Item Development and Assessment Construction Guidelines for Common Core State Standards

by
Scott A. Cunningham, B.S.
Sarah M. Callahan, Ph.D.
Jason K. Feld, Ph.D.
and John Richard Bergan, Ph.D.

Assessment
Technology
Incorporated

Item Development and Assessment Construction Guidelines for Common Core State Standards

By Scott A. Cunningham, B.S., Sarah M. Callahan, Ph.D.,
Jason K. Feld, Ph.D., and John Richard Bergan, Ph.D.
Assessment Technology, Incorporated

Table of Contents

I. Introduction

This document details the design, review, and construction procedures used by Assessment Technology Incorporated (ATI) to develop standards aligned selected-response, technology-enhanced, and constructed-response items and assessments aligned to Common Core State Standards (CCSS) and Next Generation Science Standards (NGSS)*. ATI procedures are informed in part by released materials from the Partnership for Assessment of Readiness of College and Careers (PARCC) and the Smarter Balanced Assessment Consortium (SBAC). These organizations have provided guidelines and sample items that can be used to guide the development and implementation of local assessment and instruction systems designed to align with CCSS and to promote student learning.

A. Changes Introduced by Common Core State Standards and New Science Standards Related to Item and Assessment Development

CCSS and new science standards differ from earlier state standards in a variety of ways that directly affect the development of items and assessments reflecting the new standards. First, the new standards promote the **integration of assessment and instruction**. As a consequence, assessment is no longer limited to snapshots of information collected within a brief clearly defined point in time. Rather, **assessment may occur over an extended time** period during the instructional process. For example, assessment in science might occur throughout a research project involving a sequence of steps such as hypothesis formulation, observation, data collection, data analysis, and reporting results.

Second, the new standards emphasize the development of higher order thinking skills reflecting **high Depth of Knowledge (DOK) levels**. The increased importance of high DOK levels has subject-specific effects on instruction. In the area of English language arts, there is a new emphasis on **text complexity** (Krehbiel, 2012). Students are given the opportunity to learn from content-rich informational texts that require **close reading and detailed analysis**. They are expected to conduct research, to learn and use academic vocabulary, and to craft arguments based on evidence. In the area of mathematics, there is increased focus on understanding mathematical operations as well as fluently implementing mathematical operations (Briars, 2012). For example, the longstanding practice of memorizing a formula and implementing it to solve a problem may be replaced by learning involving proofs showing why the formula produces the desired result. Thus, application of the formula is supported by **understanding how the formula works**. In both science and mathematics, there is increased emphasis on the use of **physical models to inform mental models** of processes of interest. For example, in the biological sciences, physical models of a food chain may be used to inform the student's mental representation of the chain.

Third, the new standards replace the heavy reliance on selected-response items characteristic of earlier assessments with a more balanced approach involving **selected-response items, technology-enhanced items, and constructed-response items**. This balanced approach results in an **expansion of item types** essential for assessment. Selected-response items continue to be used to assess capabilities reflecting low and moderate DOK levels. Technology-enhanced items offer new ways to measure student achievement that may more closely reflect the cognitive capability of interest. Technology enhanced items have the additional time- and labor-saving advantages associated with automated item scoring.

*Next Generation Science Standards (NGSS) is a registered trademark of Achieve. Neither Achieve nor the lead states and partners that developed the NGSS was involved in the production of, and does not endorse, this product.

Constructed-response items are preferred for capabilities involving high DOK levels because complex thinking processes are best measured by tasks that require the student to produce an extended detailed response. The benefits of constructed-response items are likely to increase as automated scoring programs become more advanced and more widely available. At present, manual scoring using rubrics is the procedure of choice.

Finally, the new standards have introduced changes in the assessment process as well as changes in item and assessment content. Whereas in the past, standards-based assessments have been limited for the most part to fixed-length criterion-referenced measures, there is currently a plan to include **computerized adaptive tests** (CAT) in the assessments developed by SBAC for CCSS. CATs require **different test-taking skills** than fixed-length tests. In a fixed-length test, the examinee typically has the opportunity to review previously answered questions and to make changes in their answers to those items. However, in a CAT, there are restrictions on the availability of previously presented items. CATs are also best administered online. Districts/charter schools that do not currently possess the technology to support **online testing** would need to increase their technology readiness before attempting to introduce CATs. Students will also need to further develop test-taking skills and strategies specific to online test-taking. CATs also impose requirements on test analysis and test scoring. More specifically, the use of **Item Response Theory** (IRT) is essential to the effective implementation of the CAT approach. In a CAT, the selection of items of appropriate difficulty to be presented to the examinee is based on estimates of examinee ability and item difficulty made using IRT.

B. ATI Support for the Transition to Common Core State Standards and New Science Standards

Implementation of CCSS and new science standards calls for major changes in item development and test construction. With that said, the new standards do not ignore the accomplishments of the past. Rather, they build on those accomplishments. Providing a link to earlier assessment practices and outcomes is beneficial in a number of ways. First, there are important skills that can be efficiently assessed by existing procedures. Second, the link to the past supports a transition period wherein districts/charter schools are able to gradually introduce new approaches to assessment while maintaining previous approaches that have proved useful and with which district/charter school educators and students are familiar. Third, the link makes it possible to compare assessment results over time. Comparisons across time provide a continuing record of changes in educational achievement.

