Kansas Assessment Program
Technical Manual
2016

University of Kansas Achievement & Assessment Institute

January 2017

Table of Contents

Table of Tables

Table of Figures

I. Statewide System of Standards and Assessments

The Kansas Assessment Program (KAP) under the Kansas State Board of Education (hereafter the State Board) is mandated by the Kansas State Legislature. KAP also complies with federal elementary and secondary education legislation. The four main purposes of the KAP, as stated in the *Kansas Assessment Examiner's Manual 2015–2016* (hereafter the *Examiner's Manual*; Kansas State Department of Education [KSDE], 2015a), are to

- measure specific claims related to the Kansas College and Career Ready Standards (KCCRS);
- provide information for calculating Annual Measurable Objectives and for state accreditation;
- report individual student scores, along with the student's performance level; and
- provide subscale and total scores that can be used with local assessment scores to assist in improving a building's or district's programs in English language arts (ELA); mathematics; science; and history, government, and social studies (HGSS).

The state statutory authority behind KAP is Kan. Stat. Ann. §72-6479, School-performance accreditation system; curriculum standards; student assessments; school site councils (2015). According to this statute, the State Board is mandated to, in part,

- design and adopt a school-performance accreditation system based upon improvement in performance that reflects high academic standards and is measurable;
- establish curriculum standards that reflect high academic standards for the core academic areas of mathematics, science, reading, writing, and social studies; and
- provide statewide assessments in the core academic areas and determine performance levels on the statewide assessments.
- Performance expectations address what students should know and be able to do. They should represent high academic standards in the academic area at the grade level to which the assessment applies.

KAP offers three summative assessments: the test for the general student population, the alternate assessment for students with significant cognitive disabilities, and the test for English Language learners. This technical manual addresses the test for the general student population, which will be referred to as the KAP. Additionally, this report follows the reporting structure recommended in the 2015 Assessment Peer Review Guidance (U.S. Department of Education, 2015).

I.1. State Adoption of Academic Content Standards for All Students

The state legislature mandated that KSDE review the Kansas curriculum every seven years. The State Board adopted the KCCRS for ELA and mathematics in 2010, for science in June 2013, and for HGSS in April 2013. The first operational administration of the KCCRS for ELA and mathematics was in 2015 and HGSS was in 2016. Science's first KCCRS operational administration will be 2017.

I.2. Coherent and Rigorous Academic Content Standards

Standards in Kansas were developed and reviewed by committees of Kansas educators and stakeholders. These standards help schools prepare students by providing the knowledge and skills needed to pursue higher education or better careers and to compete in an increasingly competitive and global work environment. The KCCRS are Kansas's coherent and rigorous academic content standards, which adherent to the State Board's mission.

> The mission of the Kansas State Board of Education is to prepare Kansas students for lifelong success through rigorous, quality academic instruction, career training and character development according to each student's gifts and talents. The Kansas CAN Vision is to Lead the World in the Success of Each Student (refer to http://www.ksde.org/Board).

I.2.1. Process and timeline. Under the direction of and feedback from Kansas educators, the KCCRS were adapted from the Common Core State Standards (CCSS). Beginning in November 2009, KSDE received drafts of the CCSS and provided feedback to the Council of Chief State School Officers (CCSSO). From January 2010 to August 2010, Kansas educators who served on the ELA or mathematics KCCRS committee provided feedback to the CCSSO and other groups involved in the development process; this feedback was incorporated into subsequent drafts of the CCSS. In September 2010, the standards for ELA and mathematics were presented to the State Board, which on October 10, 2010, adopted the KCCRS for ELA and mathematics for use in Kansas.

Kansas led the development of the Next Generation Science Standards (NGSS). Beginning in 2011, participating states and standards writers were recruited to start the development process. Between 2011 and 2013, writing teams and stakeholders reviewed and revised a series of drafts of the science standards, with periods of public review. When the new standards were completed, the Kansas standards development committee thoroughly reviewed the document to verify that feedback from the Kansas review team was acknowledged and that the standards represented the best interests of Kansas students. In May 2013, the Kansas NGSS review committee recommended the KCCRS for science to the State Board, which adopted the standards on June 11, 2013, after a month of deliberation.

The development of Kansas HGSS standards was undertaken by a committee of Kansas educators and stakeholders in May 2011 and culminated with adoption of the standards by the State Board on April 16, 2013. From the outset, the goal of the HGSS standards committee was to create a document that would emphasize and encourage the application of content in authentic situations, rather than a traditional approach to HGSS standards that focuses on dates and minutiae. To this end, the final standards represent methods of thinking rather than a document to be used as a scope and sequence. The Mission Statement in the HGSS content standards reads "The Kansas Standards for History, Government, and Social Studies prepare students to be informed, thoughtful, engaged citizens as they enrich their communities, state, nation, world, and themselves" (KSDE, 2013, p. 5).

The drafting of the content standards was an iterative process, moving from the committee to public comment and review and back to the committee. In total, the document went through three cycles of public review and revision before it was submitted to the State Board in October 2012 for review and feedback. The HGSS committee incorporated the additional changes recommended by the State Board and presented the standards for adoption in March 2013. The State Board adopted the standards in April 2013.

I.2.2. Convergence and divergence with national standards. According to the Common Core State Standards Initiative, the CCSS

> define what students should understand and be able to do by the end of each grade. They correspond to the College and Career Readiness (CCR) Anchor Standards [in the KCCRS] … by number. The CCR and grade-specific standards are necessary complements—the former providing broad standards, the latter providing additional specificity—that together define the skills and understandings that all students must demonstrate. (2010, p. 10)

The key difference between the national CCSS and Kansas's KCCRS is the Kansas 15%, the purpose of which is to emphasize concepts and teaching philosophies that are important in Kansas. Although most of the Kansas concepts are mentioned in the CCSS, KSDE wanted to highlight the importance of each one by including the concepts and teaching philosophies in the KCCRS. As part of the Kansas 15%, KSDE added the anchor standards for literacy learning, as well as four other anchor standards—two in reading and two in writing (KSDE, 2010a).

For mathematics, the Kansas additions to the CCSS were for probability and statistics and algebraic patterning. These two topics were left for each school and/or district to decide how to incorporate them (KSDE, 2010b).

The development of the NGSS was led by Kansas; thus, the KCCRS for science closely align to the NGSS. The NGSS are based on the *Framework for K–12 Science Education* developed in 2012 by the National Research Council of the National Academies. However, the intents of the NGSS are to put the *Framework* into practice by coupling the practice with content, provide performance expectations but leave curricular and instructional decisions to states and educators, and evaluate students on the degree of understanding of a full discipline core idea. The NGSS were developed because

> the world has changed dramatically in the 15 years since state science education standards' guiding documents were developed. Since that time, many advances have occurred in the fields of science and science education, as well as in the innovation-driven economy. The U.S. has a leaky K–12 science, technology, engineering and mathematics (STEM) talent pipeline, with too few students entering STEM majors and careers at every level—from those with relevant postsecondary certificates to PhD's. We need new science standards that stimulate and build interest in STEM.

> The current education system can't successfully prepare students for college, careers and citizenship unless we set the right expectations and goals. While standards alone are no silver bullet, they do provide the necessary foundation for local decisions about curriculum, assessments, and instruction.
>
> Implementing the NGSS will better prepare high school graduates for the rigors of college and careers. In turn, employers will be able to hire workers with strong science-based skills—not only in specific content areas, but also with skills such as critical thinking and inquiry-based problem solving. (Next Generation Science Standards, 2013, p. 1 of Introduction)

The mission of the KCCRS for HGSS, as described in the *Kansas Standards for History, Government, and Social Studies*, is to "prepare students to be informed, thoughtful, engaged citizens as they enrich their communities, state, nation, world, and themselves" (KSDE, 2013, p. 5). To develop the KCCRS for HGSS, the standard writing committee

> reviewed other state and national standards, researched best instructional practices, and gathered input from professionals and citizens in order to define what Kansas students should be able to know and to do in history, civics/government, geography, and economics. The committee responded to feedback on earlier versions throughout the current process. This revised document focuses on discipline-specific habits of mind that encourage the application of content in authentic situations rather than specific content, and is intended as a framework for curriculum, instruction, assessment, and teacher preparation (KSDE, 2013, p. 6).

I.2.3. Standard review committees. Committee members involved in the development of the Kansas additions to the CCSS for ELA and mathematics were recruited across states. The ELA committee comprised 22 members, and the mathematics committee was composed of nine members; most members were K–12 educators. Additionally, two representatives from postsecondary education were recruited for each subject.

The KCCRS for science were reviewed by the Kansas review team and the Kansas science education committee, a subcommittee of the review team. The review team included 60 members from across the state and comprised K–12 science educators, postsecondary science professors, and business and industry professionals. The subcommittee focused on finding ways to "build and leverage relationships between P-20[1] educators and business and industry to build state-wide capacity for science education in Kansas." (Kansas Next Generation Science Education, http://community.ksde.org/Default.aspx?tabid=5407).

The panel of HGSS committee members was the result of nominations from State Board members and the Commissioner of Education, as well as internal nominations from KSDE content staff who were familiar with top Kansas educators and community leaders. Committee members included representation from the community at large, and several state and national

[1] P-20 refers to the integrated education system that extends from preschool through higher education.

organizations. Their expertise included HGSS teaching and curriculum, special education, and English language learning (ELL). The final committee was composed of approximately 30 individuals from across the state and was facilitated by Donald Gifford of KSDE.

I.3. Required Assessments

The KAP tests students in the subject areas of ELA, mathematics, science, and HGSS. The subject areas and grades tested are

- ELA in grades 3–8 and 10;
- mathematics in grades 3–8 and 10;
- science in grades 5, 8, and 11; and
- HGSS in grades 6, 8, and 11 (tested in even-numbered years, e.g., 2016, 2018, etc.).

I.4. Policies for Including All Students in Assessments

Kansas is committed to including all students in the KAP. Students enrolled in the Kansas public schools must take one of the three summative tests: the KAP, English language proficiency test, or alternate assessment. ELL students who are recent arrivals to the United States are required to take the KAP mathematics and science tests, but their results count only toward participation. They are not required to take the ELA or HGSS tests, but must take the English language proficiency test.

Qualifying students with significant cognitive disabilities, typically not more than one percent of Kansas students, take the Dynamic Learning Maps® Alternate Assessment for ELA, mathematics, and science and a separate HGSS Alternate Assessment. Other special-needs students with Individualized Education Programs, 504 plans, or Student Intervention Team plans take the KAP but can use accommodations consistent with their personal needs profiles (PNPs). If an unapproved accommodation is used (e.g., reading aloud to student on the KAP ELA test), the student is considered "not tested." A detailed accommodation summary can be found in chapter V. Inclusion of All Students of this technical manual.

There are only a few exemptions granted to students. The exceptions include

- students serving long-term suspension;
- students who were truant for more than two consecutive weeks at the time of testing;
- students who experienced catastrophic illnesses or accidents during testing;
- students who moved during testing; and0
- students who were incarcerated during testing.

I.5. Participation Data

In 2016, the KAP operational test was administered in ELA, mathematics, and HGSS. Table I-1 shows the number and percentage of enrolled students who took each test in each grade. Most of the tested rates are 98% and above; two high school subjects demonstrated tested rates of 96%.

Table I-1. Number and Percentage of Enrolled Students Tested by Subject Test and Grade

Subject test	Grade						
	3	4	5	6	7	8	HS
ELA							
No. enrolled	38,880	37,981	37,263	37,142	37,114	36,683	36,552
No. tested	38,135	37,176	36,596	36,386	36,267	35,772	34,982
Percentage tested	98%	98%	98%	98%	98%	98%	96%
Mathematics							
No. enrolled	38,866	37,929	37,257	37,008	37,006	36,673	36,553
No. tested	38,138	37,185	36,606	36,338	36,190	35,796	35,029
Percentage tested	98%	98%	98%	98%	98%	98%	96%
HGSS							
No. enrolled				37,003		36,644	34,472
No. tested				36,463		35,851	33,230
Percentage tested				99%		98%	96%

Note. HS = high school.

II. Assessment System Operations

The development of any test requires many important decisions regarding, for example, the content and cognitive complexity, as well as the appropriate scope, sequence, and progression of that content for particular subject areas. Other decisions relate to the number of points for each test and the proportion of those points for any subscores. These decisions are not made in isolation but instead must be reasonable across all grade levels of the assessment. Together, these decisions represent the constructs that a test measures. Critical test construction related documents yielding from these decisions include development timeline and test blueprint. These documents guide the test construction process and products.

II.1. Content Standards of the Assessed Grades

KAP KCCRS content standards—except for HGSS, which labels its content hierarchy by standards and benchmarks—are defined in two levels: claims and targets, not all claims have a target sublevel, and an item may be assigned to multiple claims. Tables II-1 through II-6 show the KCCRS content standards for the four KAP subjects. ELA and mathematics have the same claims and targets across grades. Initially ELA had only two claims: reading and writing. Listening was added as the third claim in 2016. Mathematics has four claims and all of its targets are under Claim 1. Science has the same claims but different targets across grades. HGSS standards and benchmarks are identical across grades.

Table II-1. ELA Claims and Targets

Claim	Claim label	Target
1	Reading	Literary texts Informational texts Making and supporting conclusions Main ideas
2	Writing	Revising Editing Vocabulary and language use
3	Listening	

Table II-2. Mathematics Claims and Targets

Claim	Claim label	Target
1	Concepts and procedures	Operations and algebraic thinking Number and operations in base ten Numbers and operations with fractions Measurement and data Geometry The number system Expressions and equations Statistics and probability

Claim	Claim label	Target
		Algebra
2	Problem solving	
3	Communicating reasoning	
4	Modeling and data analysis	

Table II-3. Science Grade 5 Claims and Targets

Claim	Claim label	Target
1	Physical science	Structure and properties of matter Engineering design in physical science
2	Life science	Matter and energy in organisms and ecosystems Engineering design in life science
3	Earth and space science	Earth's systems Space systems Engineering design in earth and space science

Table II-4. Science Grade 8 Claims and Targets

Claim	Claim label	Target
1	Physical science	Structure and properties of matter Chemical reactions Forces and interactions Energy Waves and electromagnetic radiation Engineering design in physical science
2	Life science	Structure, function, and information processing Matter and energy in organisms and ecosystems Interdependent relationships in ecosystems Growth, development, and reproduction of organisms Natural selection and adaptations Engineering design in life science
3	Earth and space science	Space systems History of the earth Earth's systems Weather and climate Human impacts Engineering design in earth and space science

Table II-5. *Science Grade 11 Claims and Targets*

Claim	Claim label	Target
1	Physical science	Structure and properties of matter Chemical reactions Forces and interactions Energy Waves and electromagnetic radiation Engineering design in physical science
2	Life science	Structure and function Matter and energy in organisms and ecosystems Interdependent relationships in ecosystems Inheritance and variation of traits Natural selection and evolution Engineering design in life science
3	Earth and space science	Space systems History of the earth Earth's systems Weather and climate Human sustainability Engineering design in earth and space science

Table II-6. *HGSS Standards and Benchmarks*

Standard	Benchmark
Choices have consequences.	1.1 The student will recognize and evaluate significant choices made by individuals, communities, states, and nations that have impacted our lives and futures. 1.2 The student will analyze the context under which choices are made and draw conclusions about the motivations and goals of the decision-makers. 1.3 The student will investigate examples of causes and consequences of particular choices and connect those choices with contemporary issues. 1.4 The student will use his/her understanding of choices and consequences to construct a decision-making process and to justify a decision.
Individuals have rights and responsibilities.	2.1 The student will recognize and evaluate the rights and responsibilities of people living in societies. 2.2 The student will analyze the context under which significant rights and responsibilities are defined and demonstrated, their various interpretations, and draw conclusions about those interpretations.

Standard	Benchmark
	2.3 The student will investigate specific rights and responsibilities of individuals and connect those rights and responsibilities with contemporary issues. 2.4 The student will use his/her understanding of rights and responsibilities to address contemporary issues.
Societies are shaped by beliefs, ideas, and diversity.	3.1 The student will recognize and evaluate significant beliefs, contributions, and ideas of the many diverse peoples and groups and their impact on individuals, communities, states, and nations. 3.2 The student will draw conclusions about significant beliefs, contributions, and ideas, analyzing the origins and context under which these competing ideals were reached and the multiple perspectives from which they come. 3.3 The student will investigate specific beliefs, contributions, ideas, and/or diverse populations and connect those beliefs, contributions, ideas and/or diversity to contemporary issues. 3.4 The student will use his/her understanding of those beliefs, contributions, ideas, and diversity to justify or define how community, state, national, and international ideals shape contemporary society.
Societies experience continuity and change over time.	4.1 The student will recognize and evaluate continuity and change over time and its impact on individuals, institutions, communities, states, and nations. 4.2 The student will analyze the context of continuity and change and the vehicles of reform, drawing conclusions about past change and potential future change. 4.3 The student will investigate an example of continuity and/or change and connect that continuity and/or change to a contemporary issue. 4.4 The student will use his/her understanding of continuity and change to construct a model for contemporary reform.
Relationships among people, places, ideas, and environments are dynamic.	5.1 The student will recognize and evaluate dynamic relationships that impact lives in communities, states, and nations. 5.2 The student will analyze the context of significant relationships and draw conclusions about a contemporary world. 5.3 The student will investigate the relationship among people, places, ideas, and/or the environment and connect those relationships to contemporary issues. 5.4 The student will use his/her understanding of these dynamic relationships to create a personal, community, state, and/or national narrative.

II.2. Test Design and Development

The Center for Education Testing and Evaluation (CETE) worked with KSDE to determine the content to be assessed by the KAP subject-area and grade-level tests. The development leading up to the 2016 KAP test administration occurred over multiple years. Table II-7 outlines the test-development timeline for the four subjects: ELA, mathematics, science, and HGSS.

Table II-7. Development Timeline for the KAP

Milestone	Date	Note
ELA/Mathematics		
Adoption of KCCRS	October 2010	
KCCRS item development	2011 to 2016	Determined on a yearly basis
KCCRS items included in the summative assessment	Spring 2012 to spring 2014	Machine-scored items only. Included to provide schools and districts a performance snapshot on the KCCRS but not included in accountability measures.
Operational nonadaptive assessment	Spring 2015	Operational items are machine scored only. Performance tasks are field tested; not machine scored.
Standard setting	Summer 2015	
Operational three-stage adaptive assessment	Spring 2016	Operational items are machine scored only. Includes embedded field testing for machine-scored items. HGSS MDPTs also contribute to ELA scores.
Science		
Adoption of KCCRS	June 2013	
KCCRS item development	2015 to 2016	Determined on a yearly basis
Census field testing	Spring 2016	Machine-scored items only
HGSS		
Adoption of KCCRS	April 2013	
KCCRS item development	2012 to 2016	Determined on a bi-yearly basis
Census field testing	Spring 2015	Both machine-scored and human-scored (MDPT) items
Operational nonadaptive assessment	Spring 2016	Both machine-scored and human-scored (MDPT) items. No field-tested items

Milestone	Date	Note
Standard setting	Spring 2016	

Note. MDPT = multidisciplinary performance task.

II.2.1. Test blueprints.

Test blueprints that specify the number or proportion of items required for each claim (standard for HGSS) is presented in Table II-8. ELA and mathematics have the same claim proportions across grades. Science and HGSS percentages vary slightly across grades; the maximum ranges across grades are presented in the table.

The KAP does not specify total score, score point by claim, or proportion by item type in the blueprints because its test construction is operated under the principle of selecting the most appropriate items for the test (i.e., items that have better statistics and that fit the content requirement). These items may be technology-enhanced items, multiple-choice items, or performance tasks. Scores of these items cannot be categorized by item type because some technology-enhanced items have polytomous scores and some performance tasks have dichotomous scores.

