venn Diagram, 2 Circles
   This interactive tool allows students to create Venn Diagrams that contain two overlapping circles, enabling them to organize their information logically.
2. Venn Diagram Rubric
   The Venn Diagram Rubric can be used to provide students feedback on students' use of a graphic organizer to compare and contrast two things.
3. Compare and Contrast Electronic Text With Traditionally Printed Text
   Students become familiar with the similarities and differences between electronic and printed text by comparing the textual aids included in a textbook with those of an educational website.
4. Comparing and Contrasting: Picturing an Organizational Pattern
   Using picture books as mentor texts, students learn effective strategies for organizing information that compares and contrasts. Students can then apply appropriate organizational strategies to their own papers.

0 points Novice or below: The student needs instruction and support on

- Making comparisons and identifying contrasts
- Using organizational tools

1. Comparing and Contrasting: Picturing an Organizational Pattern
   Using picture books as mentor texts, students learn effective strategies for organizing information that compares and contrasts. Students can then apply appropriate organizational strategies to their own papers.
2. Compare and Contrast Chart
   This organizer can be used to help students explain similarities and differences between two things or ideas. After this organizer has been completed, it could easily be developed into a classroom discussion or writing topic on the information gathered.
3. Compare & Contrast Map
   The Compare & Contrast Map is an interactive graphic organizer that enables students to organize and outline their ideas for different kinds of comparison essays.

Item 5: Future Research

As students review various sources, they likely begin to form opinions about the sources and to ask their own questions related to the topic. The following writing standard, claims, and evidences set the stage for the next item.

> W.7.7. Conduct short research projects to answer a question, drawing on several sources and generating additional related, focused questions for further research and investigation.
>
> **Master Claim:** Students can engage appropriately in independent inquiry to investigate/research topics, pose questions, and gather and present information.
>
> ***Target 1*****:** Given sources reviewed on topic, student can recommend additional areas of research and provide an explanation to justify their recommendations and desire to investigate more.
>
> ***Evidence 1*****:** Student is able write additional research questions and explain why the questions are relevant to the topic and their desire to know more.

Real-world connections. Finally, item 5 is reflective of what a thorough researcher will do: think about and identify additional information needed. This parallels the thinking a smart consumer will apply, for example, before purchasing a product. For item 5, students will demonstrate proficiency at thinking beyond the information already gathered to the information that remains to be found. To demonstrate the knowledge, skills, and abilities as evidence, the students will be asked to reflect on all the sources they have encountered in this item set, to identify two additional research questions, and to provide reasoning for their interest in wanting to know more.

Technology enhancements. The technology features used to enhance this item include the audio version of the passage and the video with related transcription. The student is required to type in his or her text response, which can also be heard via audio once it is typewritten.

Accessibility matrix. See Tables A10 and A11.

Table A10. Identifying Accessibility Barriers: *Item 5_ResTable*

	Description of steps	Focal KSAs	Nonfocal KSAs		
Step 1	Recall or read or listen to a passage about the Baghdad Battery	Read or listen to and comprehend informational text	Perceive and utilize key word pop-ups [✓ E] [✓ V] H [✓ M]	Perceive and utilize read-aloud feature [✓ E] V [✓ H] M	Perceive text E [✓ V] H M
Step 2	Recall or watch a video about the Baghdad Battery	Watch, listen, and comprehend video describing topic	Perceive video E [✓ V] H M	Hear video [✓ E] V [✓ H] M	E V H M
Step 3	Identify and document additional areas of research, in open-ended table format	List two additional research questions and explain why questions are appropriate, answers are needed.	Perceive response area/table E [✓ V] H M	Write/keystroke entry online E [✓ V] H [✓ M]	E V H M

Table A11. Identifying Ways to Enhance Accessibility: ***Item 5_ResTable***

For each step: top row—identify relevant populations (1 per column); middle—specify the difficulty; bottom—suggest solution.