During the last several years, ATI has developed and implemented a plan to support the transition to CCSS in a manner that preserves the link to the past and provides a path to the future. The central elements of the plan are as follows:

- ATI has developed **online *Instructional Dialogs* supporting the integration of assessment and instruction** and providing real-time information on student learning. *Instructional Dialog* technology provides assessment and instruction within one interface. This approach supports brief assessments as well as extended projects involving assessments with high DOK levels that cannot be completed within the limited time frames available for the typical assessment. In addition, *Instructional Dialog* technology is ideally suited to provide students with experience and guidance related to the introduction of technology-enhanced items with which the students may be unfamiliar.
- ATI has **constructed thousands of selected-response items aligned to the** new **standards**. In addition, where appropriate, **items aligned previously to state**

standards have been mapped to the new standards. These initiatives have made it possible for districts/charter schools to align their local assessments to state standards being addressed on current statewide tests while also initiating the transition to the new standards.

- ATI has **developed a series of new technology-enhanced item types** and online tools reflecting the movement toward online assessment planned by PARCC and SBAC. ATI has also **expanded the number of constructed-response items** in the item banks. These item types and tools are designed to assess a broad range of capabilities including the higher order thinking skills emphasized in the new standards. The development of technology-enhanced and constructed-response items is a continuing project. Thus, item development by the ATI Assessment and Instructional Design Department occurs on a continuous basis.

- ATI has **implemented IRT techniques to score selected-response, technology-enhanced, and constructed-response items**. IRT techniques are also used to place new test scores on a common scale with previous test scores. This makes it possible to compare assessment scores over time.

- ATI has supported the increased use of CATs called for by SBAC by introducing **a multi-stage *Computerized Adaptive Testing* option**. This option provides greater control over assessment content than is available with a traditional CAT. Content control is important in standards-based education because it makes it possible to ensure that what is tested is what has been taught.

- ATI has designed **professional development offerings to assist districts/charter schools to make the transition to CCSS**. Training options include on-site professional development sessions, online webinars, and a series of short online webisodes dealing with specific topics critical to CCSS implementation.

C. Goals of Assessment and Requirements for Item and Assessment Development

Assessment is a central component of the Galileo® K-12 Online Instructional Improvement and Instructional Effectiveness System. The purpose of both the instructional improvement and instructional effectiveness units in the system is to provide tools that can assist educators to promote student learning. ATI has developed a *comprehensive assessment system* that serves assessment needs for both the instructional improvement and instructional effectiveness units of the system. The *comprehensive assessment system* includes benchmark assessments, formative assessments, pretests and posttests, course examinations, placement tests, and other forms of assessment. The assessment component is intended to inform instruction aimed at promoting the mastery of standards reflecting valued educational goals.

During the last year, many districts and charter schools have begun transitioning the standards selected to guide instruction to CCSS. In order for the Galileo assessment component to be useful in informing instruction reflecting CCSS, it must:

- adhere to industry standards for item and test construction;
- provide information on the mastery of standards;
- provide recommendations as to what to teach next;
- be capable of forecasting standards mastery as reflected in statewide test performance;
- be capable of reflecting the full range of variations in knowledge levels reflected in the standards guiding instruction

- be sensitive to diverse background characteristics of students that may affect student performance;
- adhere to universal design specifications for accommodating English Language Learners (ELL) and students with special needs;
- provide local control of test specifications that yields efficient measurement of standards targeted for instruction;
- reflect standards controlling the visual quality of items in a manner that accommodates both online administration and offline administration using test booklets; and
- estimate and evaluate item parameters through a continuing research program. IRT provides the generally accepted procedures for conducting research on item parameters. IRT also makes it possible to place scores from multiple assessments on a common scale facilitating the measurement of progress. In addition, IRT can be used to inform recommendations regarding what to teach next to promote learning.

The item development process for comprehensive assessments in Galileo® has been carefully designed to produce high-quality items including selected-response, technology-enhanced, constructed-response items that adhere to the conditions specified above. This process includes the development of item specifications, item construction, and item review with certification. The rationale for the adopted procedures is informed by: 1) industry standards outlined by the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education in *The Standards for Educational Testing*, 1999; 2) the unique characteristics and purposes of the many types of assessments included in the ATI *comprehensive assessment system*; and 3) a continuing research program that examines the psychometric properties of items used in ATI assessments.

The discussion that follows outlines the item development process. In addition, the item types utilized in assessments are summarized and the importance of local control of test specifications is reviewed. The ability of the assessment component to accommodate ELL and students with special needs is addressed. The discussion includes a brief outline of the design of assessments that are effective in forecasting performance on statewide assessments and also useful for guiding instruction.