Additionally, the blueprints also do not specify the proportions of depth of knowledge (DOK) levels required for the assessment. However, because each content standard is written with language that relates to DOK, the blueprint presented in Table II-8 indirectly provides the guidance for DOK coverage. For example, the first content standard of the ELA grade 3 is "L.3.1−Demonstrate command of the conventions of standard English grammar and usage when writing or speaking." Its first two sub-standards are "L.3.1a – Explain the function of nouns, pronouns, verbs, adjectives, and adverbs in general and their functions in particular sentences.", "L.3.1b – Form and use regular and irregular plural nouns." The key words, such as "explain" and "form and use", indicate the associated DOK of these content standards.

Table II-8. Test Blueprint by Subject and Claim/Standard

Claim/ Standard	Proportion of items by claim			
	ELA	Mathematics	Science	HGSS
1	60%–65%	65%–75%	32%–37%	65%–75%
2	25%–30%	8%–12%	30%–35%	25%–35%
3	10%–15%	8%–12%	29%–34%	
4		8%–12%		

II.2.2. Test design. The KAP is going through a transition period after the adoption of new content standards. This transition process dictates the test design to be different across the establishing years. The test design plan, as indicated in Table II-7, is to have a nonadaptive design in 2015 and, in 2016, a three-stage adaptive design for ELA and mathematics and a three-stage nonadaptive design for HGSS and science. For the adaptive test, assignment of the Stage 2 block is determined by the ability estimates based on students' answers to Stage 1 items; similarly, ability estimates for Stage 2 are used to determine the block assignment in Stage 3. Because students' abilities are unknown at the beginning of the test, the Stage 1 block is set at a

medium difficulty level that meet the abilities of the majority of students. Field-test items stand alone in Stage 4. Multiple field-test blocks are constructed and randomly assigned to students in order to obtain representative samples for each block.

Because of the inclusion of listening (Claim 3) in the ELA test, listening items are added to Stages 1, 2, and 3 as operational field-test items. Students respond to the same six listening items in the three stages. These items are analyzed after the testing window closes, and items with good statistics are selected to be operational items. Scores on these items are added to student total scores. Additional Claim 3 items are field-tested in Stage 4, but they do not count toward students' scores. An MDPT is given as a part of the HGSS test, but its score counts toward both ELA and HGSS scores.

The HGSS and science tests conform to the three-stage design; the intent is to provide a consistent "language" for use in the field when referring to the tests and to represent the volume of content students can be expected to complete in a single class period. Four parallel, equivalent test forms are created for HGSS. Items about historical vignettes are in one block and items about a primary source are in another block. The MDPT is a separate test administration occurring over two days. One of the four pre-equated test forms is used in the 2016 administration. The other test forms are slated for subsequent years or used as a breach form.

For each subject test, accommodations are provided to students with special needs (see chapter V. Inclusion of All Students). Each stage has a block of items designated for students who need accommodations. When review panels or accessibility experts determine that items are not appropriate for students with special needs, those items are modified. Accommodations are assigned to students who requested them during registration for the KAP assessment.

Table II-9 shows the test design of the KAP, and Table II-10 presents the number of blocks and block difficulty levels for each stage by subject. Note that HGSS has a small number of machine-scored items and an emphasis on the MDPT writing portion because it is initially intended to be a programmatic evaluation of curriculum and learning rather than a student achievement test.

Table II-9. Test Design for the KAP

		No. of items				
Subject	**Grade**	**Total**	**Stage 1**	**Stage 2**	**Stage 3**	**Note**
ELA	3–8, HS	55	25	15	15	Adaptive
Mathematics	3–8, HS	55	25	15	15	Adaptive
Science	5	60	25	5	20	Nonadaptive
	8, HS	70	25	20	25	Nonadaptive
HGSS	6	10	Vignettes	Primary	MDPT	Nonadaptive
	8	8	Vignettes	Primary	MDPT	Nonadaptive
	HS	11	Vignettes	Primary	MDPT	Nonadaptive

Table II-10. Number and Difficulty of Blocks for the KAP

	No. of blocks			Block difficulty		
Subject	Stage 1	Stage 2	Stage 3	Stage 1	Stage 2	Stage 3
ELA	1	3	4	medium	easy, medium, hard	very easy, medium easy, medium hard, very hard
Mathematics	1	3	4	medium	easy, medium, hard	very easy, medium easy, medium hard, very hard
Science	1	1	4 for grades 5–8; 5 for HS	medium	medium	medium
HGSS	1	1	1	medium	medium	medium

Note. HS = high school. For test security reasons, each stage must have multiple blocks. Stages that have only one block would have multiple blocks created by shuffling items of the block.

II.2.3. Operational test construction. Domain sampling refers to the selection of a sample of test items from a well-defined population of items (Crocker & Algina, 1986). Student performances on those sampled items are used to infer their abilities to the tested content area; therefore, the selection of items and item quality will affect the validity and reliability of student ability estimates.

Koretz (2008) noted a few important factors in assuring the generalizability of the test results: motivate respondents in testing, the wording of questions, and the content representativeness of sampled items. Content representativeness is the optimal goal of operational test construction. This goal is achieved by building a test matching with its blueprint. However, item quality, both in wording and item statistics, also plays an important role in test quality and is evaluated in each test construction process. Test construction process is similar across years. It starts with item screening: summarizing item pool quality from content and statistics aspects and identifying eligible items. The following item characteristics frame the item screening targets.

- Items and passages are approved by KSDE prior to field testing.
- Items are reviewed to eliminate enemies (e.g., items that might clue answers to other items).
- Items' psychometric characteristics meet the criteria.
- Items with the best slopes are preferentially selected.

Following are subject-related guidelines for 2016 test construction.

II.2.3.1 *ELA and mathematics test construction guidelines.*

- Different test forms include approximately the same number of items per claim.

- Different test forms include approximately the same number of items per target.
- Different test forms include approximately the same number of items per DOK.
- ELA and mathematics Stage 1 block includes a wide range of item difficulties and their average difficulty is of moderate level.
- Linking items should have robust item statistics and match the test blueprint. They are placed in Stage 1.
- In mathematics, the Stage 1 items are ordered from easiest to hardest within each claim.
- In ELA, passage-based items are ordered according to established protocol (i.e., starting with main idea and followed by specifics) and referencing the order of appearance in the text.
- In ELA, Claim 2 items are generally ordered from least difficult to most difficult.

II.2.3.2 *HGSS test construction guidelines.*

- Test forms include a wide range of item difficulties and their average difficulty is of moderate level.
- HGSS items about historical vignettes are ordered chronologically or, when no chronology is specified, in the order of the content units.
- HGSS items about a primary source items are ordered according to the rules (i.e., in sequence of source/purpose/audience, context, and content/conclusion) established by the CETE and KSDE.
- HGSS primary source and MDPT do not used the same source.

II.2.3.3 *Science test construction guidelines.*

- Different test forms use approximately the same number of items per claim.
- Different test forms use approximately the same number of items per target.
- Different test forms use approximately the same number of items per DOK.
- Each block includes a wide range of item difficulties and their average difficulty is of moderate level.
- Science items are ordered by claim.

II.2.4. Item pool evaluation. Because both ELA and mathematics use multistage adaptive tests, more items are consumed for the multiple blocks used in Stages 2 and 3. The number of quality items in the item pool are essential to the success of the design. This section addresses item pool quality from three perspectives: content alignment, item count by content standards, and simulation of paths to the different blocks of Stages 2 and 3.

II.2.4.1 *Alignment study of adaptive test item pool.* In fall 2014, edCount and CETE drafted a plan to investigate multiple facets of KAP items: the use of items, gaps between the expected and actual use of items, and alignment throughout the test-development process. After a yearlong effort, the *Kansas Assessment Program Alignment Evaluation Report 2015–2016* (the *Alignment Study* hereafter) was completed in July 2016.

The *Alignment Study* gathered a wide range of evidence to address the quality of items and performance tasks in association with test blueprints. The evidence included:

> item quality, alignment, coherence, and accessibility; blueprint quality and alignment; and test form alignment to targets and intended blueprints. This enhances item reviews of 2014-15 through the inclusion of additional items for review and the addition of blueprint and test-level reviews, which provide evidence regarding the degree of alignment between the assessments and the claims and targets. (Forte el al, 2016, p. 1)

Different from the typical alignment studies that are designed for post hoc evaluation, edCount used Forte's (2013, 2016) framework and developed a process to include items from past administrations in the early evaluation stage and emphasized the alignment between item, blueprint, and content standards. In spring 2016, approximately 355 ELA and 234 mathematics items of the 2016 KAP operational test were reviewed by panels of content experts, who were instructed to evaluate whether each item actually clearly and accurately reflects its content target. This section summarizes the item pool results.

The edCount blueprint review panel, composed of four internal content and research staff members, used the internally developed protocols to assess the connections among the KAP KCCRS, the test blueprint, and item-bank metadata. The panel concluded that the item pool for all grades of both ELA and mathematics met the following requirements: at least six items addressed each claim on the blueprint, at least one item slot in the blueprint was assigned for each target in the content emphasis document, and the percentage of items addressing each claim met expectations. Because KAP did not have DOK blueprints, averages of DOK by target were computed. ELA DOK was 2.4 for all grades; mathematics DOK ranged from 2.4 to 2.6. The values indicated there were more items in the level 3 DOK (higher cognitive complexity). Additionally, evaluation of operational pathways and items indicated that each pathway adhered to the blueprints, and operational items reflected the breadth and depth of the KCCRS.

When selecting items for the adaptive test, CETE uses stage approach in that the test blueprint is divided into three stages and blocks within each stage are parallel forms. This way all possible combination of adaptive routings will yield test forms that match the blueprint.

II.2.4.2 ***Item count by content standard.*** Table II-11 presents the number of items and proportion by claim in the ELA and mathematics item pools for the 2016 test construction. The proportions of ELA and mathematics items by claim align with their test blueprints. Table II-11 also shows that the ELA item pool contains only items for Claims 1 and 2. As mentioned in section I.4 Content Standards for Assessed Grades, listening has been included in the ELA claim since the 2016 administration; its items were not included in the 2015 tests. Because the item pool was composed of items from past administrations, listening items were not included. Instead, listening items were developed and placed in the 2016 tests as operational field-test items.

Table II-11.Percentages of Item by Claims

Grade	No. of ELA items	ELA claims (%)		No. of math items	Mathematics claims (%)			
		1	2		1	2	3	4
3	329	63	37	244	69	11	10	9
4	316	64	36	266	70	11	9	10
5	325	62	38	243	70	13	9	8
6	323	64	36	213	69	13	9	9
7	314	61	39	246	66	14	11	9
8	299	61	39	250	73	10	7	10
10	326	62	38	251	71	10	11	8

II.2.4.3 *Item statistics.* The ELA and mathematics tests use a 1-3-4 design and should yield 12 possible pathways. Block difficulty levels and possible pathways are presented in Table II-12. Note that the table shows only 10 pathways because the Stage 2 *easy* block cannot be routed to the Stage 3 *very hard* block, and the Stage 2 *hard* block cannot not be routed to the Stage 3 *very easy* block.

Table II-12. Pathways of Multistage Design for ELA and Mathematics

Pathway	Stage		
	1	2	3
1	Medium	Easy	Very easy
2	Medium	Easy	Medium easy
3	Medium	Easy	Medium hard
4	Medium	Medium	Very easy
5	Medium	Medium	Medium easy
6	Medium	Medium	Medium hard
7	Medium	Medium	Very hard
8	Medium	Hard	Medium easy
9	Medium	Hard	Medium hard
10	Medium	Hard	Very hard

Simulations for the multistage adaptive test were performed during test construction. As mentioned earlier, Stage 2 and Stage 3 block assignments depend on students' performance in the previous stage. The algorithm used to determine block assignment is the test information function (block information, in this case) of item response theory (IRT). The block that provides the most test information is selected for administration. For example, Figures II-1 and II-2 show Stage 2 and Stage 3 block information function curves. Figure II-1 shows that the easy block (difficulty level 1) has a larger information function at theta range (−4.0, −1.4) than the other two blocks, the medium block has larger information function at range (−1.4, 0.6); the hard block has larger information at range (0.6, 4.0). If a student's Stage 1 theta estimate is 0.4, the medium block will be administered because it has larger information function for this theta value.

Due to item-bank limitations, items that can be used in each block are limited and cause the information function of some blocks to be too low across the theta range when compared to other blocks. For example, the Stage 3 block information function curves presented in Figure II-2 show that the theta ranges of the four blocks by difficulty levels are (−4.0, −1.5); (−1.5, −0.5); (−0.5, 1.4); and (1.4, 4.0), respectively. The Stage 2 easy block of this example will not be routed to the Stage 3 *medium hard* block, and the Stage 2 hard block will not be routed to the Stage 3 *medium easy* block. This situation occurs in most ELA and mathematics grades. Especially, the grade 7 ELA Stage 2 information function of the medium block is lower than that of the other two blocks and will not be assigned. The results of pathway routing for ELA and mathematics can be found in Appendix D, Path Reliability.

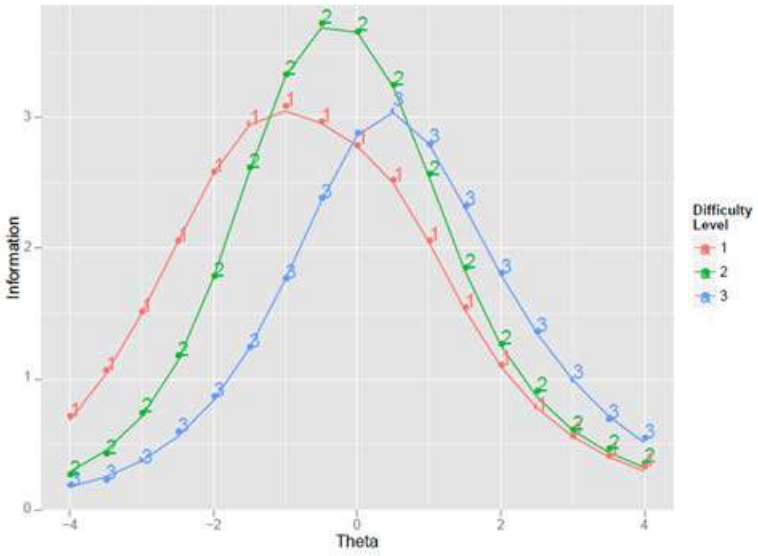


Figure II-1. An example of Stage 2 block information curves.

Note: Difficulty levels 1, 2, and 3 refer to the easy, medium, and hard blocks, respectively.

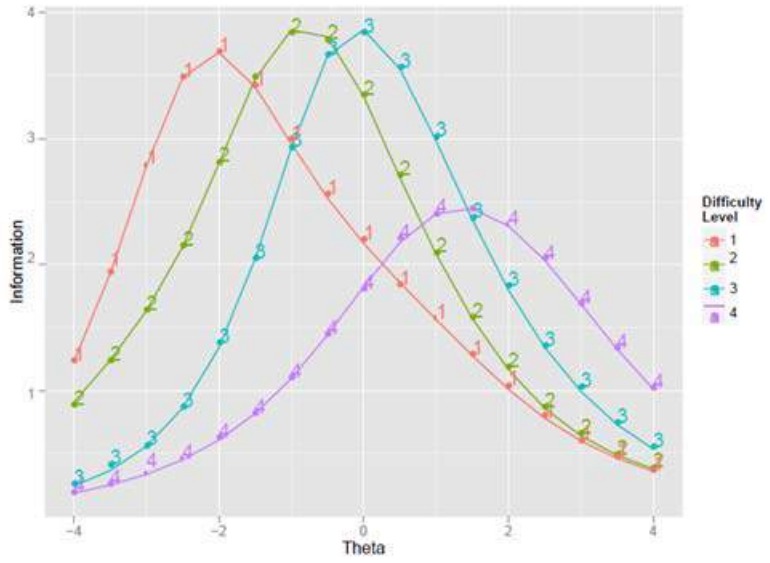


Figure II-2. An example of Stage 3 block information curves.

Note: Difficulty levels 1, 2, 3, and 4 refers to the very easy, medium easy, medium hard, and very hard blocks, respectively.

II.3. Item Development

Item development entails various efforts to ensure item quality, including research into best practices, drafting subject-area item specifications, preparing materials for item-writer training, recruiting item writers, conducting item-writer training, creating items, and reviewing items. Item review is conducted in two phases: first, when items are created, and next, after items are field tested. In the first phase, items are reviewed by both CETE content experts and trained, external item reviewers. Before appearing on any assessment, items are reviewed by content reviewers, bias and sensitivity reviewers, and KSDE staff. CETE staff use item-review feedback to revise test items as needed. Items are then prepared for field testing, according to test specifications. After field testing, item and test data are analyzed; this data analysis guides decisions about the use of items on future assessments. The following section describes typical procedures for different stages of item development.

II.3.1. Passage selection and review. ELA passages are selected before item writers are recruited because of the intense and time-consuming efforts required in identifying and acquiring quality passages. The ELA team uses several resources to analyze text complexity and guide grade-level placement. Assessment passages include commissioned, permissioned, and public domain readings. Passages from all sources undergo multiple rounds of review, including editorial, content, bias, sensitivity, and accessibility reviews. For example, KSDE accessibility specialists and CETE content specialists review passages for accessibility issues and content accuracy (e.g., inaccurate or outdated science information).

Passages that are accepted at the internal review are then reviewed by an external panel of educators. These external passage-review panels are formed by grade band: grades 3–5, grades 6–8, and high school. Each panel includes educators with backgrounds in ELL and special education. Because passage reviews are processed through a secure, online reviewing system, reviewers can follow detailed instructions to review passages at their own pace and provide feedback or placement recommendations by a given deadline.

Passage reviewers use rubrics of both qualitative and quantitative measures to examine text complexity and grade-level suitability through text structure, language features, and knowledge demands. CETE uses the Flesch-Kincaid score as a quantitative measure for longer passages. However, passages of only 350–450 words are not long enough to give an accurate Flesch-Kincaid reading. In those cases, CETE considers sentence length and complexity to gauge an initial grade placement. Qualitatively, CETE looks at each set for vocabulary, knowledge demands, topic familiarity, and interest level.

In addition, both internal and external reviewers consider the following passage components.

- Length: Are the texts of reasonable length for students? Are they long enough and rich enough to support a number of items?
- Bias or sensitivity: Are all groups portrayed accurately and fairly? Does the passage demonstrate awareness of different cultures and sensitive topics in the state (e.g., natural disasters, politics, moral values)?
- Overexposure: Is the passage topic already commonly taught in the school or district, or is it used frequently in anthologies or lesson plans?
- Interest level: Will more than half of students be at least moderately interested in the passage?
- Images: Are there any concerns related to the accessibility or content of images? Should images be added to enhance or support the passage?
- Other: For example, should introductions be included to provide historical context or background information?

Reviewers are then asked to recommend a grade for each passage, based on complexity and other considerations. After the passage-review window is completed, reviewers are invited to an optional telephone discussion of the passages. After compiling the information and summarizing the overall data collected from the review, CETE shares the results and passages with KSDE for approval of grade placement. Based on item pool needs (e.g., complexity levels, text types, topics), some passages are selected for item development. Remaining passages are held for future development.

II.3.2. Item writers. University of Kansas graduate research assistants (GRAs) who are trained in a given subject, who have prior item-writing, or who have teaching experience are recruited to be item writers. Because ELA, mathematics, science, and HGSS tests cover a wide range of knowledge and skills, GRAs who write items for the assessments major in a variety of academic

areas, including curriculum and teaching, English, mathematics, economics, pre-med, classical languages, biology, computer science, and earth and space sciences. Additionally, panels of educators and subject-area experts from outside of CETE convene to develop specialized, open-ended items like HGSS MDPTs and mathematics performance tasks.

II.3.3. Item-writing training. Before writing items for the KAP, item writers are trained in the use of KAP subject-area item specifications in the writing and reviewing of items. All item writers receive training in several topics, including

- the KCCRS,
- validity and reliability,
- alignment,
- differentiating between cognitive complexity and difficulty,
- evidence-centered design,
- principles of universal design and accessibility,
- bias and sensitivity, and
- item types.