Step 1	Population →	Visually impaired	Motor impaired	ELL	
	Difficulty →	Screen size, text size, pop-up feature	Scrolling, mouse clicks	Passage terminology	
	Solution→	Read-aloud with key word and button tags, magnification	Key board shortcut accessible	For culturally challenging terms, add pop-up definitions and graphics	
Step 2	Population →	Visually impaired	Hearing impaired	ELL	
	Difficulty →	Watching the video	Hearing the video	Video terminology	
	Solution→	Read-aloud/audio only; this video does not contain graphically intense footage	Closed captioning could be provided	Video reviewed and challenging terminology covered in passage pop-ups	
Step 3	Population →	Visually impaired	Motor impaired		
	Difficulty →	Seeing table and response space, text entry of response	Text entry of response		
	Solution→	Enlargement/magnification, assistive technology, scribe, oral response	Assistive technology, scribe, oral response		

Item 5 scoring rubric. The scoring rubric for Item 5 follows:

2 points
You have completed the table. Your two questions are relevant, and you provide clear, reasonable support for each question; your questions are different than other questions you have already answered for this task.

1 point:
You have partially completed the table. You provide only one unique question with explanation, OR your two questions are not explained, OR your questions are nearly the same.

0 points:
Your table is inaccurate or has not been completed. If there are entries in the table, they are irrelevant, inaccurate, or too similar to questions you have already answered for this task.

Item 5 instructional links. Samples of suggested instructional links for each score point in Item 5 follow:

2 points Expert/Master: The student shows readiness to

- Conduct extensive research across media forms
- Extend and communicate research across media forms

Sample suggested resources for Expert/Master:

1. Wading Through the Web: Teaching Internet Research Strategies
   The Internet is often the first stop for student researchers, but few know how to vet sources effectively. In this lesson, students learn about Internet research through an interactive PowerPoint presentation.
2. Creating a Persuasive Podcast
   Students learn how to get their voice out on the web when they research issues important to them and compose a persuasive podcast to post online.

1 point Novice: The student needs instruction and support on considering additional areas of research, and may benefit from increased exposure to research applications.

Sample suggested resources for Novice:

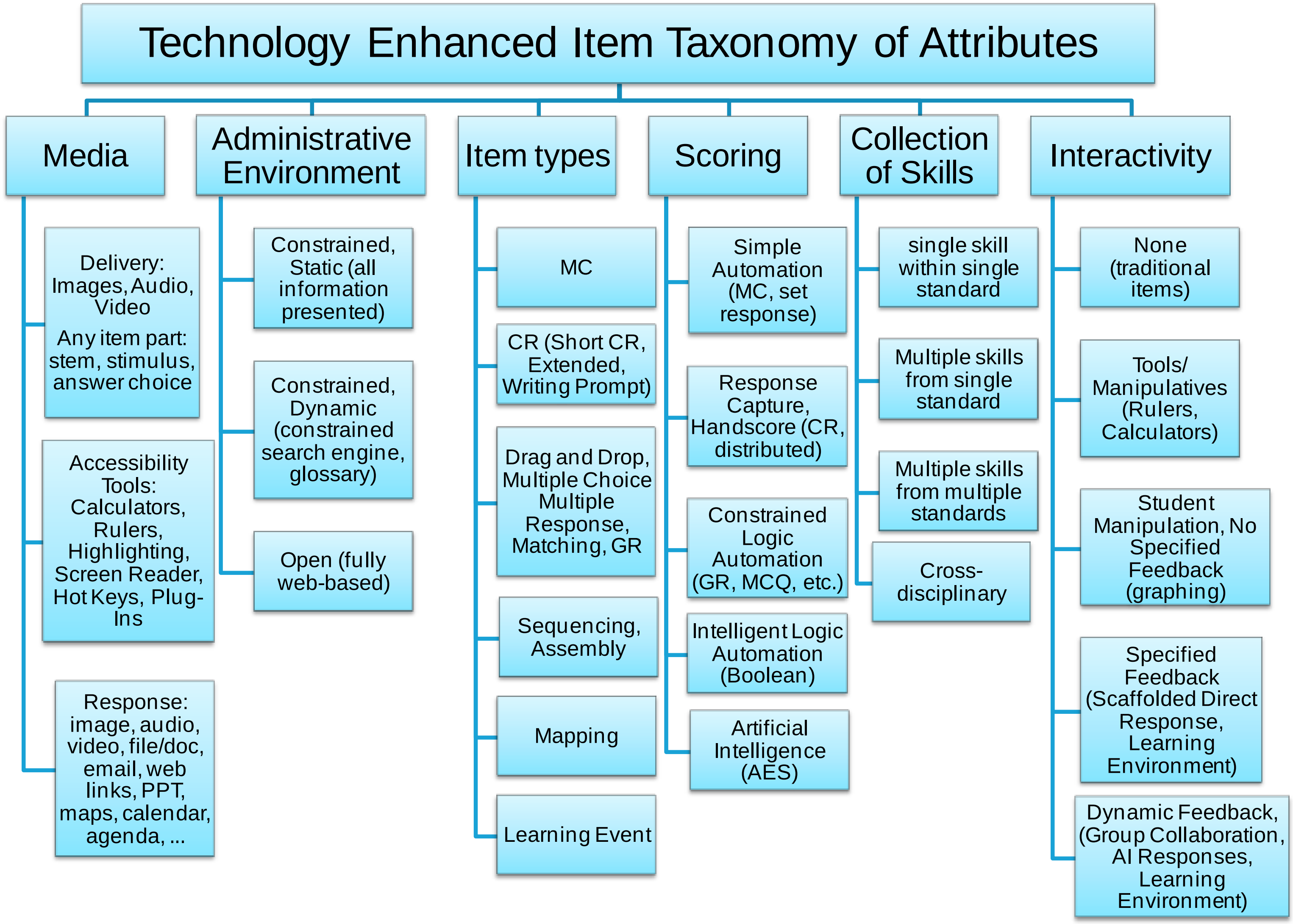
1. Inquiry on the Internet: Evaluating Web Pages for a Class Collection
   Students use Internet search engines and Web analysis checklists to evaluate online resources then write annotations that explain how and why the resources will be valuable to the class.
2. Biography Project: Research and Class Presentation
   Classroom biography study offers high-interest reading with a purpose, as students begin with inquiry and research, summarize and organize their information, and prepare oral presentations to share with the class.