II. Item Specifications

When new items are to be added to the ATI item banks, the first step is to review the standards which are to be assessed. Each standard is broken down into the skills that make up the standard. These skills are the starting point for developing an online list of item specifications defining the characteristics of the particular class of items to be written. **Item specifications indicate the defining characteristics of the item class, the rationale for the class, and the required characteristics for each item component**. Specifications might also address such factors as the cognitive complexity intended for items included in the specification class, the appropriateness of vocabulary, and Lexile® measure requirements related to readability levels. The value of creating specifications as a guide for the item development process is recognized as a critical part of a process documenting that assessments are reliable and valid indicators of the ability they are intended to measure (Haladyna, 2004). Their structure and specificity also afford many advantages for ensuring that assessments may be readily adapted as district/charter school needs and or state/federal requirements change.

Item specifications may be **designed to align very closely to guidelines and released items from assessment consortia (i.e., PARCC, SBAC)**. The careful design of item

specifications is especially critical to the development of technology-enhanced and constructed-response items similar to those students are likely to encounter on future statewide assessments. For example, the following specification describes the characteristics of an interdependent item set that requires the student to draw a conclusion and cite textual evidence in support of their analysis. As described in the specification, the item set consists of two items aligned to separate Common Core State Standards that collectively represent a task with a depth of knowledge level of three. As indicated in the following screen shots, this item specification is designed to produce item sets that mirror one of the released PARCC prototype items. To better illustrate the student's experience, the sample ATI item set is presented as viewed in the online student testing interface.

Sample item specification for an interdependent item set in seventh grade English language arts –
Confidential Screen Shot

Description: Student must choose a description of a character and then cite evidence to support the description.

Released Items:
similar to 6th grade PARCC prototypes

DoK Justification: Drawing conclusions and citing evidence are DoK 3 tasks.

Stimulus Attributes: grade-appropriate fiction

Stem Attributes:
[RL.7.2] sub item 1: Which word best describes [character name]?
[RL.7.1] sub item 2: Which quotation best supports your answer to Part 1?

Response Attributes:
sub item 1: four words or phrases
sub item 2: quotations from the story, most likely dialogue or action; the CA should be the clearest example

Depth of Knowledge: 3   Spec Question Type: Not Set

Scale Name	Performance Objective
CC-R07: 7th Grade English Language Arts	CC-RL.7.1 Key Ideas and Details: Cite several pieces of textual evidence to support analysis of what the text says explicitly as well as inferences drawn from the text.

Sample ATI interdependent item set created using the preceding item specification (as viewed in the online student testing interface) – Confidential Screen Shots

Item Development and Assessment Construction
Guidelines for Common Core State Standards
Assessment Technology, Incorporated

1.800.367.4762
ati-online.com

Corresponding released PARCC prototype English language arts item

SAMPLE ITEM

Part A

What does the word "regal" mean as it is used in the passage?

☐ a. generous ☐ b. threatening

☐ c. kingly ☐ d. uninterested

Part B

Which of the phrases from the passage best helps the reader understand the meaning of "regal?"

☐ a. "wagging their tails as they awoke" ☐ b. "the wolves, who were shy"

☐ c. "their sounds and movements expressed goodwill" ☐ d. "with his head high and his chest out"

Passage

George, Jean C. Julie of the Wolves. New York: Harper and Row, 1972. Print.

PARCC is committed to using authentic texts. Permissions are pending for the texts associated with this item.

Similarly, the sample specification in the following screen shot defines the characteristics of a constructed-response math item for eighth-grade students. The specification for this item includes the rationale for the item, the objective the item is designed to assess, the attributes of the stimulus and stem presented to students, the standard(s) to which the item is aligned, the scoring rule for the item, a prototype item, the required DOK level, and the justification for this DOK level. This item specification defines a class of items that are highly similar to a released practice item from SBAC.

Sample item specification for eighth grade math – Confidential Screen Shot

Item Type: Constructed Response

Rationale: This set of items is designed to assess Common Core standard 8.EE.5. In addition to making accurate calculations, students must apply their understanding of ratios and unit rates while explaining their answer to a real-world problem. For this reason, the spec is classified as DOK level 2. (Note: Items in this spec may *approach* DOK 3, but are classified as DOK 2 because of the guidance offered in some of the directions, as in part (a) of the prototype.)

Objective: Students will be able to solve problems by making accurate comparisons between ratios relating price to volume.

Stimulus Attributes: presentation of two 3-dimensional figures (in context); images showing the dimensions of the figures and the prices associated with them.

Stem Attributes:

(a) Students are asked to explain, using unit rates, whether one of the images represents a better value than the other.

(b) Students are asked to determine a new price for one of the objects pictured that would make both unit rates equal.

Scoring Rules:

Part a: Student will use unit rates to explain whether the image identified represents a better value than the other.

Part b: Student will correctly identify the price change that makes the unit rates of both objects equal.

Record the score: 2 points possible, 1 point per part answered correctly.

Prototype Question

An ice cream company distributes its product in two different container sizes. The prices for each container are as shown below.

Complete the following.

a. Is the large container a better value? (Use the prices per cubic inch for both containers to explain your answer.)

b. The ice cream company decides to change the price of the large container so that the prices per cubic inch are equal. What should be the new price for each large container?