To guide the item-writing process, item writers are trained in content, format, structure, stem structure, answer choice, accessibility, bias and sensitivity, etc. Besides learning fundamental principles in item writing, item writers also receive training in item review so they can objectively evaluate their own products as well as others' items. Key points of these guidelines are presented below.

II.3.3.1 *General guidelines.*

- Write items that have clearly correct answer choice(s), with other answer choices clearly incorrect.
- Ensure that items are clearly worded.
- Avoid the use of tricky or misleading items.
- Proofread items for correct grammar, punctuation, and spelling.
- Avoid the use of contractions.
- Use third-person perspective.
- Avoid the use of humor.

II.3.3.2 *Content guidelines.*

- Write items to appropriate content standards.
- Ensure that multiple-choice items measure a single concept.
- Ensure that items focus on important ideas, not trivia.
- Use vocabulary that is consistent with students' grade level.
- Align items to the cognitive complexity of content standards.
- Write items to a variety of difficulty levels.

II.3.3.3 *Format guidelines.*

- Format answer choices vertically rather than horizontally.

- Ensure that items include enough white space and are not cramped.
- Create clear layouts.
- Write clear instructions.

II.3.3.4 *Structure guidelines.*

- Avoid complex-format items.
- Write items in the form of a question.
- Avoid window-dressing of items (e.g., excessive verbiage).

II.3.3.5 *Stem construction guidelines.*

- Write stems positively whenever possible.
- Avoid asking for and expressing opinions in stems.
- Ensure that the central idea is in the stem.
- Place the question as close to the answer choices as possible.
- Minimize the use of qualifying words (e.g., always, never).

II.3.3.6 *Answer-choice development guidelines.*

- Order answer choices logically.
- Create independent answer choices that do not overlap.
- Write answer choices that are roughly of the same length and parallel in structure.
- Do not offer all of the above, none of the above, or I don't know as answer choices.
- Avoid cluing between the stem and answer choices.
- Avoid specific determiners like always or never.
- Create plausible distractors.
- Create distractors that take advantage of common errors and misconceptions.
  Answer keys should be roughly uniform in distribution.

II.3.3.7 *Accessibility guidelines.*

- Consider the access needs of special populations and how accommodations affect an item's intent.
- Use simple sentence structures.
- Minimize the use of words with multiple meanings.
- Avoid the use of slang and regional dialect.
- Avoid the use of complicated names or names that could be confused with other nouns.
- Clearly label graphics.

II.3.3.8 *Bias and sensitivity guidelines.*

- Avoid the use of stereotypes.
- Consider the regional and cultural nuances of words.
- Avoid the use of demeaning or offensive materials, particularly in the stimulus.
- Avoid the use of offensive or religious references.
- Ensure that items are not related to socioeconomic status or family attributes.
- Use artwork that reflects the diversity of the student population.

Item-writing training also includes extensive practice. Participants discuss DOK for specific standards, examine practice items for alignment to content standards, and determine whether practice items are written to the appropriate difficulty level. Participants also practice writing items and receive feedback from CETE staff.

Specialized training for the HGSS and mathematics performance tasks is provided to content experts. For example, HGSS content experts identify an array of primary and secondary source documents and create sets of documents (or excerpts) from these sources to serve as the basis for both the items about the primary source and the MDPTs. These experts also write the writing prompts for each set of documents identified and the Document Focus questions. After item prompts are completed, CETE GRAs and content staff starts writing items for the prompts.

II.3.4. Item writing. CETE relies on teacher expertise throughout the item development process. While all items are written according to these guidelines, special care is taken with HGSS MDPTs. CETE develops sets of MDPT per grade level (3–5, 6–8, high school) for teacher review. Each MDPT set contains two to three passages, a graphic, and several writing prompts. The writing prompts are developed into one of the three categories: informative/explanatory, argumentative/opinion, and narrative. Then CETE brings in several teachers and state-level writing specialists to review and comment on these initial sets, and teachers make several recommendations that guide later development. One suggestion is to provide students in lower grades with more prompts to finish a story (rather than write a story on their own). Teachers also suggest that familiar topics be used in the lower grades.

Teacher input is incorporated into MDPT writing prompts; therefore, at the lower grades, the informative/explanatory and argumentative/opinion sets use familiar topics and straightforward prompts. At the upper grades, informative/explanatory and argumentative/opinion prompts are often based on sets with less-familiar topics that required students to think more critically about their answers. Due to time constraints and younger students' less-developed typing skills, narrative prompts at the lower grades ask students to finish a story rather than write an entire story on their own. In contrast, some middle-grade narrative prompts ask students to develop an entire story, with a sample story provided in the set.

II.3.5. Item reviewers. The item-review process involves several stages.

- Internal content review
- Psychometric review
- Accessibility review
- Editorial review
- KSDE review
- External content review, using multiple panelists
- External bias and sensitivity review, using multiple panelists
- Internal content team resolution, in consultation with KSDE

CETE content experts and KSDE staff recruit two types of item reviewers from Kansas educators: content reviewers and bias and sensitivity reviewers. Prospective item reviewers complete an online survey in which they indicate their demographic information, teaching experience, professional qualifications, content expertise, knowledge of the standards, and special education or ELL endorsements or training.

Content-review panels for ELA and mathematics are formed by grade band: grades 3–5, grades 6–8, and high school. Content-review panels for HGSS and science are formed by grade, but some reviewers serve on more than one panel since domain content knowledge extended above or below grade levels. Bias and sensitivity panels are assembled and include members of various minority groups. Similar to the passage review process, item reviews are processed through a secure, online reviewing system. After completing a web-based training session, reviewers go through items at their own pace and provide feedback by a given deadline.

II.3.6. Item review. All item reviewers must complete two web-based sessions of item-review training: bias and sensitivity training and content-review training. The training sessions include information about the KSDE–CETE partnership, test and item security, item-writing guidelines, and the item-review process. Item-review training also provides participants with practice items and CETE staff contact information.

Bias and sensitivity reviewers are given a code sheet that provides code categories and descriptions for possible concerns. When reviewers flag items for bias or sensitivity concerns, they use codes to provide details. A code is also assigned to indicate there is no barrier, bias, sensitivity, or other concerns for clarity and record keeping purposes. Descriptions of concerns are given below.

- Possible bias related to gender, race or ethnicity, socioeconomic factors, or other
- Possible barrier related to uncommon or unfamiliar language, linguistic complexity or lack of clarity, assumed prior knowledge, cultural restrictions, accessibility, or other
- Possible sensitivity concern related to stereotype, religion, socioeconomic factors, status, specific topic, or other
- Other concern

Content reviewers also attend to the alignment of items to assessment targets, checking that items adequately address part of the target and elicit evidence for at least part of one evidence statement. Furthermore, content reviewers check items for

- appropriate, grade-level vocabulary;
- a clear, complete statement or question;
- grammatically correct text;
- a correct key;
- accurate, relevant graphics; and
- well-designed answer choices that do not require background knowledge outside of the content area and that are free from clang associations. (Clang occurs when words from an

item's stem appear in one or more response options.)

Based on their analysis of items, reviewers advise that items be accepted, revised, or rejected, and give specific reasons for their decisions (e.g., "item does not align").

II.3.7. Universal design in test development. Universal design (UD) in item and test development not only allows for the participation of the widest range of students, but it also should bolster the validity of score inferences. KAP's comprehensive inclusion rules mean that KAP tests include virtually all Kansas students. While initially intended to meet the interests of special-needs students, the benefits of universally designed assessments should apply to all students with diverse characteristics.

Item-writer training teaches participants about UD concepts, including a definition of UD and examples of test items that adhere to UD principles. Additionally, the item-writer guidelines include many UD principles. The following are some focuses of UD in the KAP's development.

- Item writers are trained to become aware of and sensitive to issues of cultural and regional diversity.
- Both internal and external reviewers of items and test specifications strive to ensure that no barriers stem from a lack of sensitivity to ability, culture, or other characteristics.
- The tests are developed to be compatible with many accommodations and a variety of widely used adaptive equipment and assistive technology without changing the meaning or difficulty of test items.
- The language used in test materials is direct and concise. Additionally, unnecessary images and text are omitted to avoid distracting students.

II.3.8. Field testing. In general, field testing of new KAP items uses the embedded-model approach. ELA and mathematics 2016 field test uses a separate, non-adapting block of 15 items to field-test machine-scorable items. Multiple blocks of field-test items are developed to supplement the adaptive item pool. Each field-test block mirrors at least one of the operational blocks in structure and domain coverage in order to mask its identity as the field-test block. This step is critical in ensuring the main advantage of an embedded field test: Examinees cannot differentiate items that count toward their score from field-test items, thereby using the same care to answer the field-test items. This trait improves the field-test item data quality and provides more robust item-parameter estimates.

Because HGSS is assessed only in even-numbered years, a stand-alone field test is administered in an off year to provide the item statistics needed to construct operational tests. For example, the items used to build the spring 2016 tests were field-tested in spring 2015. The spring 2015 HGSS administration had only field-test items and consequently HGSS scores were not used for accountability or reported to the field. The 2015 field testing also included the writing portion of the HGSS assessment, MDPTs.

II.3.9. Field-test data analysis. Field-test item analyses include classical item analysis, IRT calibration, model fit, and differential item functioning (DIF). Items that are too easy or too difficult, that do not discriminate ability well, that fail to meet the IRT model, or that have large DIF are flagged according to predetermined criteria. The statistics and flags are added to the item pool for use in test construction. Note that because this report focuses on the quality of operational items, field test statistics are not presented.

II.3.10. Data review. Following field-test item analyses and prior to test construction, the content team reviews item statistics. Items with statistical flags are used only when the item pool does not have other items for blueprint coverage. When flagged items are used, they undergo extra review and discussions.

II.4. Test Administration

Large-scale assessment requires a standardized test-administration process to prevent the unintended effects of administration differences. The standardized test-administration procedures are described in the *Examiner's Manual* and *Tools and Accommodations for the Kansas Assessment Program (KAP) 2015–2016* (hereafter *Tools and Accommodations*). The *Examiner's Manual* provides information on standardized test administration for districts, schools, and teachers; *Tools and Accommodations* provides guidance on the use of available accessibility tools and features for assessments. Teachers who administer the KAP assessment are required to sign an agreement to follow the guidelines and to show that they have learned about test security and ethical test practices.

In the *Examiner's Manual*, test security procedures are described in multiple sections. "Test Security Plan" and "Test Security Guidelines" are found in "Section 2: Test Coordinators." "Test Security and Administration" is in "Section 4: Teachers." The Test Security Guidelines section of the *Examiner's Manual* explicitly explains to the district coordinator the test security practices and actions after detected test security breach, loss of materials, or any other deviation. A breach form is constructed for each subject and grade. In the case of a major test breach, the breached test will be removed and replaced with the breach form.

II.4.1. Test administration and security training. All Kansas district coordinators must take the test administration and security training during the preconference in October KSDE Annual Conference or with online training materials available from the KSDE assessment website. District coordinators will train building-level personnel before the local test. All local personnel administering state assessments must read the *Examiner's Manual* and sign an agreement to abide by state ethical testing practices. See Appendix A for the training PowerPoint.

II.4.2. Monitoring test administration. District and building test coordinators can monitor student test progress via the Kansas Interactive Testing Engine (KITE) Educator Portal. The *Examiner's Manual* describes the process on page 33.

During the testing window, KSDE staff and members of the Kansas Assessment Advisory Council visits 5% of Kansas schools at random to monitor administration and test security. The

State Monitor Quality Assurance Checklist for Test Security and Ethics is posted on the KSDE website, along with other assessment-related documents and resources to assist districts and schools in understanding the KAP administration.

Provision of accommodations is handled in two ways, by test administrators and by the online test portal. Information of accommodations handled by test administrators are not available. Accommodations build-in the online testing portal are reported at section V.3 Accommodations. Evaluation of the consistency between the accommodations included in the individual education plan (IEP) and during the assessment cannot be conducted because the IEP information is not available.

II.5. Systems for Protecting Data Integrity and Privacy

The electronic item bank, online administration system, and student responses are stored at the KITE® Suite designed and maintained by the Agile Technology Solutions (ATS) of the Achievement & Assessment Institute (AAI). Multiple portals are designed within the KITE® Suite to serve the needs for item and test development (i.e., Content Builder), for educators to input and access test and student information (i.e., Educator Portal), and for online testing (i.e., Kite Client).

AAI fully understands the importance of test security in both protecting student information and ensuring valid interpretation of test data. The physical security requirements are met by using hosting providers that conform to SAS 70 auditing standards for physical access and PCI compliance. Most of the project management, test development, and data analysis activities take place at CETE. CETE's on-campus offices are in a secure wing that can only be accessed with a key. ATS's off-campus offices are accessible only with an electronic key card. In general, most work is done at one of our sites using secure server systems. CETE and ATS staff access those servers via a secure VPN connection when they need to work remotely.

All KITE applications handle educator and administrative passwords using industry-standard encryption techniques; users must create strong passwords and may change their own passwords at any time in accordance with the password policy. All applications generate access records that can be reviewed by system administrators to track access. All released items exist in a separate pool from items used for summative purposes, ensuring that no items are shared among secure and non-secure pools. Only authorized users of the KITE assessment system have access to view items.

In accordance with FERPA, students', teachers', operators', and administrators' access to personal student data is limited to student records in which that person has a legitimate educational interest. All users are provided the minimum amount of access necessary. Throughout the lifetime of the product, security levels, groups, and access will be reviewed periodically to ensure continued compliance.

Operational access to all servers is controlled by keys that are provided only to system

administrators who manage the production data center in the operations team. Access to the networking equipment and hardware consoles is limited to the data center itself; remote access to these devices is limited to the data center-specific administration host.

Access to individual KITE applications is controlled according to the policies set forward for that application and the data the application maintains. All access policies and accounts are reviewed periodically to ensure that access to systems is limited to the appropriate populations.

In addition to physical and electronic security measures, test security is promoted through required training and certification requirements for test administrators. Test administrators are expected to deliver assessments with integrity and to maintain the security of assessments. State, district, and school users are expected to complete the security agreement within Educator Portal each year. By accepting the security agreement, users agree not to store or save assessment materials to computers or personal storage devices, to not print, and to not share personal passwords with others.

III. Technical Quality—Validity

As defined in the *Standards for Educational and Psychological Testing* (the *Standards* hereafter), validity refers to "the degree to which evidence and theory support the interpretation of test scores for proposed uses of tests" (American Educational Research Association [AERA], American Psychological Association, & National Council on Measurement in Education, 2014, p. 11).

The *Standards* (AERA et al., 2014) provide a framework for describing the sources of evidence that should be considered when evaluating test score validity. These sources include evidence based on (a) test content, (b) response processes, (c) internal test structure, (d) relationships between test scores and other variables, and (e) consequences of testing. Other sources of evidence also can bolster the validity argument. For example, when IRT is used to analyze data, validity considerations related to the use of IRT should be explored. When cut scores are critical to the interpretation of test results, the procedural validity of the processes used to establish those scores also should be addressed. The validation process involves the ongoing collection of a variety of evidence to support the proposed test score interpretations and uses. Much of this technical manual describes aspects of the KAP tests that support KAP test score interpretations and uses.

III.1. Overall Validity, Including Validity Based on Content

Because the intended uses of the test scores are the targets of a validity study, the purposes of the test should be identified before providing evidence to support test validity. The purposes of the KAP described at the beginning of this manual include (a) measuring specific claims related to the KCCRS, (b) providing Annual Measureable Objectives for state accreditation, (c) reporting student's academic performances, and (d) using with local assessment scores to assist in improving education program in the four subject areas.

Evidence on content validity, alignment study, cognitive process, and internal structure supports the use of KAP test to measure the KCCRS content as defined in the test blueprints. Information of test reliability, fairness and accessibility, and scoring and scaling justify the use of KAP test scores for Annual Measureable Objectives and reporting student's academic performances. Validity with other topics, such as using KAP scores to predict ACT scores, used local assessments to assists educators.

III.1.1. Content validity. Evidence of content validity for the KAP depends on the alignment between KAP items and the KCCRS and between test and test blueprint. The following procedural steps are used to evaluate the content validity of the KAP.

- Evaluate the alignment between KAP items and KCCRS.
- Evaluate the degree to which KAP test blueprint represent and align with the knowledge and skills described in the KCCRS.
- Conduct content reviews of KAP items using a panel of content experts to see if the items measure the intended construct or if sources of construct-irrelevant variance exist.

- Conduct fairness reviews of KAP items to avoid bias and sensitivity issues related to specific subpopulations.

The first two chapters of this technical manual present validity evidence related to test development and alignment study. As described in those chapters, all KAP items are developed and aligned with the KCCRS, and item development followed well-established procedures. After items are developed, they undergo multiple rounds of content and bias reviews. After field-test administration, items' statistical properties are reviewed. Those items pass content, psychometric, and KSDE reviews before selection for operational use. Tests also are administered according to standardized procedures, with accommodations for students with special needs. Specific efforts to ensure content validity are summarized below.

- Webb's (1997) Depth of Knowledge model is used to identify the cognitive complexity of KAP items, ensuring that items cover different cognitive complexities. Although DOK distribution is not specified in the test blueprint, description of content standards as the L.3.1, L.3.1a, and L.3.1b presented in II.2.1 Test Blueprints provides a direction of the expected DOK. Item writers used it to write items that matched with the DOK expectation of each content standards. The analyses of DOK distributions by subject and grade are presented in Tables IV-7 through IV-9.
- Qualified item writers are selected and trained to ensure they write high-quality items.
- Detailed item- and passage-development guidelines are established and used to train item writers, who also participate in guided item writing.
- CETE content specialists and editors review each new item to make sure all items align with the KCCRS; they also consider grade-level appropriateness, depth of knowledge, graphics, grammar and punctuation, language demand, and distractor reasonableness.
- Content committees composed of Kansas educators then review items and consider, among other elements,
    - overall quality and clarity,
    - KCCRS alignment,
    - grade-level appropriateness,
    - difficulty level,
    - depth of knowledge,
    - appropriate sources of challenge (e.g., item difficulty is not related to unintended content or skills),
    - answer correctness,
    - quality of distractors,
    - graphics,
    - appropriate language demand, and
    - absence of bias.
- An external bias, fairness, and sensitivity committee reviews items for issues related to diversity, gender, and other factors.
- Before items are selected for operational use, several statistical analyses are conducted, including classical item analysis, distractor analysis, and DIF. CETE staff again carefully

review items' statistical characteristics.

- Administration of the KAP tests is standardized and includes accommodations. Students are given ample time to complete the tests to avoid speededness issues).

III.1.2. Alignment study recommendations. Section II.2.4.1 Alignment study of adaptive test item pool of this manual describes the alignment of operational items to the test blueprints. The *Alignment Study* conducted by edCount provided satisfactory results; however, test improvements can still be made. Below are edCount's recommendations for KAP.

III.1.2.1 *Item recommendations.*

- Improve clarity between the targets for identified subject and grade levels.
- Revise grade-10 ELA passages that have lower agreement among panelists on the appropriateness of the reading passage.
- Review feedback on audio pacing and clarity for listening passages.
- Ensure that accommodations are provided for students who cannot read Braille and students who do not use sign language.

III.1.2.2 *Performance task recommendations.*

- Review the intended targets for each component within the mathematics performance tasks.
- Identify a primary claim and target for each performance task as a whole.
- Have panelists holistically review the DOK of the mathematics performance tasks; researchers believe the DOK of the mathematics performance tasks as a whole will be higher than the DOK levels of each component.
- Specify the intended targets for ELA performance tasks; currently, the ELA performance tasks are written to the claim level.

III.1.2.3 *Blueprint recommendations.*

- To provide greater context for both mathematics and ELA blueprints, include DOK values, item types, points, and range of proportions of content emphasis's high, medium, and low levels by target.
- Use a DOK range, rather than a maximum, for each of the targets on the content emphases document.
- For mathematics claims 2–4, provide more specificity on the blueprint to avoid oversampling targets.