0 points Novice and below: The student needs instruction and support on considering additional areas of research, and may benefit from increased exposure to research applications.

Sample suggested resources for Novice and below:

1. Scaffolding Methods for Research Paper Writing
   In this lesson, students use a scaffold to help them compile information to write a solid research paper.
2. Picture Books as Framing Texts: Research Paper Strategies for Struggling Writers
   Students use picture books as framing texts for research, freeing them from the language of encyclopedia sources and allowing them to focus their attention on the content of their papers.
3. Hughes Middle School: Writing Research Questions
   Writing Research Questions - When doing research, it is very important to write very clear research questions.
4. Writing Research Questions « Research Rundowns
   Writing Research Questions .pdf version of this page. This review is a collection of views and advice on composing research questions from problem statements.

Appendix B: Technology Enhanced Item Taxonomy of Attributes



Technology Enhanced Item Taxonomy of Attributes

- Media
  - Delivery: Images, Audio, Video
    Any item part: stem, stimulus, answer choice
  - Accessibility Tools: Calculators, Rulers, Highlighting, Screen Reader, Hot Keys, Plug-Ins
  - Response: image, audio, video, file/doc, email, web links, PPT, maps, calendar, agenda, ...
- Administrative Environment
  - Constrained, Static (all information presented)
  - Constrained, Dynamic (constrained search engine, glossary)
  - Open (fully web-based)
- Item types
  - MC
  - CR (Short CR, Extended, Writing Prompt)
  - Drag and Drop, Multiple Choice Multiple Response, Matching, GR
  - Sequencing, Assembly
  - Mapping
  - Learning Event
- Scoring
  - Simple Automation (MC, set response)
  - Response Capture, Handscore (CR, distributed)
  - Constrained Logic Automation (GR, MCQ, etc.)
  - Intelligent Logic Automation (Boolean)
  - Artificial Intelligence (AES)
- Collection of Skills
  - single skill within single standard
  - Multiple skills from single standard
  - Multiple skills from multiple standards
  - Cross-disciplinary
- Interactivity
  - None (traditional items)
  - Tools/ Manipulatives (Rulers, Calculators)
  - Student Manipulation, No Specified Feedback (graphing)
  - Specified Feedback (Scaffolded Direct Response, Learning Environment)
  - Dynamic Feedback, (Group Collaboration, AI Responses, Learning Environment)

K-12 Center at ETS

Invitational Research Symposium on

Technology Enhanced Assessments

The Center for K–12 Assessment & Performance Management at ETS creates timely events where conversations regarding new assessment challenges can take place, and publishes and disseminates the best thinking and research on the range of measurement issues facing national, state and local decision makers.