Small Container: Large Container:

    $3.60 $9.60

Depth of Knowledge: 2   Spec Question Type: Not Set

Scale Name	Performance Objective
CA CC-M08: 8th Grade Math	CA-CC.8.EE.5 Graph proportional relationships, interpreting the unit rate as the slope of the graph. Compare two different proportional relationships represented in different ways. For example, compare a distance-time graph to a distance-time equation to determine which of two moving objects has greater speed. [From cluster: Understand the connections between proportional relationships, lines, and linear equations]
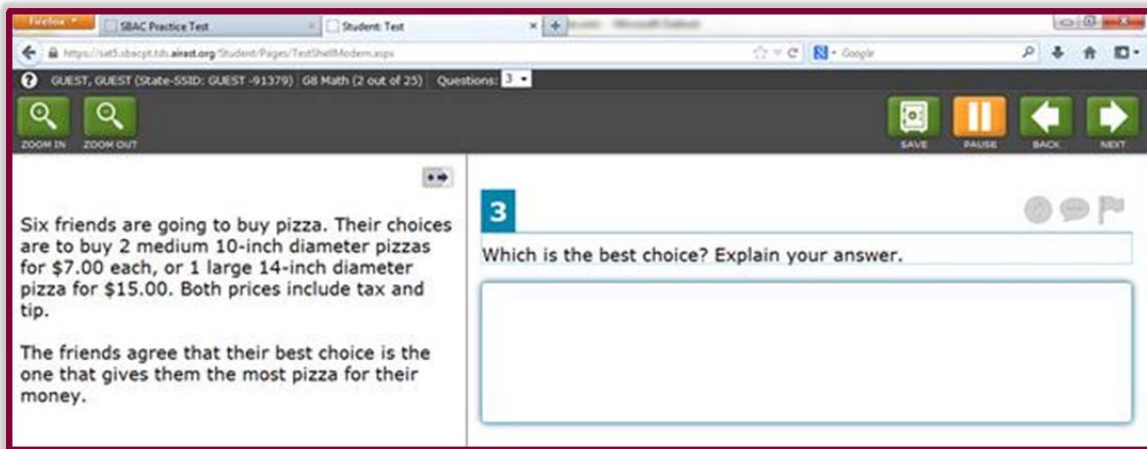
III. Item Construction

After specifications have been written, **items are constructed corresponding to the specifications using the online *Bank Builder* utility**. *Bank Builder* includes features that promote item quality. For example, images constructed using *Bank Builder* conform to size and resolution standards that yield items that are visually suited to online and offline administration. Similarly, fonts available in *Bank Builder* are chosen to work across platforms, for readability both onscreen and in printed material. *Bank Builder* also allows for the use of item families, or groups of items that refer to the same contextual material. For example, several language arts items may refer to the same reading passage.

To ensure reliable and efficient measurement of standards, passage length is controlled to accommodate the administration of tests of adequate length within the time constraints typically found in school settings. Reliability is a direct function of test length. ATI research indicates that adequate reliabilities can typically be achieved for assessments of approximately 40 items in length (Bergan, Burnham, Bergan, & Bergan, 2008). When excessively long passages are selected or when excessive numbers of passages are selected, assessment length is reduced and reliabilities may fall to unacceptable levels.

The item construction process also includes attention to important principles of item writing. For example, when writing selected-response items, incorrect alternatives must be plausible answers that may provide clues to the manner in which the student conceptualizes the problem reflected in the question. Distractors should resemble the correct choice in grammatical form, style and length (Haladyna, 2004). Care is also taken to ensure that items are written to accommodate students from diverse backgrounds. Sensitivity to diversity reflected in the needs of local programs is addressed in the test review process which enables districts/charter schools to select items appropriate for use with their students. For example, items are written to include a diversity of ethnic and multi-ethnic names so that districts/charter schools will have choices available to meet local needs when deciding which items to include.

When writing items including contextual material such as text passages, care is taken to ensure that the passages are not ones that the student is likely to have read previously. In many instances, original texts are utilized. In those cases in which previously published material is used, districts/charter schools are encouraged to select passages that are not likely to have

been previously read by students. When previous passage exposure is not controlled, assessment validity may be compromised. Student performance may reflect differential exposure to content rather than comprehension of the material presented (Haladyna, 2004).

Careful attention is also paid to ensure that the material is at the appropriate reading level. The level of text vocabulary and the readability of the text need to be considered in assessing the reading comprehension and reading skills of the students. Although the reading level of a passage may be quantified in several ways, there is a subjective component as well. Neither the complexity of sentence structure, the difficulty in comprehending the concepts presented in a passage, nor the level of abstraction can be quantified. For example, "I think, therefore I am." results in a very easy readability level, but the concepts addressed by the sentence are not so easy to understand. The appropriateness of a given passage for a given grade level is also an issue. Therefore, while we do use one measure of readability in the form of Lexile® measures, we also analyze the reading passages more subjectively to ensure that reading ability does not get in the way of, for example, correctly answering a math question.

IV. Review of Items and Related Materials

The review of items and associated materials is a multi-stage process. As indicated above, each item is written to conform to an item specification. The first step in the process is to review the item specification. The second step includes an internal item review and item certification. This step may be accompanied by an independent review of materials such as texts that may be common to a number of items. The third step in the process is external review.

A. Review Criteria

The review process is informed by criteria established for the development of item specifications and for the items designed in accordance with those specifications. The following list outlines the criteria addressed in reviewing item specifications.