III.1.2.4 *Test-level recommendations.*

- Visit those ELA grades in which the operational form's target distribution did not reflect blueprint or the intended emphasis of a particular target.
- Consider using additional item types for certain mathematics and ELA targets.

III.1.3. Item pool for adaptive test. Similarly, the item pool analyses described in section II.2.4.2 Item Count by Content Standards of this manual show that each claim has an adequate

number of items to cover test blueprints for all subjects and grades.

III.2. Validity Based on Cognitive Process

Response-process evidence examines the extent to which the cognitive skills and processes students use to answer an item match those targeted by item writers. While studies that investigating students' cognitive processes, such as read-aloud, are not planned, alternative evidence can be provided during item development process and performance level descriptions (PLDs).

During item development process, items were written by content experts who have been trained on proper item writing approaches then reviewed by content experts who had direct experience with students. As mentioned earlier, the content standards are written to cover both DOK and content specification. The DOK component guided (a) item writers to use language that soliciting the cognitive process required by the content standards, (b) item reviewers to evaluate the cognitive process required by items.

The performance level descriptors are also written to reflect the cognitive process required for the specific content area. For example, the PLDs of HGSS Ancient World History presented in Appendix F has the three statements in the second row:

- Level 2; Students can recognize the role of politics and power in the choices, consequences, rights, responsibilities, beliefs, ideas, diversity, continuity, change, and dynamic relationships in the spread and/or transformation of civilizations and cultures.
- Level 3: Students can recognize and analyze the role of politics and power in the choices, consequences, rights, responsibilities, beliefs, ideas, diversity, continuity, change, and dynamic in the spread and/or transformation of civilizations and cultures.
- Level 4: Students can evaluate the role of politics and power in the choices, consequences, rights, responsibilities, beliefs, ideas, diversity, continuity, change, and dynamic relationships in the spread and/or transformation of civilizations and cultures and can trace cultural and ideological transformation across societies.

At level 2, the cognitive process of "recognize" is required; at level 3, "recognize and analyze" are required; and at level 4, "evaluate" is required. The PLDs were written and reviewed by content experts and educators.

III.3. Validity Based on Internal Structure

As described in the *Standards* (AERA et al, 2014), internal-structure evidence refers to "the degree to which the relationships among test items and test components conform to the construct on which the proposed test score interpretations are based" (p. 13). For each KAP test, one total test score and several claim scores are reported. Multiple sources provide internal-structure evidence relating to the use of both types of scores are presented.

III.3.1. Internal construct. Item-test correlations (indicators of item discrimination) are reviewed in this manual in section IV.3.1 Classical Item Statistics. The range of acceptable

correlations for adaptive tests is broader than nonadaptive tests because extreme easy and difficult items are included to provide better theta estimates on the two ends of the scale. These extreme easy or difficult items tend to have low discrimination values because of attenuated by their difficulty. The only rule is not to use items having negative correlation. Summary of item discrimination presented in section IV.3.1 show that all items have positive correlations and the lowest correlation of HGSS, the nonadaptive test, are above .21.

III.3.2. IRT and model assumptions. The KAP ELA and mathematics tests are analyzed using IRT. IRT is an industry standard for item analysis in large-scale K–12 assessment programs because of its item and person invariance claims. However, it has several model assumptions that need to be fulfilled: model fit, unidimensionality, and local independence. The resulting inferences from any application of IRT depend on the degree to which the underlying assumptions are met. HGSS does not require equating, however, because this is its first year tested using the KCCRS. The classical test theory is used for its item analyses and scaling.

The current section introduces the IRT models and calibration procedures used for ELA and mathematics. Evaluation of IRT assumptions is presented as evidence of the appropriateness of model selection and is part of score validity.

III.3.2.1 *Samples.* The 2016 KAP ELA and mathematics tests use a three-stage adaptive design. Ten pathways are designed for the multistage adaptive design (see chapter II. Assessment System Operations). Each student takes a total of 55 items regardless the pathway. Initially, the equating is a pre-equating design: item parameters obtained prior to the current administration are used to estimate the thetas of this year's test. This method does not require re-estimation of item parameters using the current year's data; thus, it can be done prior to test administration and earns the name pre-equating. However, because some items do not have item parameters (e.g., listening items) or some items' 2015 parameters are not optimal (e.g., negative discrimination values), their item parameters must be estimated using current administration data. Because the equating must be done after those data are available, it is called post-equating. When post-equating is employed for the KAP, all items on the test that have proper prior parameters are used to anchor the items that need parameters.

The single-group concurrent calibration is used to place all item parameters across pathways onto the same scale. All operational items of the same subject and grade are compiled into one file to create a student-by-item data matrix, which is then fed into flexMIRT Version 2.80 (Cai, 2013) for concurrent calibration.

The student data file is cleaned prior to calibration and equating. For example, the estimation sample includes all students who completed all three stage blocks, except students who needed accommodations. It is expected that students with special needs perform differently from the general population and therefore could affect the common item estimates. Thus, records marked yes on accommodations are excluded in the calibration samples. Since each subject and grade is calibrated with a single-group concurrent method, the sample size for concurrent calibration

equals the number of valid cases after excluding certain accommodations. Table III-1 provides the number of students by subject and grade.

Table III-1. Sample Size for Concurrent Calibration by Grade and Subject

	Sample size	
Grade	ELA	Mathematics
3	34,229	34,010
4	32,840	32,666
5	32,067	31,904
6	32,165	31,931
7	32,818	32,568
8	31,930	31,670
10	33,204	33,195

III.3.2.2 *Missing data.* Missing responses require special attention because the coding of missing data can affect item parameter estimates. There are two types of missing responses: *omitted* and *not administered*. Omitted items appeared on the test but students did not answer them; thus, they are scored as incorrect answers (coded as *0*). Not-administered items did not appear on the test form students took but did appear on other test forms and are therefore coded as *missing*.

III.3.2.3 *Excluded items.* Few items are excluded during IRT calibration due to negative discrimination parameters and deemed by content experts to be bad items. Excluding those bad items lowers the possible maximum raw score of the form by 1 to 2 points.

III.3.2.4 *IRT models.* ELA and mathematics are calibrated with a two-parameter logistic (2PL) model (Birnbaum, 1968) and a graded response model (GRM; Samejima, 1969). The 2PL and GRM are applied to dichotomous and polytomous scored items, respectively. Model wise, the 2PL is a special case of GRM that handles dichotomous items. It defines the probability that a student of ability θ will answer item i correctly (u) as

$$P(ui = 1|\theta) = \frac{e^{[a_i(\theta - b_i)]}}{1 + e^{[a_i(\theta - b_i)]}}, \qquad \text{(III-1)}$$

where a_i is the discrimination parameter and b_i is the difficulty parameter. Discrimination indicates how well the item distinguishes between students with higher and lower levels of proficiency; difficulty is the degree of item difficulty on the same scale as theta.

Under the GRM, the probability that u_i is equal to any observed score category v equals the cumulative probability of scores 0 to $v - 1$, minus the cumulative probability of scores v to maximum score. The probability that the score is v or higher is

$$P(ui = v|\theta) = \frac{e^{[a_i(\theta - b_{iv})]}}{1 + e^{[a_i(\theta - b_{iv})]}}, \qquad \text{(III-2)}$$

where a_i is the discrimination parameter and b_{iv} is the difficulty parameter for score category v. One discrimination parameter is estimated for each item; this parameter may be interpreted as the strength of association between the item and theta. For m response categories, there are $m - 1$ GRM b parameters. The b for category v is interpreted as the point on theta where the probability of scoring in category v or higher is .5.

III.3.2.5 *Evaluating IRT assumptions.* The validity inferences from the IRT results depend on the degree to which assumptions of the models are met and on how well the models fit the data. The assumptions about IRT model fit, unidimensionality, local independence, and item-parameter invariance are evaluated.

III.3.2.5.1 IRT model fit. All operational items have been screened for statistical properties. One of the statistics checked is the model fit evaluation. Yen's (1981) Q_1 fit statistic is used to evaluate item model fit during field testing. The Q_1 fit statistic is the sum of the differences between the expected and observed numbers of students when conditioned by ability: the students are divided into 10 ability groups across the theta range, and the sum of differences is computed across the 10 groups. The Q_1 fit statistic for an item follows the chi-square distribution. Its degrees of freedom equal the number of groups minus the number of score categories for that item.

The chi-square-based item-fit statistic has one drawback: it is sensitive to sample size (Hambleton & Swaminathan, 1985). When the sample size is large, the chi-square test tends to over reject models because the statistical power is increased to the point that it is oversensitive to even very small discrepancies. On the other hand, the adaptive test causes Stage 2 and Stage 3 items to be restricted by theta range. The sample size, along with the restricted theta-range issues, might cause items to have significant Q_1 fit chi-square statistics although they actually fit the IRT model. Thus, the empirical judgment method based on the Q_1 fit graph was used to determine the model fit of an item. Figure III-1 is an example of a Q_1 fit graph of KAP item.

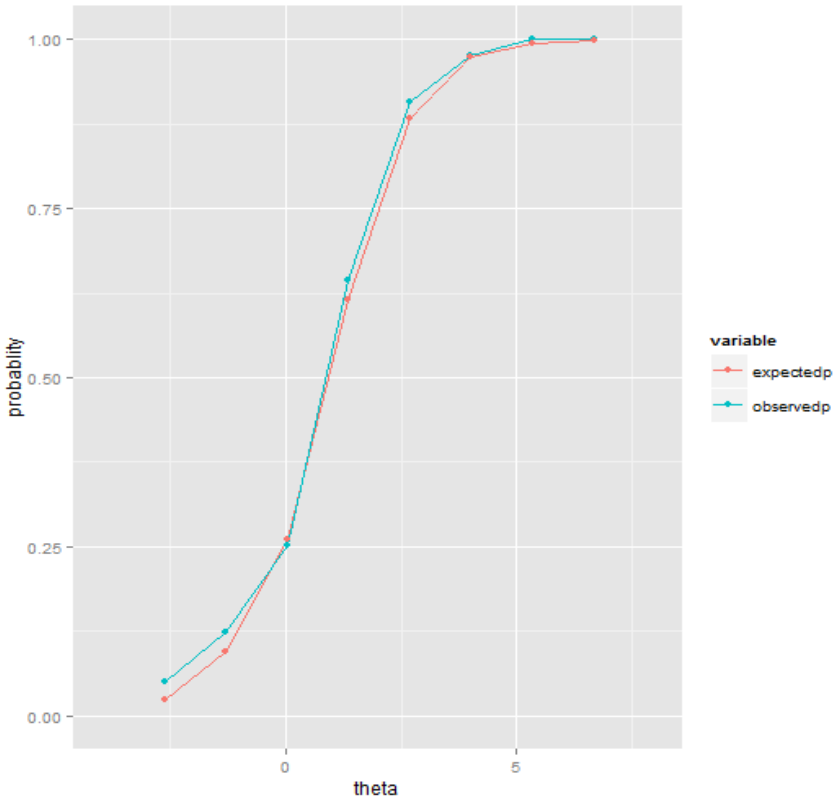


Figure III-1. Q_1 fit graph.

In Figure III-1, the blue line is the observed probability and the red line is the expected probability. The evaluation is conducted by gauging the discrepancy between the two lines. The small differences across the theta range indicate this item fit the model. Table III-2 summarizes the number of items that do not fit the 2PL or GRM based on the visual inspection. As mentioned earlier, some items are excluded from the operational item sets due to unsatisfactory statistics. The counts presented in the table include items used in the calibration only. The table indicates that, in general, ELA has more misfit items than mathematics does.

Table III-2. Number and Percentage of Misfit Items by Grade and Subject

	ELA			Mathematics		
Grade	Total items (*N*)	Misfit items (*n*)	%	Total items (*N*)	Misfit items (*n*)	%
3	144	32	22	134	15	11
4	135	23	17	135	14	10
5	131	32	24	129	20	16
6	132	12	9	120	15	13
7	111	24	22	120	19	16
8	116	10	9	122	17	14
10	129	21	16	143	14	10

Note. The total items include all items used in the calibration.

III.3.2.5.2 Unidimensionality. Both the 2PL and GRM assume that test items measure a single, dominant, latent variable. With an adaptive design, about two thirds of the ELA and mathematic items are assigned to low, mid, and high ability ranges and cause them to have restricted score range. Statistically, restricted range will cause underestimation of correlation; consequently, unidimensionality testing models that use correlation as foundation, such as principal component analysis and confirmatory factor analysis, are not appropriate in testing the dimensionality of adaptive tests. Due to the lack of proven statistical models to test the dimensionality of adaptive design, the IRT unidimensional and multi-dimensional model fit comparison is used to check test dimensionality. Additionally, a simulation study was conducted to evaluate the method.

The IRT model fit is implemented by fitting the data into a unidimensional and a bifactor IRT model that employs 2PL and GRM. To evaluate the results, the consistent Akaike's information criterion ([CAIC]; Bozdogan, 1987) is calculated. Lower CAIC values indicate better model fit. One of the characteristics of the CAIC is it penalizes complex models and favors parsimonious models (Rijmen, 2010). In other words, the CAIC will favor the unidimensional model due to fewer parameters in the model by default.

Tables III-3 and III-4 summarize the CAIC values of dimensionality model fit for ELA and mathematics. Among fourteen bifactor analyses, only two bifactor models converged (ELA grades 7 and 10). The failed-to-converge bifactor cases suggest the data are unidimensional. The CAICs of the two converged bifactor models are slightly larger than the CAICs of unidimensional models. Based on the assumption that the model with lower CAIC means the data fit the model better, the results indicate the two grades are also unidimensional. To support the conclusions of the comparison method, a simulation study was conducted.

Table III-3. Unidimensional-Model and Bifactor-Model Fit Indexes for ELA

Grade	Model	No. of parameters	CAIC
3	Unidimensional	306	2353702
	Bifactor	738	--
4	Unidimensional	289	2155144
	Bifactor	694	--
5	Unidimensional	287	2169534
	Bifactor	680	--
6	Unidimensional	304	2224146
	Bifactor	700	--
7	**Unidimensional**	**248**	**2030080**
	Bifactor	**581**	**2030886**
8	Unidimensional	286	2310065
	Bifactor	634	--
10	**Unidimensional**	**294**	**2338715**
	Bifactor	**678**	**2339555**

Note. When both unidimensional and bifactor models are converged, the rows are in boldface.

Table III-4. Unidimensional-Model and Bifactor-Model Fit Indexes for Mathematics

Grade	Model	No. of parameters	CAIC
3	Unidimensional	306	2350136
	Bifactor	738	--
4	Unidimensional	289	2720055
	Bifactor	694	--
5	Unidimensional	287	2428569
	Bifactor	680	--
6	Unidimensional	304	2367662
	Bifactor	700	--
7	Unidimensional	248	2705820
	Bifactor	581	--
8	Unidimensional	286	2529117
	Bifactor	634	--
10	Unidimensional	294	2370043
	Bifactor	678	--

The simulation study models after the ELA 8th grade design. A total of 32,000 true thetas are generated from a standard normal distribution [$N(0, 1)$], then apply to the unidimensional 2PL and GRM models to generate item responses using the empirical ELA grade 8 item parameters. The simulated responses follow the KAP adaptive rules in that the Stage 2 and Stage 3 blocks are assigned based on the interim theta estimate using the prior stage(s). Then simulated responses of all three stages of items are fitted into the unidimensional and bifactor IRT models that employed 2PL and GRM. This procedure is replicated one hundred times.

Investigation of item parameter recovery of the one hundred simulated data shows satisfying recovery outcome; which suggests the data reflect the empirical data. Out of 100 replications, convergences are observed for all unidimensional models yet only two bifactor models. This outcome suggest that when the data are unidimensional, the bifactor model will not converge in most cases. The few converged cases observed on the bifactor model are more likely caused by random chance. Both the empirical and simulation results indicate that ELA and mathematics tests meet the unidimensionality assumption.

III.3.2.5.3 Local independence. Local independence refers to the response to an item is not affected by responses to other items. This definition is necessary because it secures the foundation of the IRT model: the probability of answering an item correctly is affected only by the item's characteristics and student ability. If an item's response is affected by other items, then the IRT model cannot be used because it fails to incorporate the effects of other items. Local independence is violated when a group of independently scored items is written yet the responses of items in the latter positions depend on the responses of their predecessors. In this case, when the first item of the group is answered incorrectly, it will cause the answers to the remaining items to be incorrect.

Evaluation of local independence starts during item development. As long as all test items are written so that they do not depend on the responses of other items, local independence is assured.

During test construction, all items on a test are reviewed to ensure neither the items nor the answers clue students to other items on that test.

III.3.2.5.4 Invariance. IRT models claim that item parameter estimates are invariant up to a linear transformation for all examinees. Bivariate scatter plots and Pearson product-moment correlations are used to evaluate the relationship between the item parameters estimated from subgroups that are expected to have the same ability distributions. Due to multistage adaptive test design, Stage 2 and Stage 3 items are administered to students of a small range of abilities and skew the item parameter estimates. To avoid statistical bias caused by outliers, any items with discrimination parameters smaller than 0 or greater than 4, or with difficulty parameters greater than |6|, are excluded from the comparison. The invariance assumption is met if the estimated item parameters for female and male samples are highly correlated.

Here, the subgroups are determined by gender. The scatter plots presented in Figures III-2 through III-5 indicate that both ELA and mathematics, except for a few items, have strong linear relationships between item parameter estimates for female and male samples. The items with large discrepancies suggests potential gender DIF. Table III-5 shows that many of the Pearson correlations are above .90. These results strongly support the invariance assumption for the KAP ELA and mathematics, especially for item difficulty parameters.

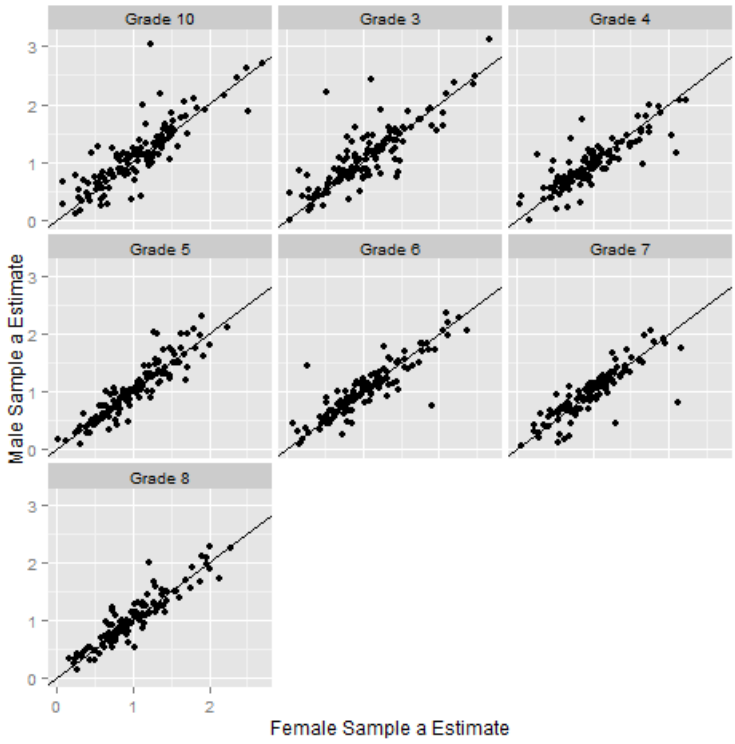


Figure III-2. ELA item discrimination parameter scatter plot by grade.