- The item specification must identify a class of items and detail the characteristics of items in the class.
- The item specification must include one or more content standards to which items in the specification are to be aligned.
- The specification must include at least one prototype item representing the class identified in the specification.
- The content standards will typically be comprised of a number of skills identified as part of the process of constructing the item specification. The item specification must be designed to support the assessment of at least one of the skills comprising each standard.
- Each skill included in the specification must be linked to a task to be accomplished by the students being assessed.
- The item specification must detail the defining features of the task. For example, if the task takes the form of a selected-response item, the specification must include the defining characteristics of the stem and alternatives.
- The specification must indicate the DOK of the task and may address demands of the task expected to affect item difficulty. For example, the cognitive complexity of the task and its effect on difficulty might be addressed in developing the specification.

The list below details criteria related to the review of individual items.

- Each item must reflect the class of items defined by the specification.
- Each item must be aligned with the content standards and specific skills within the standard included in the item specification.
- The items must reflect the DOK level and task demands associated with the task defined in the specification.
- Each item must meet standards with respect to font size and type. Determination of these item attributes is controlled automatically by Galileo® K-12 Online.
- Each item must conform to image size and quality standards. Adherence to these standards is controlled automatically by Galileo K-12 Online.
- Each item developed for the specification must meet all specification criteria.
- Each item must be evaluated for clarity of expression and adherence to grammatical rules.
- Each item must be evaluated for grade-level appropriateness of language.
- Each item must be evaluated with respect to sensitivity to the needs of students from diverse backgrounds.

In some cases, common material called an item family is developed for use with multiple items. For example, a common text may be used with a number of items assessing various aspects of text comprehension. **The ability to link a single text to multiple items increases the number of skills that can be assessed within a given time period. This enhances test reliability and is associated with increased forecasting effectiveness.** The common material comprising an item family is reviewed independently as well as in conjunction with the specific items linked to the family. Item families are typically linked to items reflecting multiple specifications. As the following list of criteria shows, this linkage affects item family review:

- Item families involving text must conform to stylistic norms associated with the genre they are intended to represent.
- Item families involving text must conform to grammatical conventions associated with the genre that they reflect.
- In some cases, item families are required to reflect errors in reasoning and/or grammatical errors.
- Typically, item families must be written in a manner that accommodates multiple items reflecting different item specifications. For example, an item family might accommodate questions addressing specifications related to character, plot, and setting.
- Text comprising an item family should be written concisely. Excessive text length may adversely affect both reliability and validity.
- Text should be written at an appropriate readability level.

B. The Review Process

The initial review process for item specifications, items, and item families occurs in-house and is composed of three steps. The process is completed when the item has been certified. The item review process continues once a certified item is made available for use in an assessment. Items placed in district/charter school designed assessments are subjected to review by each member of the district/charter school's review team during the test review process. The following discussion details both the internal and the district/charter school review processes.

The internal review process for item specifications, items, and item families occurs in three steps. Step one is an independent review using the online bank review tool in Galileo® K-12 Online. Reviewers may accept the material being reviewed, reject it, or accept it with modifications. If the material is accepted with modifications, the reviewer provides the writer with a list of suggested modifications, which are recorded in the online review. The modifications guide revision. Following the initial review, the material is subjected to a final review. **When consensus is reached between the two reviewers and the writer, the material under review is accepted for use.** In the case of item review, the item is certified. Certified items are locked so that they cannot be changed by Galileo users. Only certified items are provided by ATI for use by school districts/charter schools.

When a certified item is made available for use in an assessment, it is subjected to test review by districts/charter schools planning assessments that may use the item. Test review is a two-stage process. In the initial stage, any number of district/charter school reviewers may review the assessment. In stage two, a final reviewer assembles the reviews from stage one and submits to ATI a final review of the assessment. Test review provides the district/charter school with the option to accept an item or replace it with an item that more closely meets their needs. The district/charter school may also suggest item changes. The development of new items based on district/charter school suggestions is one source of the continuing expansion of ATI item banks. ATI item banks, which currently contain over 118,000 items, are among the largest item banks aligned to standards in the nation. **During the course of a single school year, ATI items are reviewed by hundreds of educators.** The combination of a multi-material (specification, item, and item family) multi-step internal review and item certification accompanied by continuous external review by educators representing districts/charter schools with a variety of needs provides a rigorous approach to evaluation producing assessments that meet local needs and that are highly effective in forecasting and guiding instruction toward standards mastery.

V. Selected-Response, Technology-Enhanced, and Constructed-Response Item Formats

Galileo **supports selected-response, technology-enhanced, and constructed-response item formats**. Selected-response items have held a dominant place in educational assessment for a very long time. In the past, selected-response items have had significant advantages over other formats with respect to assessment efficiency, and flexible support for online and offline assessment. Selected-response assessment with automated scoring minimizes the potential for error inherent in hand scoring and avoids the additional time and resources necessary to administer and manually score constructed response items (Downing, 2006; Haladyna, 2004). Given their many advantages, it is not surprising that selected-response items continue to play an important role in assessment. Despite these advantages, selected-response items have significant shortcomings. In some cases, they may not provide a valid indicator of the skill to be assessed. In addition, they are not well suited to the task of assessing skills involving high DOK levels.

Advances in technology coupled with the adoption of CCSS have promoted the development of a variety of **new technology-enhanced items that are well suited to assessing competencies involving high DOK levels and that can be scored automatically**. ATI has initiated an ongoing development project producing several of these item types. As one example, an **interdependent item set** similar to a released PARCC prototype item was illustrated in Section II. The interdependent item set enables the inclusion of multiple sub items

aligned to individual standards in the context of a high DOK level task (e.g., drawing a conclusion and citing textual evidence in support of the analysis). As additional examples, the following screen shots illustrate an **interactive text item type** similar to a released SBAC practice item that enables students to answer a question by selecting the answer from within the text passage itself and a **sequencing item type** that enables students to arrange a series of events from a text passage in the correct order via drag-and-drop technology.

Interactive text item type: Student selecting answer within the text passage – Confidential Screen Shot

Item Development and Assessment Construction
Guidelines for Common Core State Standards
Assessment Technology, Incorporated

1.800.367.4762
ati-online.com

Constructed-response items are useful in addressing skills involving high DOK levels. Moreover, these items often have significant advantages related to validity. More often than not, a valid measure of a skill requires the examinee to do something from scratch rather than selecting an answer from a series of alternatives. Galileo® provides technology that supports the use of constructed-response items. For example, *Test Builder* supports extended response items such as essays. In addition, scoring guidelines can be entered into the system and made available online. Finally, hand written student responses can be scanned into Galileo and scored manually online. There are considerations that should be addressed in order to utilize constructed-response items effectively. For example, constructed-response items typically must be scored manually. ATI has the capability to offer automated scoring. However, because of current limitations in automated scoring, it is not widely used. Regardless of the approach to scoring, clearly defined scoring rules must be established. In addition, when constructed-response items will be scored by teachers, it is useful to assess the reliability of scores. This may be accomplished by assessing agreement between two or more observers.

VI. Writing Assessment

To assess writing without automated scoring, it is necessary to apply a standardized rubric and a writing prompt that allows the student to express their responses in a manner that indicates an accurate representation of their ability to compose, convey and communicate to match both the purpose of the text and the information that they possess relevant to the topic. Technology-enhanced and selected-response items may also be used to assess some aspects of writing. ATI has developed a wide variety of item types that can be used to assess writing as consistent with the needs, technological capabilities, and available resources of each district/charter school. For example, the sample specification in the following screen shot defines the characteristics of a constructed-response writing item for third grade students. This item specification defines a class of items that are highly similar to a released practice item from SBAC. Within Galileo® K-12 Online, students can respond to constructed-response items online via the online testing interface or offline via printed worksheets. If students respond offline, the worksheet with the student's response can be scanned into the system where it can be accessed, stored, and scored online. Alternatively, teachers can score the items offline and enter the scores into the program for each student.

Sample item specification for third grade writing – *Confidential Screen Shot

Description:
Student must write a topic sentence for a given paragraph.

Work Sample type.
Points: 1

Dok Justification: Connecting ideas and thinking about organizational structure is a level 2 writing task.

Stimulus Attributes:
a persuasive paragraph w/no topic sentence

Stem Attributes:
Read the paragraph. Write a topic sentence for the paragraph.

Scoring Rules:
Student writes a topic sentence that introduces the paragraph. Record the score: 1 point possible.

Prototype Question

Read the paragraph.

They are very fun to play with. For example, they will chase a string or a small ball.
Cats are also very cuddly. They may like to lay in your lap and purr or sleep with you.
Cats are also easy to take care of. You should get a cat!

Write a topic sentence for the paragraph.

Depth of Knowledge: 2   Spec Question Type: Not Set

Scale Name	Performance Objective
AZ CC-R03: 3rd Grade English Language Arts	**AZ-W.3.1a Text Types and Purposes: Introduce the topic or text they are writing about, state an opinion, and create an organizational structure that lists reasons.**
CA CC-R03: 3rd Grade English Language Arts	**CA-CC-W.3.1a Text Types and Purposes: Introduce the topic or text they are writing about, state an opinion, and create an organizational structure that lists reasons.**
CA-R03: 3rd Grade Reading	**CA-3WS1.1 Create a single paragraph: a) Develop a topic sentence; b) Include simple supporting facts and details.**
CC-R03: 3rd Grade English Language Arts	**CC-W.3.1a Text Types and Purposes: Introduce the topic or text they are writing about, state an opinion, and create an organizational structure that lists reasons.**
CO CC-R03: 3rd Grade English Language Arts	**CO-03.WC.1a.i Introduce the topic or text they are writing about, state an opinion, and create an organizational structure that lists reasons. (CC-W.3.1a)**
HI CC-R03: 3rd Grade English Language Arts	**HI-CC-W.3.1a Text Types and Purposes: Introduce the topic or text they are writing about, state an opinion, and create an organizational structure that lists reasons.**
IL CC-R03: 3rd Grade English Language Arts	**IL-W.3.1a Text Types and Purposes: Introduce the topic or text they are writing about, state an opinion, and create an organizational structure that lists reasons.**
MA CC-R03: 3rd Grade English Language Arts	**MA-W.3.1a Text Types and Purposes: Introduce the topic or text they are writing about, state an opinion, and create an organizational structure that lists reasons.**
MO CC-R03: 3rd Grade English Language Arts	**CC-W.3.1a Text Types and Purposes: Introduce the topic or text they are writing about, state an opinion, and create an organizational structure that lists reasons.**
RI CC-R03: 3rd Grade English Language Arts	**CC-W.3.1a Text Types and Purposes: Introduce the topic or text they are writing about, state an opinion, and create an organizational structure that lists reasons.**
WA CC-R03: 3rd Grade English Language Arts	**CC-W.3.1a Text Types and Purposes: Introduce the topic or text they are writing about, state an opinion, and create an organizational structure that lists reasons.**