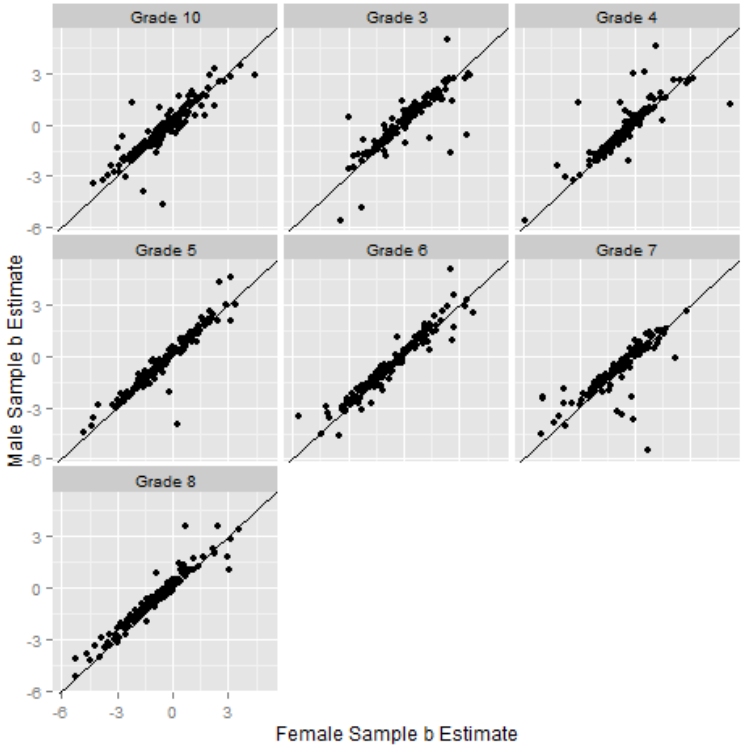


Figure III-3. ELA item difficulty parameter scatter plot by grade.

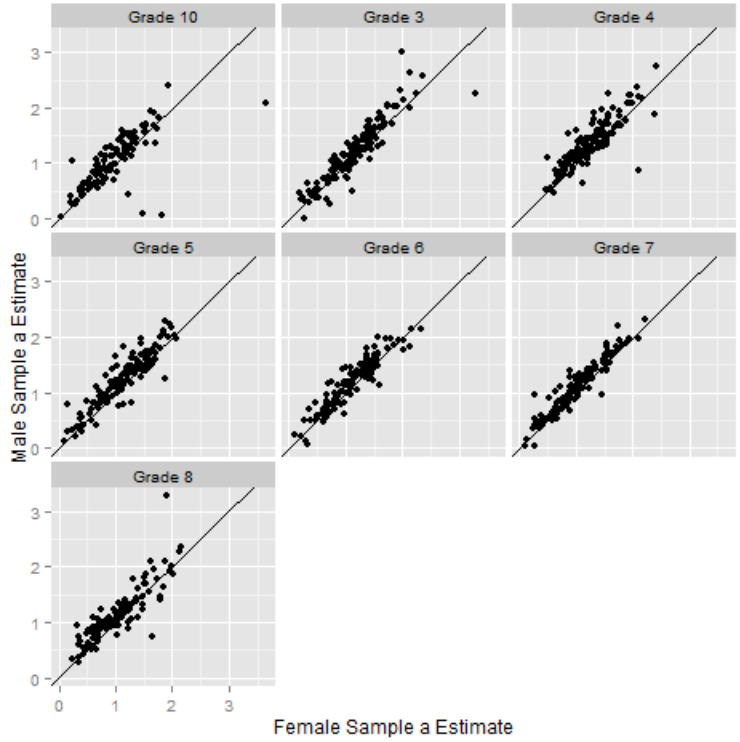


Figure III-4. Mathematics item discrimination parameter scatter plot by grade.

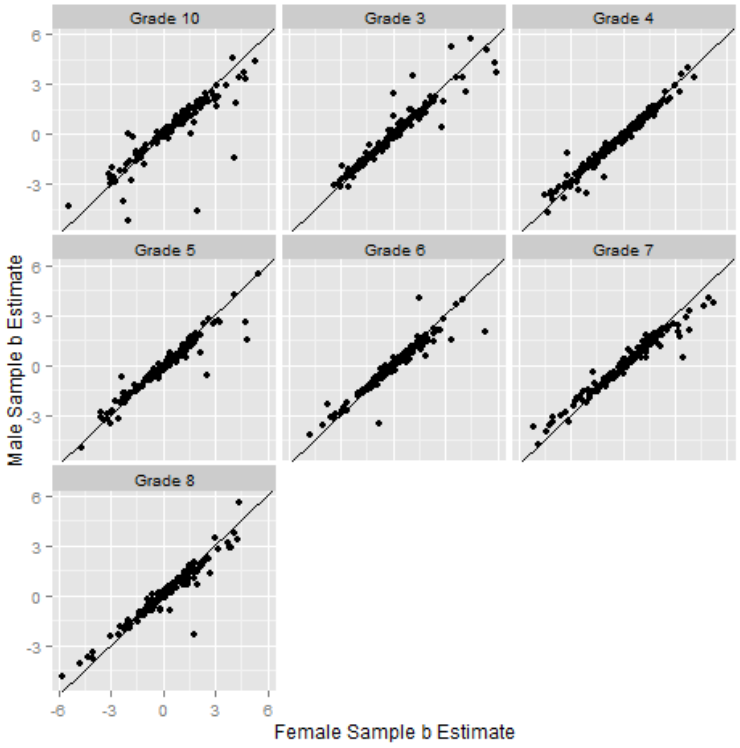


Figure III-5. Mathematics item difficulty parameter scatter plot by grade.

Table III-5. Item-Parameter Correlations Between Female and Male Samples

	Item discrimination		Item difficulty	
Grade	ELA	Mathematics	ELA	Mathematics
3	.83	.91	.86	.95
4	.84	.86	.85	.98
5	.92	.91	.95	.96
6	.89	.92	.96	.94
7	.86	.95	.80	.98
8	.92	.87	.96	.96
10	.85	.75	.90	.86

III.3.3. Differential item function (DIF). DIF examines whether an item shows statistical difference between two groups of students, after ability effect is removed. Logistic regression is used to detect DIF. Based on Jodoin and Gierl's (2001) DIF classification criteria, when the DIF test is significant, moderate DIF has R^2 change between .035 and .070, and large DIF has R^2 change greater than .070.

DIF is examined across gender (female vs. male) and race (Black vs. White). Tables III-6 through III-8 show the number of items identified as having DIF, by grade, for ELA, mathematics, and HGSS. As seen in the tables, the number of items with DIF is close to or equal to zero for all three subjects.

The low DIF item count is expected because CETE has been proactive in improving item quality. Item statistics are used to help writing better items over the years. Among them, DIF has been addressed by providing effective item bias and sensitivity training and guidance to item writers and item reviewers. The concept has been emphasized during item writing training, item writing period, and both internal and external item reviews. The effort results in decreasing the number of DIF items over time.

Table III-6. ELA DIF Item Count by Grade

Grade	No. of items	Gender DIF		Race DIF	
		Moderate	Large	Moderate	Large
3	168	0	0	0	0
4	172	0	0	0	0
5	172	1	1	0	0
6	170	0	0	0	0
7	171	3	0	0	0
8	161	1	0	0	0
10	170	2	0	0	0

Table III-7. Mathematics DIF Item Count by Grade

Grade	No. of items	Gender DIF		Race DIF	
		Moderate	Large	Moderate	Large
3	102	0	0	0	0
4	102	0	0	0	0
5	127	0	0	0	0
6	99	0	0	0	0
7	125	0	0	0	0
8	100	0	0	0	0
10	106	0	0	0	0

Table III-8. HGSS DIF Item Count by Grade

Grade	No. of items	Gender DIF		Race DIF	
		Moderate	Large	Moderate	Large
6	11	0	0	0	0
8	9	0	0	0	0
11	12	0	0	0	0

III.4. Validity Based on Relations to Other Variables

As described in the *Standards*, "evidence based on relationships with other variables provides

evidence about the degree to which these relationships are consistent with the construct underlying the proposed test score interpretations" (AERA et al., 2014, p. 16).

This kind of evidence refers to external evidence and is classified into three types: convergent, discriminant, and criterion related. Convergent evidence is provided by the relationships between students' performance on different assessments intended to measure similar constructs. Discriminant evidence is provided by the relationships between students' performance on different tests intended to measure different constructs. Criterion-related evidence, either predictive or concurrent, is provided by relationships between students' test scores on a criterion measure (Cronbach, 1971; Messick, 1989).

Convergence validity requires that another test measures a similar construct; it is not available for the KAP at the moment. Discriminant validity can be evaluated using the correlation between subjects, such as ELA and mathematics. Past studies showed high correlations between subjects, which indicates some common traits are shared across subjects; however, the correlations should not be too high. The correlations presented in Table III-9 are between subjects of the same grade, and the values range from .63 to .79. Correlation is not computed between different grades.

Table III-9. Correlations Among ELA, Mathematics, and HGSS Scores

Grade	ELA vs. mathematics	ELA vs. HGSS	Mathematics vs. HGSS
3	.79		
4	.79		
5	.76		
6	.77	.72	.65
7	.75		
8	.76	.70	.63
10	.73		
11			

A predictive study between the KAP and ACT scores was conducted in fall 2016. According to ACT, the ACT test measures what students learn in high school and scores are used to determine students' academic readiness for college. The KAP adopted the KCCRS, which are also an indicator of college readiness. Scores of the two tests refer to somewhat different content specifications, but have the same intention. Among the ACT scores, English, reading, mathematics, and composite scores (the average of scores of the four multiple-choice subjects: English, mathematics, reading, and science) were used to correlate with the KAP ELA and mathematics scores.

This study used student ACT scores from 10 school districts. After data cleaning, about 5,369 ACT scores taken after the KAP spring 2015 administration were kept to analyze with 2015 KAP scores. When a student had multiple ACT scores, only one score was selected. Two score selection approaches were used: the first composite score and the highest composite score. The

first ACT score was used because its testing date was closer to the KAP testing window. The highest score was used because it is typically accepted by colleges regardless of the number of times students took the test. Results produced from these two samples (i.e., the first ACT score sample and the highest ACT score sample) are reported in Table III-10.

As shown in Table III-10, the correlation between the tests is greater than .62; the highest correlation of ACT and KAP scores is .85. Logically, KAP ELA scores correlate better with ACT English and reading scores than with ACT mathematics scores, and KAP mathematics correlates better with ACT math scores. Both KAP ELA and mathematics scores correlate well with ACT composite scores (.77–.79).

*Table III-10. Correlations Among KAP and ACT Scores (*N = *5,369)*

	KAP correlation with first ACT scores		KAP correlation with highest ACT scores	
	ELA	Mathematics	ELA	Mathematics
ACT score				
Composite	.78	.78	.77	.79
English	.77	.69	.76	.70
Reading	.73	.61	.73	.62
Math	.64	.85	.64	.85

IV. Technical Quality—Others

IV.1. Reliability

Reliability is a test score consistency index. It is based on the sampling theory that a test is only a sample of all possible items in a content area. To use test scores to infer the knowledge and skills of the content area, the tested content must be representative of the entire content area as defined by the content standards. Additionally, factors that can affect performance, such as allocated testing time, computer environment, and supporting materials, should be standardized to remove undesirable effects. *Standards for Educational and Psychological Testing* (AERA et al., 2014) states that the first step in examining test reliability is to investigate the specifications of replications of the testing procedures. The KAP has standardized its testing procedures, and the same procedures are applied to all students. Specific accommodations are provided to students with special needs. The testing specifications can be found in the *Examiner's Manual*.

Because reliability theory defines each test form as only a sample of the tested content area, different test forms of a subject are different samples of the content area and may yield different observed scores. In sampling theory, the mean of repeated samples' means can infer the population mean. In testing theory, the mean of repeated testing scores is the test taker's true score of the defined content area. However, it is impractical to test the same content area repeatedly because test takers cannot maintain the same knowledge, physical condition, and mental status across test administrations. Factors such as learning, fatigue, and motivation may affect test takers at different rates, making reliability through empirical study unlikely. Therefore, reliability index is derived through theories.

A fundamental reliability theory is defined by the classical test theory. Classical test theory has established that observed score is the composite of true score and measurement error. Measurement error can cause by learning effect and change in motivation, among other factors. Reliability is positively correlated with the proportion of true score. If the proportion of error increases, then reliability will decrease. This relationship has been the basis for many reliability models. Because neither the true score nor error can be observed, these values have to be derived theoretically. Typically, reliability values range from 0 to 1. Higher values indicate better test reliability.

IV.1.1. Test reliability. ELA and mathematics tests use IRT models to estimate students' latent ability (theta), which is then transferred to a scaled score. A standard error is also estimated for each theta and is then transferred to the conditional standard error of measurement (CSEM). CSEMs are computed through their inverse relationship with test information function. Graphic presentation of CSEM curves can be found in Appendix B. The information function and the CSEM are computed using all operational items in a grade, not by block or path. Typical CSEM values are low at the center and gradually increase toward the two ends of the scale, whereas scaled scores become very low and very high and result in a U-shaped pattern. However, some CSEM curves presented in Appendix B have lower values at the low scaled-score side, which may cause by the large number of items and sufficient number of items with lower difficulty

levels.

Standard errors (*SE*s) and their scaled values, CSEMs, indicate reliability by scaled-score points. Green, Bock, Humphreys, Linn, and Reckase (1984) used the standard errors of theta (θ) to derive an index for test-level reliability:

$$\bar{\rho} = \frac{\sigma_\theta^2 - \overline{SE_\theta^2}}{\sigma_\theta^2}. \qquad \text{(IV-1)}$$

Green et al. called this index marginal reliability. The equation shows that marginal reliability, $\bar{\rho}$, is defined by two values: the variance of theta (σ_θ^2) and standard errors of theta (SE_θ^2). Because standard errors are different across thetas, the mean of squared standard errors, $\overline{SE_\theta^2}$, is used in the equation.

The HGSS test, on the other hand, uses raw scores to derive scaled scores. Its reliability is estimated through Cronbach's alpha (1951). Cronbach's alpha (α) is developed using the split-half method. Because repeated testing cannot be actualized, the concept of splitting a single test into two halves to create two tests was created. However, how to split the test became an issue because different item collection of the halves would result in different reliabilities. A solution of using the average of all possible halves was proposed but was deemed impractical because of its computation intensity. Cronbach (1951) found that reliability derived from all possible split halves could be calculated with a simple equation,

$$\alpha = \frac{k}{k-1} \times \frac{\sigma_x^2 - \sum_{i=1}^{k}[p_i(1-p_i)]}{\sigma_x^2}. \qquad \text{(IV-2)}$$

The Cronbach's alpha equation uses κ to represent the number of items on the test. The term σ_x^2 is the variance of the raw scores, and p_i is the p value or item mean for item i. The term $p_i\,(1 - p_i)$ is the variance of item i. Table IV-1 presents the test reliabilities of ELA, mathematics, and HGSS assessments.

Table IV-1.Test Reliability by Grade and Subject

	Subject		
Grade	ELA	Mathematics	HGSS
3	.92	.94	
4	.91	.95	
5	.91	.95	
6	.91	.94	.60
7	.90	.93	
8	.91	.94	.57
High School	.92	.92	.38

Reliabilities of ELA and mathematics tests are above .90. This high reliability range may reflect

the benefit of multistage design. The HGSS test has relatively low reliability. In addition to not using a multistage design, the HGSS test also has significantly fewer items than the other two subject tests. Since reliability is driven by the number of test items, shorter tests tend to have lower reliability. The HGSS high school test has even lower reliability compared to the tests used in grades 6 and 8. Post hoc investigation shows that a higher percentage of students received scores of 0 on the MDPT item because responses did not attempt to answer the question. Additionally, weights are used to combine MDPT scores with the rest of the HGSS items. Both weights and high proportion of zeros affect the reliability of the HGSS high school test.

IV.1.2. Classification consistency and accuracy. How accurately students are classified into performance categories has been a great interest for accountability testing programs. Classification consistency refers to the agreement between two parallel forms and classification accuracy refers to the agreement between true scores and observed scores (Livingston and Lewis, 1995). Tables IV-2 and IV-3 presents the possible classification results of consistency and accuracy of two performance levels, respectively. Both tables indicate that when students are classified into two levels, four possible outcomes are yielded each by the parallel forms and by the true scores and observed scores. Among the four possible outcomes, two of them are consistent (accurate) and two of them are inconsistent (false).

Table IV-2. Cross-Tabulation of Classification Consistency

		Observed Score Parallel Form 2	
		Level X	Level Y
Observed Score Parallel Form 1	Level X	Consistent Classification	Inconsistent Classification
	Level Y	Inconsistent Classification	Consistent Classification

Table IV-3. Cross-Tabulation of Classification Accuracy

		True Score	
		Level X	Level Y
Observed Score	Level X	Accurate Classification	False Negative
	Level Y	False Positive	Accurate Classification

As mentioned earlier, true scores are unobservable and repeat testing is not feasible. In order to evaluate the classification consistency and accuracy of single administration, alternative statistical procedures are developed. Among them, Livingston and Lewis (1995) procedures are broadly used because they are not limited to dichotomous items and do not assume equal weight on items. Livingston and Lewis method uses (a) test reliability to estimate "effective length", (b) user selected true score model to predict the parallel form's observed score distribution for consistency comparison, and (c) user selected model to predict the true score distribution for accuracy estimates.

The results for overall consistency across all four performance levels as well as for the dichotomies created by the three cut scores are presented in Table IV-4. BB-CLASS software (Brennan, 2004) is used to derive the information. Due to being short tests, all HGSS tests have low classification outcomes.

Table IV-4. Classification Consistency and Accuracy by Subject and Grade

	Cut Score Category							
	Overall		1 vs 2, 3, 4		1, 2 vs 3, 4		1, 2, 3 vs 4	
Grade	Consistency	Accuracy	Consistency	Accuracy	Consistency	Accuracy	Consistency	Accuracy
ELA								
3	0.58	0.78	0.73	0.93	0.75	0.91	0.71	0.94
4	0.57	0.79	0.69	0.94	0.74	0.91	0.67	0.94
5	0.55	0.76	0.71	0.93	0.74	0.91	0.68	0.93
6	0.57	0.78	0.74	0.93	0.74	0.91	0.53	0.96
7	0.55	0.77	0.72	0.92	0.72	0.90	0.51	0.96
8	0.60	0.81	0.72	0.93	0.73	0.92	0.56	0.97
10	0.61	0.81	0.76	0.93	0.75	0.92	0.51	0.96
Mathematics								
3	0.63	0.81	0.73	0.95	0.79	0.93	0.76	0.95
4	0.69	0.85	0.75	0.95	0.81	0.94	0.77	0.97
5	0.68	0.83	0.77	0.94	0.81	0.94	0.78	0.97
6	0.65	0.82	0.74	0.93	0.80	0.94	0.77	0.97
7	0.64	0.83	0.70	0.93	0.78	0.93	0.71	0.98
8	0.66	0.84	0.77	0.92	0.81	0.95	0.76	0.98
10	0.61	0.81	0.74	0.91	0.78	0.94	0.74	0.98
HGSS								
6	0.23	0.58	0.37	0.90	0.45	0.80	0.23	0.86
8	0.23	0.63	0.29	0.89	0.38	0.78	0.21	0.96
11	0.12	0.48	0.20	0.68	0.22	0.76	0.03	0.98

IV.1.3. Subgroup reliability. Subgroup reliabilities are presented in Appendix C. Consistent with the previous reliability analysis methods, ELA and mathematics subgroup reliabilities are analyzed using marginal reliability, while HGSS uses Cronbach's alpha. Appendix C shows that race analysis has smaller numbers of students than other subgroups because students whose information about race was not provided were excluded from the analysis.

Both ELA and mathematics have very high subgroup reliabilities. The majority of ELA subgroup reliabilities are in the lower .90 range. The three reliabilities that are lower than .90 are still in the upper .80 range. Mathematics subgroup reliabilities are in the mid-.90 range. HGSS has lower subgroup reliabilities, but they are close to overall HGSS test reliabilities.

IV.1.4. Path reliability. Path reliability is the product of multistage test design; thus it applies only to ELA and mathematics tests. Multistage test design dictates that different sets of items

(blocks) are assigned to students at Stages 2 and 3. The different paths mean that students take item sets with different levels of difficulty. Analytically, multiple test forms are taken by students. Conceptually, path reliability is equivalent to the reliability of different test forms. The results of path reliability can be found in Appendix D.