VII. Local Control of Benchmark and Formative Assessment Specifications

If assessments are to be used to measure standards targeted for instruction in school districts/charter schools, districts/charter schools need to have a significant degree of control over test specifications (Standard 3.3 AERA, APA, NCME, 1999). For example, if a district/charter school targets a particular set of standards for instruction during a given time period, the district may elect to administer a benchmark test to assess those standards. In order to ensure the adequacy of the assessment for each standard, the district/charter school will probably want to control the number of items selected to assess each standard. In addition, the district/charter school may wish to select items that reflect varying depths of knowledge (see e.g., Webb, 2006) reflected in the specific skills targeted for instruction. The district/charter school may also want items that are sensitive to the experiences of students from diverse backgrounds represented in the district/charter school. Finally, the district/charter school will typically need to control the length of the assessment to cover the targeted content and to meet time constraints such as the length of periods during which assessments can be scheduled. The Galileo® K-12 Online Instructional Improvement and Instructional Effectiveness System includes a wide variety of tools that support local control of local assessments.

VIII. Universal Design

Thurlow, Thompson, and Lazarus (2006) present the basics of universal design as it applies to assessments in the following statement: "The essential idea behind universally designed assessments is that they are to ensure that the assessment measures what is really intended-the relevant constructs-rather than construct irrelevant information." Construct irrelevant information might include the extent to which a student's comprehension of English impairs their ability to understand what is asked in a word problem on a math test or it might include fatigue that impairs ability to perform calculations. In both cases, the validity of the final score as a measure of math skill might be unduly impacted if adequate attention is not given to assessment design and to the provision of appropriate accommodations.

A. Objectives of Universal Design

Thompson, Johnstone, and Thurlow (2002), lay out guidelines to ensure that tests meet the objectives of universal design. The discussion that follows illustrates ATI's application of those guidelines in the construction of assessments:

1. Precisely defined constructs: As indicated above, an essential part of the item development process followed by ATI is the construction of item specifications. Item specifications are written first before any items are produced. The specifications serve as a blueprint to guide both the writing and later quality assurance review. Item specifications include detailed definitions of the exact capability that is intended to be measured. This level of detail helps to make sure that the items that are ultimately available for inclusion on an assessment provide data on the intended ability.

2. Accessible, nonbiased items: Items are reviewed by ATI content specialists to ensure that they are free from potential sources of bias. The review process includes two phases. The first occurs before the item is made available for use. ATI content specialists review whether the items conform to the specifications and are free from identifiable sources of bias. The second phase occurs after administration of the assessment. It is well known that items can behave quite differently across a variety of student populations. ATI conducts research to ensure that districts/charter schools are provided with valid, reliable scores.

3. Amenable to accommodations: The ATI system of assessment design and delivery has been constructed to be supportive to a variety of different accommodations. These are listed in detail in the next section in this document.

4. Simple clear and intuitive instructions and procedures: In addition to a clear statement of the construct to be measured, item specifications also include guidance about the instructions to be provided to the student. These are reviewed as part of the review process.

5. Maximum readability and comprehensibility: One of the fundamental objectives of the item specification is to ensure that the items that are produced are as clear and comprehensible as possible. ATI assesses the readability of each of the items in its banks in multiple ways: One involves the use of a readability index in the form of Lexile® measures. Another includes reference to vocabulary lists that indicate grade-appropriate vocabulary.

6. Maximum legibility: Legibility is supported in several ways. First, items are designed to be presented in a consistent fashion when administered online. For example, all images are constructed to conform with specifications that ensure legibility. Second, printing algorithms are also designed to use consistent fonts and item layouts. Third, assessments may be printed in large print formats.

B. Providing the Full Range of Accommodations

In order to effectively assess the capabilities of ELL students and students with special needs, the following standard accommodations may be allowed for ATI benchmark and formative assessments. Determination regarding the use of these accommodations should be based on the student's ELL status and/or the specific needs outlined in his or her Individualized Education Program (IEP) if one is in place.

1. Extended Time: Those students whose IEPs indicate that they would benefit from extra time can be allowed to have the time that they require.

2. Large Print: Large print versions of the tests may be easily provided for visually impaired students from the printing interface.
3. Use of a scribe: A scribe may be provided to assist in recording a student's answers to test items. The scribe must limit their assistance to recording the student's responses as given.
4. Use of a screen reader: A screen reader may be provided to assist visually impaired students to obtain information from text. The Galileo® K-12 Online testing interface has been designed to support the use of screen readers.
5. Oral presentation or translation of test items and directions: It is permissible for the district/charter school to provide a proctor to read the directions and items aloud for ELL and special needs students. Oral translation may also be used.
6. Settings accommodations: If a student's IEP includes recommendations that the student be allowed to complete assessments in an alternative setting, then this should be allowed.