The stages in the Appendix D tables provide block information for each stage, student count and percentage, and reliability. For example, the path reliability of the ELA test in grade 3 presented in Table IV-5 indicates 10 paths (forms). Stage 1 has only one block of items, with a medium level of difficulty. Stage 2 has three blocks of items: easy, medium, and hard. Stage 3 has four blocks of items: very easy, medium easy, medium hard, and very hard. Students who take the medium block at Stage 2 may receive any of the four blocks at Stage 3. However, those who take the easy or hard block at Stage 2 may assign to only three of the four blocks at Stage 3.

Paths with zero students are a result of low information function of the block. Multistage test design uses the IRT information function to determine block assignment at Stages 2 and 3. The block with the highest information function at the given theta is administered to the student. Due to item bank limitations, some blocks may have low information function across the entire theta range compared to the information function of other blocks of the same stage and will not be selected. This situation is apparent for the ELA test in grade 7, in that the Stage 2 medium block was not assigned to any students.

Table IV-5. ELA Grade 3 Path Reliability

Path	Stage 1	Stage 2	Stage 3	*N*	Percentage	Reliability
				38,208		
1	Medium	Easy	Very easy	11,220	29.4%	.94
2	Medium	Easy	Medium easy	1,268	3.3%	.94
3	Medium	Easy	Medium hard	0		
4	Medium	Medium	Very easy	989	2.6%	.94
5	Medium	Medium	Medium easy	16,430	43.0%	.93
6	Medium	Medium	Medium hard	6,442	16.9%	.89
7	Medium	Medium	Very hard	1,631	4.3%	.83
8	Medium	Hard	Medium easy	0		
9	Medium	Hard	Medium hard	9	0.0%	.86
10	Medium	Hard	Very hard	219	0.6%	.79

IV.1.5. Subscore reliability. Besides the total test score, scores of subsets of ELA and mathematics items are also reported for students. The number of items in each subscore category varies; some items are assigned to multiple subscores. The subscores are categorized into two levels: claim and target. ELA has a total of 10 subscores, three of which are claims: reading, writing, and listening. The other subscores are targets. In the report card, *overall* is added to the claim (e.g., *Overall reading*), and the claim wording is added to the target (e.g., *Reading: Literary texts*). All of the ELA grades report the same 10 subscores.

The 12 mathematics subscores are categorized into concepts and procedures, problem solving, communicating reasoning, and modeling and data analysis. The concepts and procedures category has nine targets; the other categories do not have any targets. Each mathematics grade reports the four claims and two to four targets.

These subscores are reported in three proficiency levels: *below*, *meets*, and *exceeds*. When a student failed to respond more than 40% of the items in a claim, *insufficient data* will be reported instead of a proficiency level. Subscore proficiency levels are assigned according to subscore scaled scores. The procedure for computing subscore scaled scores is similar to that for computing test scaled scores: Student latent abilities (thetas) in each subscore category are estimated using IRT models and are then linearly transformed to scaled scores using the test's scaling constants. Item parameters derived at test level are used to derive thetas here. Cuts of 300 and 325 (one *SE* above 300) are chosen to define students' subscore proficiency levels.

Two analyses are conducted to determine the reliability of ELA and mathematics subscores. First, the subscore marginal reliabilities are computed. In general, statistical estimations are affected by sample sizes. For a test, estimations are affected by both the number of items and student sample sizes. Because the KAP is given to a large number of students, which leaves the subscore reliability mainly driven by number of items. It is expected that the reliability of some subscores may affect by smaller item counts. Second, the classification consistency and accuracy of subscores are examined because the subscores are reported by proficiency levels.

Table IV-6 reports a summary of the subscore reliability and classification consistency and accuracy. Most subscore reliabilities are within good range. The consistency indices have an average of around .40. The average of accuracy indices is above .70 for both subjects.

Table IV-6. Summary of Subscore Reliability, and Classification Consistency and Accuracy by Subject

Subject	No. of items	*Mean*	*SD*	Min	P_{25}	P_{50}	P_{75}	Max
Reliability								
ELA	70	.64	.15	−.05	.57	.67	.72	.87
Mathematics	46	.67	.22	.00	.61	.72	.82	.92
Consistency								
ELA	70	.38	.09	.14	.34	.39	.43	.59
Mathematics	46	.43	.16	.11	.34	.42	.54	.69
Accuracy								
ELA	70	.73	.06	.44	.70	.73	.76	.85
Mathematics	46	.76	.09	.52	.72	.75	.82	.91

Note. Three classification indices for ELA and two for mathematics are missing. P_{25} = 25th percentile; P_{50} = 50th percentile; P_{75} = 75th percentile.

IV.2. Fairness and Accessibility

According to *Standards for Educational and Psychological Testing*, "the central idea of fairness in testing is to identify and remove construct-irrelevant barriers to maximal performance for any examinee" (AERA et al., 2014, p. 74).

This language identifies fairness as an issue related to the validity of test score inferences. Evidence in support of any assertion about the fairness of an assessment can come from several sources, such as item and test development, inclusion and accommodations, and DIF.

Universal design (UD) was used as a guide during the development of items, test formats, and online interface. UD refers to principles that provide equal access to all students. While initially designed to meet the interests of students with special needs, universally designed assessments provide benefits to all students. Implementation of UD started during item-writer training. Using appropriate item- and test-development processes is an excellent start to help ensure fairness. However, some barriers, such as blindness, cannot be addressed by UD. Test inclusion and accommodations policies help address these needs. Many accommodations are provided in the online test system, including magnification, text-to-speech, and image contrasts, among others. Some students will require Braille tests, which are made available to students who need them. (For details about accommodations, see V: Inclusion of All Students.)

Further evidence of the fairness and accessibility of the KAP is seen in DIF analysis. DIF analysis examines whether an item shows statistical difference between two groups of students after ability effect is removed. The DIF analysis results presented in III: Technical Quality—Validity show that, out of nearly 2,000 operational items for all subjects and grades, only seven items show moderate DIF and one shows large DIF.

IV.3. Full Performance Continuum

The KAP was developed with the goal that assessment of each subject area and grade level would provide a reasonably precise estimation of student performances across the full performance continuum (i.e., from low-achieving to high-achieving students). This goal is fulfilled by using items that cover different DOK levels and a wide range of difficulties. As mentioned earlier that although the proportions of each DOK level are not specified in test blueprints, the expected DOK level is explicitly stated in each content standard. When test items are written to each content standard, the items also have to reflect the expected DOK level as specified by the content standard. This expectation is emphasized throughout the item writing and both internal and external item reviews. Consequently, when the items selected for a test meeting the blueprint, those items also meet the underlining DOK requirements.

During test construction, there is no constraint on item p-values or mean scores. Item quality is screened through item-total correlation, DIF, option analyses, and IRT parameters. This approach not only ensures the quality of items to be used on the test, but also provides the widest range possible in measuring student abilities. Additionally, curves of test characteristic, test information, and conditional standard error of measurement are plotted during test construction

to gauge the ability range each test covers. It is noteworthy that one of the advantages of the adaptive test design is it enables the test to extend from the extremely low to the high ability range that is typically ruled out by nonadaptive design.

In order to confirm that the tests efficiently cover the full performance continuum as expected, classical and IRT item statistics are presented here as evidence.

IV.3.1. Classical item statistics. Item difficulty and item discrimination are the two statistics provided. Item difficulty refers to how easy or difficult an item is, and item discrimination indicates the degree to which an item differentiates students of high and low abilities. Item difficulty of classical test theory is expressed as a *p* value or mean score. A *p* value is the percentage of students that correctly answers a multiple-choice item (score 0 and 1); mean score is the average score earned by students on a polytomous item. The range of *p* values is 0 to 1; higher values indicate easier items (Equations IV-1 and IV-2). Because the mean score fluctuates with the possible score range of an item, the mean score can be divided by the possible maximum score to bring it to the range of 0 to 1 (Equation IV-1).

$$p \text{ value / average proportion score} = \frac{\frac{1}{n}\sum_{i=1}^{n} x_i}{\text{item max score}}, \qquad \text{(IV-1)}$$

$$\text{mean score} = \frac{1}{n}\sum_{i=1}^{n} x_i \,, \qquad \text{(IV-2)}$$

where *x* refers to observed score, *i* refers to student *i*, and *n* refers to the total number of students who take the item.

For difficult multiple-choice items with four response options, completely random guessing by students would lead to an expected *p* value of ¼ point (.25). That means there is a 25% chance a student will guess the correct response without any related prior knowledge. For multiple-choice items with five response options, the guessing *p* value would be ⅕ point (.20), and so on for other numbers of response options.

Summaries of item difficulty for ELA, mathematics, and HGSS tests are presented in Tables IV-7 through IV-9. ELA grade-level item difficulty averages are around .60, ranging between .05 and .99 across grades. Mathematics and HGSS have average item difficulties around .50. Mathematics item difficulty averages range from .01 to .97, and HGSS averages range from .21 to .71. The item difficulties for both ELA and mathematics are lower than random guessing. This finding indicates the usefulness of multistage design: Items are assigned to students according to their abilities, thereby reducing the incidence of guessing. Note that the P_{25} and P_{75} in the following tables refer to the 25th and 75th percentiles, respectively.

Table IV-7. Summary Statistics for Classical Item Difficulties for ELA

Grade	No. of items	*M*	*SD*	Min	P_{25}	Median	P_{75}	Max
3	144	.58	.17	.16	.47	.58	.72	.94
4	135	.58	.19	.09	.47	.58	.70	.99
5	130	.59	.17	.26	.46	.58	.71	.96
6	132	.63	.18	.20	.49	.65	.76	.99
7	110	.66	.16	.30	.57	.65	.76	.93
8	115	.60	.15	.23	.50	.60	.70	.90
10	129	.60	.17	.05	.48	.64	.72	.91

Table IV-8. Summary Statistics for Classical Item Difficulties for Mathematics

Grade	No. of items	*M*	*SD*	Min	P_{25}	Median	P_{75}	Max
3	134	.52	.21	.06	.36	.50	.68	.91
4	135	.56	.20	.10	.38	.59	.72	.92
5	129	.49	.18	.05	.37	.48	.63	.90
6	120	.48	.17	.01	.37	.47	.56	.97
7	117	.48	.18	.05	.36	.49	.61	.93
8	122	.50	.19	.05	.37	.51	.66	.90
10	112	.46	.17	.03	.35	.45	.58	.93

Table IV-9. Summary Statistics for Classical Item Difficulties for HGSS

Grade	No. of items	*M*	*SD*	Min	P_{25}	Median	P_{75}	Max
6	10	.50	.16	.23	.45	.53	.61	.71
8	8	.54	.17	.21	.50	.57	.64	.70
11	11	.56	.12	.32	.49	.57	.64	.73

Note. P_{25} = 25th percentile; P_{75} = 75th percentile.

Item discrimination reflects an item's ability to differentiate students of high and low abilities. Ideally, high-achieving students (i.e., those with high raw scores) should be more likely to answer any given item correctly, whereas low-achieving students (i.e., those with low raw scores) should be more likely to answer the same item incorrectly. The Pearson's product-moment correlation coefficient between student item scores and test scores is used to compute item discrimination. It is also known as item-total correlations, or point-biserial correlations when items have dichotomous (0, 1) scores. The item-total correlation ranges from −1.0 to 1.0. Positive values indicate that students with high raw scores are more likely to answer an item correctly than are students with low raw scores; negative values mean the opposite. The magnitude of the correlation indicates the degree of discrimination, in that higher values have better discrimination power.

Tables IV-10 through IV-12 present item discrimination for the three subjects. The medians of item discrimination for ELA and mathematics are around .30 across grades, a good range for

item discrimination. HGSS has an even higher median of item discrimination across grades, around .50.

Table IV-10. Summary Statistics for Classical Item Discrimination for ELA

Grade	No. of items	*M*	*SD*	Min	P_{25}	Median	P_{75}	Max
3	144	.31	.10	.04	.24	.31	.37	.60
4	135	.30	.10	.05	.24	.30	.37	.59
5	130	.31	.09	.09	.25	.31	.38	.59
6	132	.31	.10	.10	.23	.30	.40	.52
7	110	.37	.14	.11	.27	.37	.46	.82
8	115	.32	.09	.09	.26	.32	.38	.54
10	129	.33	.11	.11	.24	.29	.40	.62

Note. P_{25} = 25th percentile; P_{75} = 75th percentile.

Table IV-11. Summary Statistics for Classical Item Discrimination for Mathematics

Grade	No. of items	*M*	*SD*	Min	P_{25}	Median	P_{75}	Max
3	134	.33	.12	.09	.25	.32	.41	.66
4	135	.35	.12	.12	.26	.34	.45	.62
5	129	.35	.13	.11	.26	.33	.46	.69
6	120	.35	.12	.07	.26	.34	.46	.62
7	117	.35	.11	.11	.27	.35	.42	.64
8	122	.35	.12	.14	.25	.34	.43	.63
10	112	.32	.11	.06	.26	.32	.39	.55

Note. P_{25} = 25th percentile; P_{75} = 75th percentile.

Table IV-12. Summary Statistics for Classical Item Discrimination for HGSS

Grade	No. of items	Mean	*SD*	Min	P_{25}	Median	P_{75}	Max
6	10	.50	.16	.23	.45	.53	.61	.71
8	8	.54	.17	.21	.50	.57	.64	.70
11	11	.56	.12	.32	.49	.57	.64	.73

Note. P_{25} = 25th percentile; P_{75} = 75th percentile.

IV.3.2. IRT item statistics. The KAP ELA and mathematics tests are calibrated and equated using a pre-equating method: Last year's item statistics are used to calculate thetas. Because of item pool limitation, a few items with negative discrimination parameters were used for 2016 operational tests for blueprint coverage. These items were recalibrated in 2016 and again screened for negative discrimination parameters. If the recalibrated items still had negative discrimination parameters, they were excluded from equating and scoring.

Tables IV-13 through IV-16 summarize the difficulty, *b*, and discrimination, *a*, parameter estimates for operational items. Most items are dichotomous, but some items have as many as 11 score categories (thus, 10 *b* parameters yet still only one *a* parameter); therefore, the numbers of *b* and *a* parameters are different in these tables. Parameters for all items, irrespective of the number of categories, are included together in the tables below.

The mean item difficulty increases as the grade increases for mathematics, but remains close to −0.5 in all grades for ELA. The large standard deviations of difficulty parameters indicate a large variability of item difficulties. The minima and maxima for the difficulty parameters indicate that the items included in the KAP assessments adequately cover the full performance continuum. Although item discrimination is not usually too far from 1.0 on average, it clearly varies over items, justifying the use of the 2PL that permits discrimination parameter to vary over items. The median item discrimination declines as the grade increases for mathematics, but remains close to 0.9 in all grades for ELA. Overall, mathematics has better discrimination parameters than ELA does.

Table IV-13. Summary Statistics for IRT Item Difficulty for ELA

Grade	No. of *b* parameters	*M*	*SD*	Min	Q1	Median	Q3	Max
3	162	−0.08	1.37	−2.94	−1.01	−0.34	0.70	3.53
4	154	−0.52	1.39	−3.84	−1.38	−0.76	0.29	5.10
5	155	−0.54	1.56	−5.36	−1.53	−0.67	0.62	3.36
6	172	−0.70	1.62	−4.32	−1.79	−0.95	0.38	4.80
7	135	−0.75	1.45	−4.21	−1.57	−0.83	−0.01	7.94
8	168	−0.82	1.60	−6.04	−1.79	−0.84	0.05	3.67
10	172	−0.52	1.57	−3.87	−1.55	−0.68	0.38	4.26

Note. *b* = difficulty parameter; Q1 = first quartile; Q3 = third quartile.

Table IV-14. Summary Statistics for IRT Item Difficulty for Mathematics

Grade	No. of *b* parameters	*M*	*SD*	Min	Q1	Median	Q3	Max
3	158	−0.14	1.93	−10.76	–1.20	−0.20	0.82	5.59
4	212	−0.38	1.58	−4.04	–1.52	−0.40	0.72	4.08
5	169	−0.07	1.43	−3.65	–0.86	−0.09	0.94	5.13
6	145	0.05	1.68	−6.60	–0.77	0.07	0.93	6.02
7	156	0.09	2.01	−8.33	–0.89	0.34	1.36	5.50
8	157	0.25	1.76	−5.68	–0.69	0.25	1.37	4.54
10	138	0.51	1.94	−4.32	–0.51	0.31	1.58	7.67

Note. *b* = difficulty parameter; Q1 = first quartile; Q3 = third quartile.

Table IV-15. Summary Statistics for IRT Item Discrimination for ELA

Grade	No. of *a* parameters	*M*	*SD*	Min	Q1	Median	Q3	Max
3	144	0.96	0.43	0.26	0.66	0.89	1.23	2.50
4	135	0.88	0.37	0.25	0.61	0.83	1.05	2.06
5	130	0.92	0.37	0.26	0.64	0.90	1.20	1.89
6	132	0.94	0.38	0.20	0.68	0.90	1.18	2.01

Grade	No. of *a* parameters	*M*	*SD*	Min	Q1	Median	Q3	Max
7	110	0.97	0.39	0.22	0.67	0.95	1.20	2.11
8	115	0.94	0.36	0.23	0.74	0.91	1.19	2.00
10	129	0.90	0.33	0.22	0.67	0.90	1.12	2.21

Note. *a* = discrimination parameter; Q1 = first quartile; Q3 = third quartile.

Table IV-16. Summary Statistics for IRT Item Discrimination for Mathematics

Grade	No. of *a* parameters	*M*	*SD*	Min	Q1	Median	Q3	Max
3	134	1.11	0.42	0.30	0.80	1.11	1.39	2.26
4	135	1.19	0.36	0.57	0.91	1.14	1.41	2.02
5	129	1.19	0.34	0.34	0.96	1.17	1.41	2.10
6	120	1.12	0.45	0.38	0.81	1.10	1.44	2.36
7	117	0.95	0.42	0.23	0.63	0.85	1.20	2.01
8	122	1.00	0.38	0.25	0.72	0.94	1.23	2.04
10	112	0.89	0.40	0.20	0.61	0.84	1.16	2.00

Note. *a* = discrimination parameter; Q1 = first quartile; Q3 = third quartile.

IV.3.3. Cognitive complexity. KAP items are categorized by cognitive complexity, as described by Webb's depth of knowledge (DOK; Webb, 1997). A description of Webb's DOK follows.

- Level 1 (recall) requires simple recall of such information as a fact, definition, term, or simple procedure.
- Level 2 (skill/concept) involves some mental skills, concepts, or processing beyond a habitual response; students must make some decisions about how to approach a problem or activity. Keywords distinguishing a Level 2 item include classify, organize, estimate, collect data, and compare data.
- Level 3 (strategic thinking) requires reasoning, planning, using evidence, and thinking at a higher level.
- Level 4 (extended thinking) requires complex reasoning, planning, developing, and thinking, most likely over an extended time. Cognitive demands are high, and students are required to make connections both within and among subject domains.

Item cognitive complexity is affected by the familiarity of the constructs being measured. Constructs that were previously taught in the same grade or earlier than described by the KCCRS are likely to appear easier in the early years of the assessment than constructs that were previously taught in higher grades or not addressed in previous content standards. The DOK associated with each content standard identifies the maximum DOK for an item. Items at Level 4, extended thinking, are not typically seen in most assessments unless extended performance tasks are included.

Tables IV-17 shows the percentage of operational items by DOK level, subject, and grade. This information also reveals the proportions of DOK requirements according to content standards. Most ELA items are at Level 1 and Level 2; fewer items are at Level 3. In mathematics, most items are at Level 1 and Level 2 as well, with relatively fewer Level 3 items. For ELA and

mathematics, there are only one or two Level 4 items per grade, but as noted earlier, Level 4 items are very rare in most assessments unless extended performance tasks are included. In HGSS, due to the small number of items tested, the number of items per level is small. However, HGSS has a performance task that contributes to the Level 4 DOK.