To assist in the analysis and interpretation of test scores involving ELL and/or special needs students, the district/charter school is provided with an interface in the online application to indicate the accommodations that were used, the students with whom they were used, the particular assessment to which the accommodations were applied, and the reason for the accommodation.

The interface also provides space for documenting any non-standard accommodations that may be required. It is recommended that these be covered with the district/charter school's ATI Field Services Coordinator prior to the administration of the assessment as this will assist ATI to plan subsequent data analyses appropriately.

IX. Assessment Design and Forecasting Effectiveness

Customized benchmark assessments constructed in the Galileo assessment component are used to inform instruction and to forecast performance on statewide assessments. Data on forecasting effectiveness can be found in the technical manual for the Galileo K-12 Online Instructional Improvement and Instructional Effectiveness System (Bergan, Burnham, Bergan & Bergan, 2008). The discussion here focuses on assessment design requirements related to forecasting. Assessment design supporting forecasting effectiveness is based on two considerations: First, the benchmark assessments used to forecast standards mastery must provide credible information regarding the likelihood that students will meet standards as determined by their performance on the statewide test. The technical manual provides evidence related to this issue. Second, benchmark assessments must assess standards that are targeted for instruction. If this were not the case, the assessment would not provide information that could be used to guide instruction. There are a variety of ways to design assessments to reflect what is targeted for instruction. The Galileo K-12 Online Instructional Improvement and Instructional Effectiveness System has the technology that makes it possible to customize assessments to reflect the district/charter school curriculum. This technology makes it possible to customize thousands of assessments each year. Reports obtained from these assessments identify students at risk of not meeting standards. In addition, recommendations are made as to which standards should be addressed next in order to bring at-risk students on course to meet standards.

X. Summary of Item Development and Assessment Construction Guidelines

The material in this document provides a summary of guidelines ATI uses to inform item development and assessment construction for a wide variety of types of assessments (e.g., benchmark assessments, formative assessments). Implementation of these guidelines has produced one of the largest assessment item banks in the nation. Tools in the Galileo® system have made it possible to produce thousands of reliable and valid customized assessments aligned to standards reflecting district/charter school curriculums. The guidelines have promoted the development of efficient assessments of standards mastery within the context of the time constraints and logistical needs of schools. The precision and quality assurance features of the development process have minimized errors. The ongoing research program that accompanies item development and test construction has produced thousands of tests that are both effective in diagnosing learning needs and in predicting standards mastery on statewide assessments.

The adoption of CCSS has introduced new requirements related to item development and assessment construction. ATI has implemented a plan to respond to these new requirements by developing new content including selected-response, technology-enhanced, and constructed-response items as well as new approaches to the format of assessments including multi-stage computerized adaptive tests. These innovations will assist districts and charter schools in transitioning instruction and assessment to align with the new standards. These innovations will also ultimately assist districts and charter schools in elevating student achievement by providing students with the opportunity to master the content addressed in the new standards as well as to practice the online test-taking skills required for success in the next generation of statewide assessments created by PARCC and SBAC.

XI. Text References

AERA, APA, & NCME (1999). Standards for educational and psychological testing. Washington, D.C.: Author.

Bergan, J.R., Burnham, C.G., Bergan, J.R. & Bergan, K.C. (2008). *Benchmark Assessment Development in the Galileo Educational Management System.* Tucson, Arizona: Assessment Technology Incorporated

Briars, D. J. (2012). The Common Core Transition in Mathematics: What States, Districts, and Schools Can Do Now.

Downing, S.M. (2006). Selected-Response Item Formats in Test Development. *Handbook of Test Development* (pp. 287-302). Mahwah, N.J.: Lawrence Erlbaum Associates.

Haladyna, T.M. (2004). *Developing and Validating Multiple-Choice Test Items* (3rd ed.). Mahwah, N.J.: Lawrence Erlbaum Associates.

Krehbiel, C. (2012). The Common Core Transition in English Language Arts: What Districts, Schools, and Teachers Can Do Now

Thompson, S. J., Johnstone, C. J., & Thurlow, M. L. (2002). *Universal Design Applied to Large Scale Assessments* (Synthesis Report 44). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Item Development and Assessment Construction
Guidelines for Common Core State Standards
Assessment Technology, Incorporated

- 19 -

1.800.367.4762
ati-online.com
Copyright © Assessment Technology, Incorporated 2013. All rights reserved.

Thurlow, M.L., Thompson, S.J., Lazarus, S.S. (2006). Considerations for the Administration of Tests to Special Needs Students: Accommodations, Modifications, and More. *Handbook of Test Development* (pp. 653-676). Mahwah, N.J.: Lawrence Erlbaum Associates.

Webb, N.L. (2006). Identifying Content for Student Achievement Tests. In S.M.Downing and T.M.Haladyna (Eds.). *Handbook of Test Development* (pp. 155-180). Mahwah, N.J.: Lawrence Erlbaum Associates.