Table IV-17. Number of Items by DOK Level, Subject, and Grade

	ELA					Math					HGSS				
		DOK Level, %					DOK Level, %					DOK Level, %			
Grade	**Total Items**	**1**	**2**	**3**	**4**	**Total Items**	**1**	**2**	**3**	**4**	**Total Items**	**1**	**2**	**3**	**4**
3	147	29	59	12	1	135	39	56	5	0					
4	140	24	59	16	1	135	36	55	8	1					
5	132	28	55	17	1	129	42	54	4	0					
6	134	30	47	22	1	120	40	59	1	0	11	18	36	36	9
7	114	18	68	12	2	120	40	58	2	0					
8	120	19	61	19	1	122	36	61	3	0	9	22	56	11	11
10	129	27	63	10	0	114	32	64	4	0					
11											12	25	42	25	8

IV.4. Scoring and Scaling

This section discusses the procedures for scoring individual items, scoring the test as a whole, and scaling. Student responses go through these three procedures to produce a reliable and valid report score.

IV.4.1. Rubrics. KAP items are scored with two scoring methods: machine scoring and human scoring. Machine scoring compares student responses to the correct keys and assigns the predetermined scores accordingly. Human scoring uses pre-determined rubrics to determine the scores of student responses. The KAP has two types of performance tasks scored by human raters: mathematics performance tasks and MDPTs. The rubrics of mathematics performance tasks are confidential; thus only the MDPT rubrics are presented here as examples.

MDPT is scored holistically (Figure IV-1); however, three analytical rubrics (Figures IV-3 through IV-5) are used to guide raters during scoring process. The analytical rubrics provide rationales to justify score assignment, and the holistic rubric provides the general meaning for score interpretation. Each MDPT aligns with one of the three analytical rubrics: opinion/ argument, information explanatory, and narrative. The scoring rubrics for grades 6 through 8 are presented below as examples. The holistic rubric and condition codes (Figure 4-2) for responses that did not reply to the question are identical across grades. However, the analytical rubric is different for each grade.

In the response, the student demonstrates:	
4	consistent and effective command of the skills needed to complete an on-demand writing task.
3	mostly consistent and adequate command of the skills needed to complete an on-demand writing task.
2	somewhat consistent and minimal command of the skills needed to complete an on-demand writing task.
1	inconsistent and ineffective command of the skills needed to complete an on-demand writing task.

Figure IV-1. Holistic rubric for multidisciplinary performance task.

BL	Blank	The response is blank.
IN	Insufficient	The response does not include enough student writing to score.
OT	Off Task	The response is unrelated to the resources and/or prompt.
OL	Other Language	The response is in a language other than English.

Figure IV-2. Condition codes for multidisciplinary performance task.

6-8 Argument	Student's Response…			
PL:	4	3	2	1
Focus/Argument	☐ States a clear argument related to the resources and prompt, and maintains it throughout the work	☐ States a clear argument related to resources and prompt and mostly maintains it throughout the work	☐ States a somewhat clear argument, which may lose focus sporadically throughout the work	☐ Does not state a clear argument, or stated argument is unrelated to resources or prompt
Evidence	☐ Uses relevant and accurate details/evidence from two or more resources to support argument	☐ Uses mostly relevant and accurate details/ evidence from two or more resources to support argument	☐ Uses some relevant and accurate details/evidence from one or more resources to support argument	☐ Does not use relevant and accurate details or evidence from resources to support argument
Argument	☐ Consistently uses grade-appropriate strategies to clarify relationships between and among ideas, and to connect evidence to argument ☐ Consistently and accurately uses domain-specific words to develop and support argument	☐ Adequately uses grade-appropriate strategies to clarify relationships between and among ideas and to connect evidence to argument ☐ Adequately uses domain-specific words to develop and support argument	☐ Inconsistently uses grade-appropriate strategies to clarify relationships between and among ideas and to connect evidence to argument ☐ Inconsistently uses domain-specific words to develop and support argument	☐ Shows little or no attempt to clarify relationships between and among ideas or connect evidence to argument ☐ Uses few or no domain-specific words to develop and support argument
Conventions	☐ Is readable with most grade-level conventions used correctly and may use them creatively to enhance the message; minor mistakes do not impede the reader's ability to understand the writer's meaning	☐ Is readable with most grade-level conventions used correctly; mistakes do not affect the reader's ability to understand the writer's meaning	☐ Is readable but some errors in grade-level conventions negatively impact the reader's ability to understand the writer's meaning	☐ Is nearly unreadable due to pervasive errors in grade-level conventions

Figure IV-3. Analytical Rubric (Opinion/Argument) for multidisciplinary performance task.

6-8 Inf./Expl.	Student's Response…			
PL:	4	3	2	1
Focus	☐ States and maintains a clear controlling idea that directly addresses the resources and prompt	☐ States and maintains a clear controlling idea that mostly addresses the resources and prompt	☐ States a controlling idea somewhat related to the resources and prompt	☐ Does not state a clear controlling idea, or stated controlling idea is largely unrelated to resources or prompt
Support	☐ Uses relevant and accurate facts, definitions, and details from two or more resources to help explain the controlling idea	☐ Uses mostly relevant and accurate facts, definitions, and details from two or more resources to help explain the controlling idea	☐ Uses some relevant and accurate facts, definitions, and details from one or more resources to help explain the controlling idea	☐ Does not use relevant or accurate facts, definitions, or details from the resources to help explain the controlling idea
Connections and Audience	☐ Consistently uses grade-appropriate strategies to clarify relationships between and among ideas, and to help explain the controlling idea ☐ Consistently and accurately uses domain-specific words to develop and explain ideas	☐ Adequately uses grade-appropriate strategies to clarify relationships between and among ideas, and to help explain the controlling idea ☐ Adequately uses domain-specific words to develop and explain ideas	☐ Inconsistently uses grade-appropriate strategies to clarify relationships between and among ideas, and to help explain the controlling idea ☐ Inconsistently uses domain-specific words to develop and explain ideas	☐ Shows little or no attempt to clarify relationships between and among ideas, or to help explain the controlling idea ☐ Uses few or no domain-specific words to develop and explain ideas
Conventions	☐ Is readable with most grade-level conventions used correctly and may use them creatively to enhance the message; minor mistakes do not impede the reader's ability to understand the writer's meaning	☐ Is readable with most grade-level conventions used correctly; mistakes do not affect the reader's ability to understand the writer's meaning	☐ Is readable but some errors in grade-level conventions negatively impact the reader's ability to understand the writer's meaning	☐ Is nearly unreadable due to pervasive errors in grade-level conventions

Figure IV-4. Analytical Rubric (Information Explanation) for multidisciplinary performance task.

6-8 Narrative	Student's Response…			
PL:	4	3	2	1
Storyline and Characters	☐ Effectively establishes a setting and a narrator/character(s) OR maintains a setting and a narrator/character(s)	☐ Adequately establishes a setting and a narrator/character(s) OR maintains a setting and a narrator/character(s)	☐ Unevenly or minimally establishes a setting and a narrator/character(s) OR unevenly or minimally maintains a setting and a narrator/character(s)	☐ Shows little or no attempt to establish a setting and a narrator/character(s) OR shows little or no attempt to maintain a setting and a narrator/character(s)
Development	☐ Effectively uses dialogue and/or descriptions to develop characters and/or situations ☐ Effectively uses grade-appropriate sensory details to convey experiences and events	☐ Adequately uses dialogue and/or descriptions to develop characters and/or situations ☐ Adequately uses grade-appropriate sensory details to convey experiences and events	☐ Unevenly or minimally uses dialogue and/or descriptions to develop character(s) and/or situations ☐ Unevenly uses grade-appropriate sensory details to convey experiences and events	☐ Shows little or no attempt to use dialogue and/or descriptions to develop character(s) and/or situations ☐ Uses few or no grade-appropriate sensory details to convey experiences and events
Sequencing	☐ Purposefully uses grade-appropriate temporal words and phrases to signal event order	☐ Uses some grade-appropriate temporal words and phrases to signal event order	☐ Uses few grade-appropriate temporal words and phrases to signal event order	☐ Does not use any grade-appropriate temporal words or phrases to signal event order
Introduction and Conclusion	☐ Includes an effective and grade-appropriate introduction and conclusion	☐ Includes an adequate and grade-appropriate introduction and conclusion	☐ Might include a grade-appropriate introduction or conclusion, but one or both are weak.	☐ Does not include an introduction or a conclusion.
Conventions	☐ Is readable with most grade-level conventions used correctly and may use them creatively to enhance the message; minor mistakes do not impede the reader's ability to understand the writer's meaning	☐ Is readable with most grade-level conventions used correctly; mistakes do not affect the reader's ability to understand the writer's meaning	☐ Is readable but some errors in grade-level conventions negatively impact the reader's ability to understand the writer's meaning	☐ Is nearly unreadable due to pervasive errors in grade-level conventions

Figure IV-5. Analytical rubric (Narrative) for multidisciplinary performance task.

IV.4.2. Rater scoring training. All mathematics performance tasks are scored by Questar, using scoring guides that are unique to each grade and task. All raters receive scoring training. After receiving an introductory group training, raters review the rubric and then practice scoring items. Practice begins with group scoring and follows with individual scoring practice. During group scoring practice, rationales for score assignments are discussed to improve alignment among raters. For validity, during individual scoring, rater scores are evaluated against scores assigned by trainers. After training, raters begin scoring real student responses.

The MDPTs are scored by educators from across the state. Educators complete a PowerPoint training created by CETE that provides information about the MDPTs and familiarizes educators with the holistic rubric used at all grades, the analytical rubrics used at each grade band, and the ScorePoint scoring system. After completing the PowerPoint training, educators log into ScorePoint to score practice responses prior to scoring actual student responses. Rater scoring training materials can be found in Appendix E.

The interrater reliability is estimated by sampling 10% of student records to score by a second rater. It is acceptable when the scores assigned by two raters are identical or adjacent. When two scores differ by more than one point, a third rater is assigned. Raters whose scores show signs of drifting are retrained and monitored.

IV.4.3. Test scoring. Test scoring uses a psychometric model to derive item scores on the test to produce a single score indicating a student's proficiency level. For ELA and mathematics tests, IRT models are used. For HGSS tests, classical test theory (CTT) is used.

IV.4.3.1 *ELA and mathematics test scoring.* The IRT ability estimates (thetas) are computed using the 2PL model and GRM. Because the total score is derived using the number-correct method—in which scores for each item are added together to derive the raw score—thetas have one-to-one correspondence with raw scores (i.e., each raw score has only one matching theta). Using the test characteristic curve function of the IRT models, the theta for each raw-score point is obtained for a test form (Press, Flannery, Teukolsky, & Vetterling, 1989).

IV.4.3.2 *HGSS test scoring.* The HGSS test uses the CTT model for scoring and scaling; thus, its initial scores are raw scores. The test comprises three parts. Part 1 is vignette-based historical questions; Part 2 is document-based historical questions. Parts 1 and 2 are machine scored. Part 3 has an MDPT and is hand scored. Scoring weights are assigned to the three parts of the test to yield weighted raw scores. Table IV-20 presents the highest possible unweighted raw scores, the weight for each section, and the highest possible weighted raw scores for each part by grade.

Table IV-18. HGSS Maximum Unweighted and Weighted Raw Scores and Weights.

Grade		Part 1	Part 2	MDPT	Total
6	Max. unweighted score	7.00	3.00	4.00	
	Weight	5.00	13.33	6.25	
	Max. weighted score	35.00	40.00	25.00	100
8	Max. unweighted score	5.00	3.00	4.00	
	Weight	9.00	10.00	6.25	
	Max. weighted score	45.00	30.00	25.00	100
11	Max. unweighted score	8.00	3.00	4.00	
	Weight	2.50	10.00	12.50	
	Max. weighted score	20.00	30.00	50.00	100

IV.4.4. Scaling. Scaling is the procedure of transforming thetas or raw scores to a scale; the purpose is to facilitate the use and interpretation of test scores. The scale is also the basis for setting performance levels. The section below addresses the procedures for constructing scaled scores. Procedures used to establish ELA and mathematics performance-level cut scores can be found in the 2015 Technical Manual. Procedures used to establish HGSS cut scores are described in the current manual in VI: Academic Achievement Standards and Reporting.

IV.4.4.1 ***Scale transformation and cut scores.*** Kolen and Brennan (2004) used the following formula to derive scaling constants:

$$SS(y) = \frac{\sigma(SS)}{\sigma(Y)} y + [SS(y_1) - \frac{\sigma(SS)}{\sigma(Y)} y_1] \quad , \qquad \text{(IV-3)}$$

where $SS(y)$ is the scaled score, $\sigma(SS)$ is its standard deviation, $\sigma(Y)$ is the standard deviation of the original scores, y_1 is an original score, and $SS(y_1)$ is the scaled score equivalent to the original score, y_1. This equation can be structured to

$$SS = A \times y + C, \text{ where} \qquad \text{(IV-4)}$$

$$A = \frac{\sigma(SS)}{\sigma(Y)} \text{ and} \qquad \text{(IV-5)}$$

$$C = SS(y_1) - \frac{\sigma(SS)}{\sigma(Y)} y_1. \qquad \text{(IV-6)}$$

A and C are the slope and intercept of the scaling constants, respectively. KSDE has predetermined the scaled score to have a slope, A, of 25 for all subjects and grades.

The KAP has four performance levels, Level 1 through Level 4; higher levels indicate higher performance. Students in Level 3 or 4 are considered proficient. A scaled score of 300 is determined by KSDE as the cut that separates Levels 2 and 3 (Level 2/3). In other words, a scaled score of 300 separates students into proficient and nonproficient groups. The original scores of Level 2/3 cuts of each subject and grade are set by standard-setting panels. With the

original cut score (y_1), equivalent scaled score (i.e., SS(y_1) = 300), and a scaled-score standard deviation of 25 (i.e., σ(ss) = 25) identified, the intercept, C, can be derived using Equation IV-6 after the standard deviation, σ(Y), is computed.

IV.4.4.2 *ELA and mathematics scale transformation.* Equating of ELA and mathematics is conducted with IRT models; thus, their initial ability estimates are the IRT thetas. The theoretical values of theta range from negative infinity to positive infinity. In other words, thetas can be negative values and have decimal points. One can imagine the difficulty of using and interpreting negative test scores with multiple decimal points. To ease score interpretation, it is crucial to transform thetas to a scale composed of positive integers.

Since thetas are used for ELA and mathematics, the y_1 in Equation IV-6 is the theta associated with a scaled score of 300. The grade-level theta cuts for ELA and mathematics were set by standard-setting panels in 2015 (see theta cuts in Tables IV-21 and IV-22). Using Equation IV-6, the C for each grade is found (see Table IV-23). Since A and C are known, the other two scaled-score cuts can be derived using Equation IV-4. Note that the scaled-score cuts are rounded up rather than to the nearest integer. The rationale for rounding up is that students need to have scores equal to or higher than the cut score to pass a given level.

Table IV-19. ELA Cut Scores.

	Theta cuts			Scaled-score cuts		
Grade	Level 1/2	Level 2/3	Level 3/4	Level 1/2	Level 2/3	Level 3/4
3	−1.015	−0.050	1.020	276	300	327
4	−1.457	−0.275	1.107	271	300	335
5	−1.085	−0.064	0.952	275	300	326
6	−0.756	0.181	1.594	277	300	336
7	−0.800	0.219	1.610	275	300	335
8	−0.940	0.495	1.850	265	300	334
10	−0.785	0.465	1.800	269	300	334

Table IV-20. Mathematics Cut Scores.

	Theta cuts			Scaled-score cuts		
Grade	Level 1/2	Level 2/3	Level 3/4	Level 1/2	Level 2/3	Level 3/4
3	−1.225	−0.230	0.906	276	300	329
4	−1.215	0.160	1.375	266	300	331
5	−0.885	0.219	1.245	273	300	326
6	−0.882	0.215	1.340	273	300	329
7	−1.055	0.321	1.980	266	300	342
8	−0.527	0.530	1.968	274	300	336
10	−0.497	0.530	1.830	275	300	333

Table IV-21. ELA and Mathematics Scaling Constants.

	ELA		Mathematics	
Grade	*A*	*C*	*A*	*C*
3	25	301.25	25	305.75
4	25	306.87	25	296.00
5	25	301.59	25	294.53
6	25	295.48	25	294.63
7	25	294.53	25	291.98
8	25	287.63	25	286.75
10	25	288.38	25	286.75

Note. *A* = slope; *C* = intercept.

IV.4.4.3 *HGSS scale transformation.* Due to the adoption of new content standards, a new scale was established for HGSS in 2016. HGSS uses raw scores as initial ability estimates, rather than the thetas of the IRT model. To transfer raw scores onto the scale with $SS(y_1)$ equal to 300 and a standard deviation of 25, the y_1 in Equation IV-6 is replaced with raw-score cuts.

The raw-score cuts of HGSS were set in May 2016 during a standard-setting workshop, as described in VI: Academic Achievement Standards and Reporting. Table IV-24 shows the raw-score and scaled-score cuts, and Table IV-25 presents the scaling constants of HGSS.

Table IV-22. HGSS Cut Scores

	Raw-score cuts			Scaled-score cuts		
Grade	Level 1/2	Level 2/3	Level 3/4	Level 1/2	Level 2/3	Level 3/4
6	25	52	74	273	300	329
8	29	54	83	274	300	336
11	37	65	93	275	300	333

Table IV-23. HGSS Scaling Constants

Grade	*A*	*C*
6	1.228553	236.1153
8	1.348829	227.1632
10	1.154917	224.9304

Note. *A* = slope; *C* = intercept.

IV.4.4.4 *Properties of scaled scores.* Whether scaling constants are applied to thetas for ELA and mathematics or to raw scores for HGSS, the derived scaled scores are decimal numbers and must be rounded up to the nearest integers.

The IRT model cannot estimate the thetas of extreme scores (e.g., 0 and perfect raw scores) because responses to all items are identical. A theta of −99 or 99 is typically assigned to those raw-score points by software. In order to keep the scaled score meaningful, the lowest obtainable scaled score (LOSS) and the highest obtainable scaled score (HOSS) are set to cap scaled scores

within a reasonable range. KAP's LOSS and HOSS are set at 220 and 380, respectively.

IV.4.5. Operational test results. Summaries of scaled scores by subject and grade are presented in Tables IV-26 through IV-28; summaries by demographic subgroups are presented in Appendix C. Tables IV-26 through IV-28 indicate that the minimum and maximum values are within the LOSS and HOSS values of 220 and 380, respectively. The differences between (a) P_{50} and P_{25} and (b) P_{75} and P_{50} are indicators of the shapes of score distributions: The larger of the two differences indicates the direction of any skewness in the distribution (a negative skew when the first difference is larger and a positive skew when the second difference is larger). If the two differences match, the distribution is symmetric. In ELA, the distributions for grades 6 and 10 are symmetric in shape, the distributions for grades 4, 5, 7, and 8 are negatively skewed, and the distribution for grade 3 is positively skewed. In mathematics, all distributions are positively skewed. In HGSS, grade 6 is negatively skewed, and grades 8 and 11 are positively skewed.

Table IV-24. Scaled-Score Descriptive Statistics by Grade for ELA

Grade	*Mean*	*SD*	Min	P_{10}	P_{25}	P_{50}	P_{75}	P_{90}	Max
3	298.7	28.0	220	262	278	297	318	336	380
4	303.0	29.3	220	266	282	303	322	342	380
5	298.2	29.3	220	259	278	298	317	337	380
6	293.1	28.4	220	255	274	293	312	328	380
7	292.7	27.9	220	256	274	293	311	328	380
8	286.7	28.5	220	250	266	287	305	323	380
10	285.3	29.6	220	246	265	285	305	323	380

Note. P_{10}, P_{25}, P_{50}, P_{75} and P_{90} are 10th, 25th, 50th, 75th, and 90th percentiles, respectively.

Table IV-25. Scaled-Score Descriptive Statistics by Grade for Mathematics

Grade	*M*	*SD*	Min	P_{10}	P_{25}	P_{50}	P_{75}	P_{90}	Max
3	304.3	27.8	220	270	285	302	322	341	380
4	293.3	28.2	220	259	272	291	311	332	380
5	292.2	27.4	220	260	272	288	310	330	380
6	292.2	27.2	220	261	273	288	308	329	380
7	289.4	28.4	220	256	269	286	306	328	380
8	285.5	28.4	220	254	266	281	301	324	380
10	285.0	28.2	220	256	266	279	299	324	380

Note. P_{10}, P_{25}, P_{50}, P_{75} and P_{90} are 10th, 25th, 50th, 75th, and 90th percentiles, respectively.

Table IV-26. Scaled-Score Descriptive Statistics by Grade for HGSS

Grade	*M*	*SD*	Min	P_{10}	P_{25}	P_{50}	P_{75}	P_{90}	Max
6	296.8	24.7	236	263	278	298	316	329	359
8	292.7	24.7	227	261	276	292	309	326	362
11	280.3	24.9	225	248	262	279	298	315	340

Note. P_{10}, P_{25}, P_{50}, P_{75} and P_{90} are 10th, 25th, 50th, 75th, and 90th percentiles, respectively.

The scaled-score means presented in Appendix C show that, across all subjects and grades, Asian

students have the highest mean scores, followed by White students. Students of American Indian and non-Hispanic Pacific Island backgrounds have similar performances. Black students do not perform as well as other groups do. The gaps between the highest and smallest subgroup mean scores range between 22 and 28 scaled-score points for ELA, 31 and 36 for mathematics, and 18 and 23 for HGSS.

The proportion of students in each performance level (Levels 1 through 4) and passing rate (combined Levels 3 and 4) are provided by subject and grade in Table IV-29 and Figures IV-6 through IV-8. The passing rates ranged from 24% to 55% across subjects and grades. All three subjects tended to have lower passing rates in higher grade levels.

Table IV-27. Percentage of Students in Each Performance Level by Subject and Grade

	ELA (%)					Mathematics (%)					HGSS (%)				
Grade	1	2	3	4	Pass	1	2	3	4	Pass	1	2	3	4	Pass
3	22	31	29	17	**46**	15	30	36	19	**55**					
4	14	33	39	14	**53**	16	46	27	10	**37**					
5	22	32	29	18	**47**	26	38	23	13	**36**					
6	27	30	35	7	**42**	25	41	24	10	**34**	13	40	35	13	**48**
7	26	33	34	7	**41**	20	48	26	5	**31**					
8	23	45	27	5	**32**	40	34	21	6	**27**	14	48	35	4	**39**
10	28	40	26	6	**32**	41	34	17	7	**24**					
11											32	44	22	2	**24**

Note. Passing rates are in boldface.

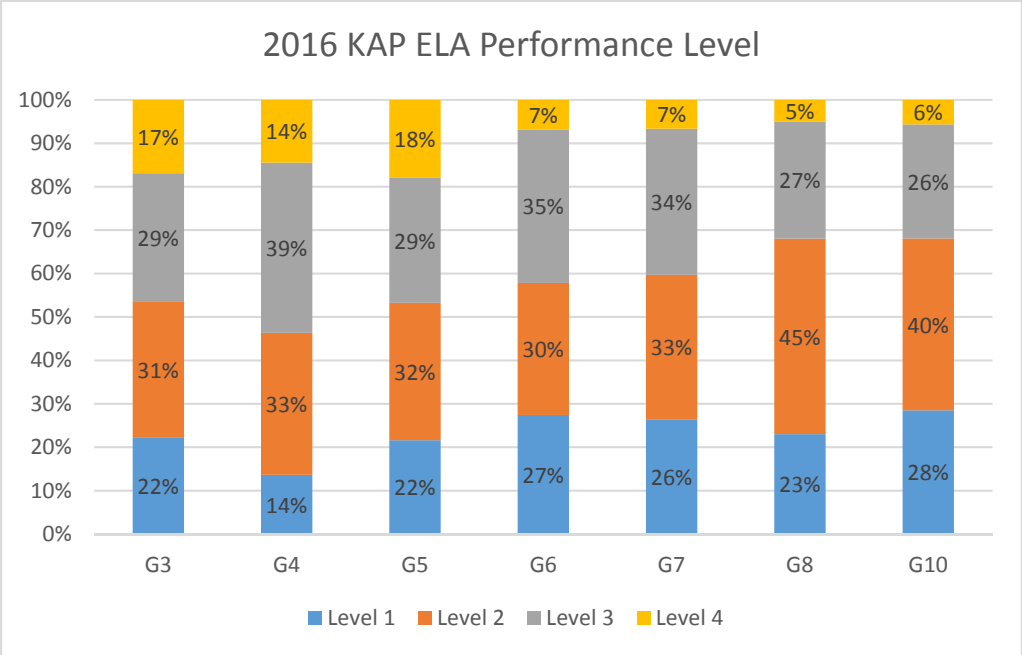


Figure IV-6. Performance-level results for ELA.

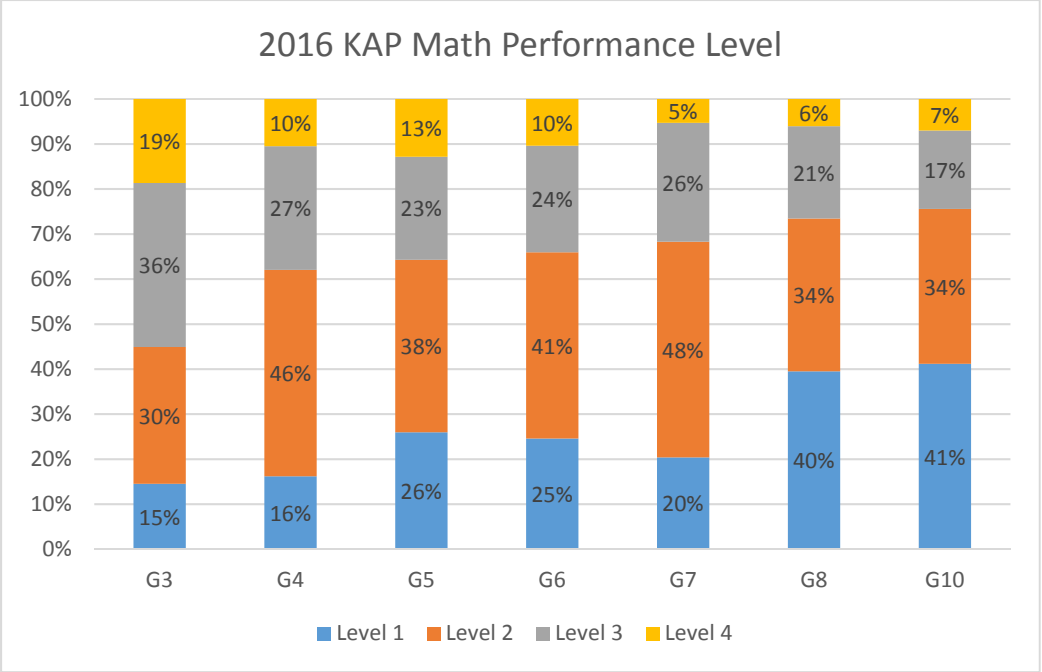


Figure IV-7. Performance-level results for mathematics.

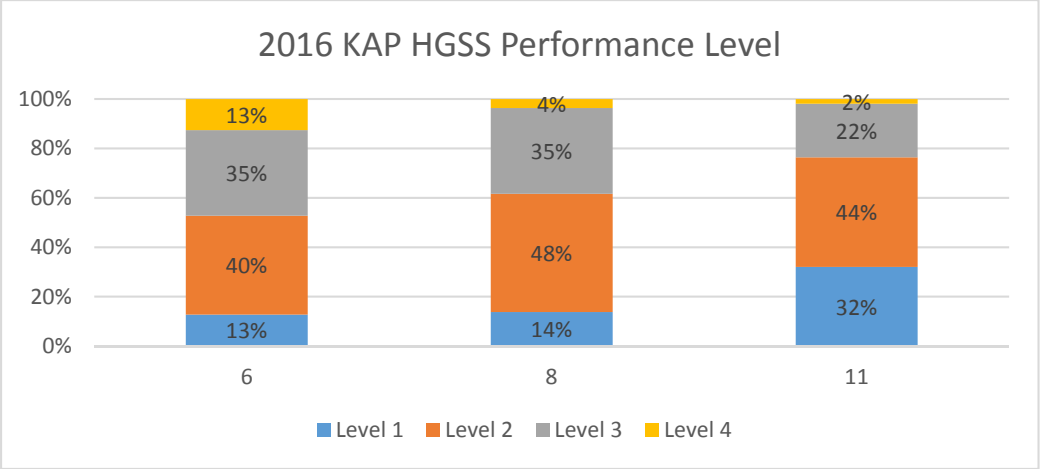


Figure IV-8. Performance-level results for HGSS.

ELA and mathematics scaled-score and performance-level trends across years are presented in Tables 4-27 and 4-28 and Figures 4-7 through 4-10. The tables present the scaled-score mean, standard deviation, and *N* count across administration years by grade. Figures 4-7 and 4-8 present the trends by performance level, and Figures 4-9 and 4-10 present passing-rate trends. The longitudinal trend cannot be computed for HGSS because 2016 is the first year of administration after adopting the KCCRS content standards. Figures 4-9 and 4-10 show that ELA passing rates declined in lower grade levels but increased in higher grade levels; mathematics

passing rates increased in all grades except grade 10.

Table IV-28. Longitudinal Scaled-Score Trend for ELA

	2015			2016		
Grade	*M*	*SD*	*N*	*M*	*SD*	*N*
3	298.3	24.7	37,723	298.7	28.0	38,370
4	303.6	24.9	37,200	303.0	29.3	37,366
5	298.6	25.0	36,965	298.2	29.3	36,803
6	292.9	24.7	37,270	293.1	28.4	36,732
7	291.3	25.1	36,875	292.7	27.9	36,589
8	285.9	24.7	36,784	286.7	28.5	36,193
10	286.7	24.7	35,593	285.3	29.6	35,653

Table IV-29. Longitudinal Scaled-Score Trend for Mathematics

	2015			2016		
Grade	*M*	*SD*	*N*	*M*	*SD*	*N*
3	303.2	24.4	37,740	304.3	27.8	38,343
4	293.0	24.7	37,261	293.3	28.2	37,448
5	292.2	24.5	36,986	292.2	27.4	36,806
6	292.6	23.9	37,268	292.2	27.2	36,657
7	289.6	24.0	36,878	289.4	28.4	36,583
8	285.7	23.9	36,821	285.5	28.4	36,169
10	285.7	23.7	35,603	285.0	28.2	36,831

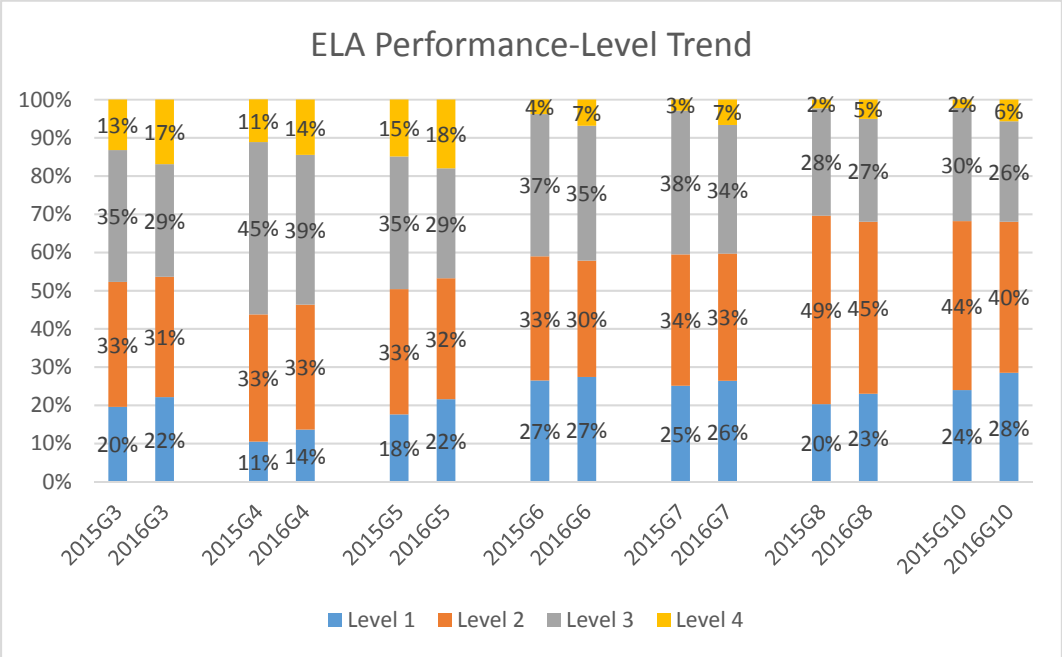


Figure IV-9. Performance-level trend for ELA.

G = grade.

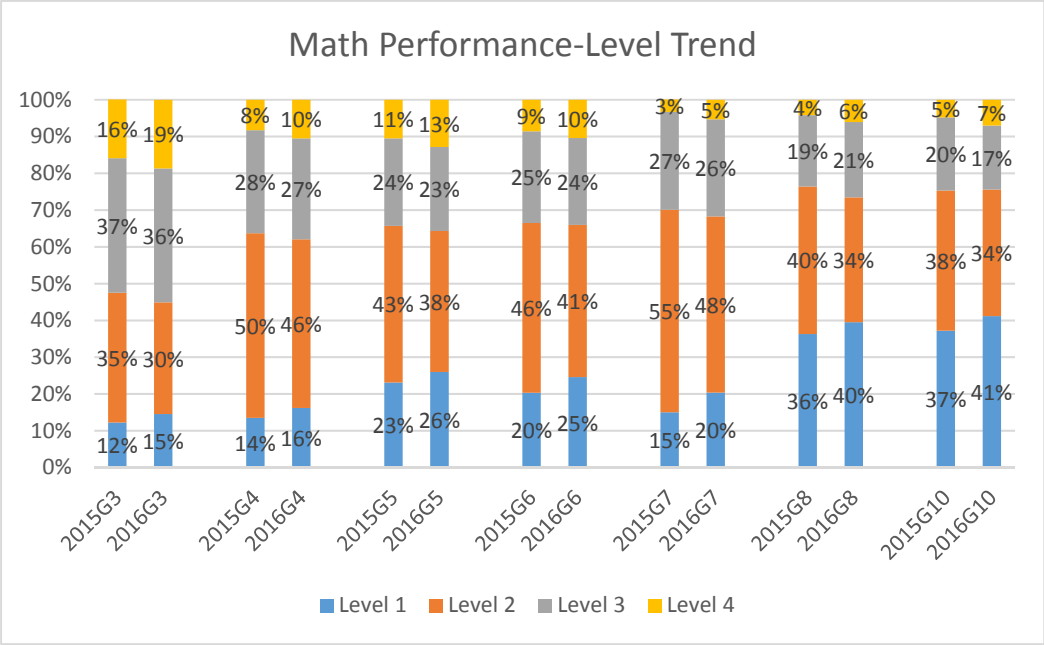


Figure IV-10. Performance-level trend for mathematics. G = grade.

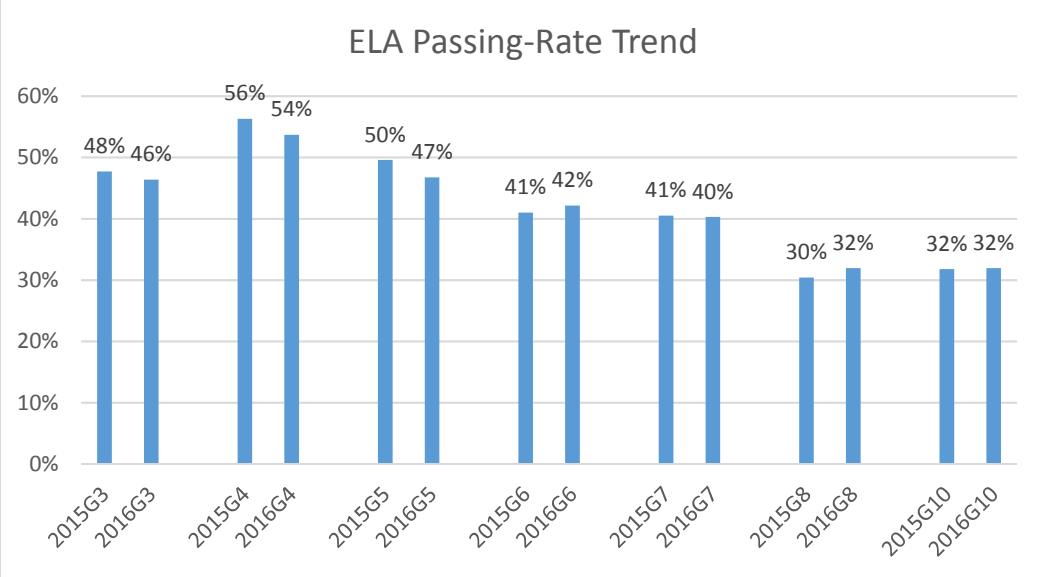


Figure IV-11. Passing-rate trend for ELA. G = grade.

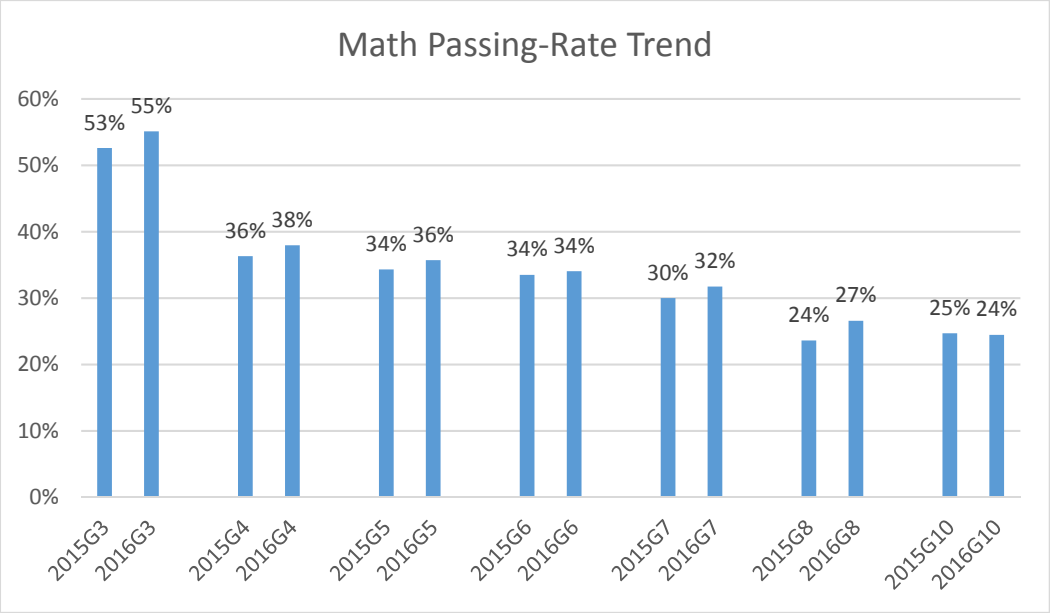


Figure IV-12. Passing-rate trend for mathematics. G = grade.

IV.5. Multiple Assessment Forms

In large-scale assessment programs, different item sets might be used on test forms both within and across years. Linking the scores from these different test forms puts the form scores on a common scale and ensures that all forms for a given grade level and subject area provide comparable scores. This outcome means that students will not have an unfair advantage or disadvantage simply because they took an easier or harder test form than other students did.

The 2016 KAP administration is the second year of KCCRS administration for ELA and mathematics. Both subjects have multiple operational forms, so their linking involves both within- and cross-year equating procedures. However, 2016 is the first year of KCCRS administration for HGSS; because each grade has only one form, there is no need for linking HGSS. Section 2.1.2 describes the test-design elements and associated data collection, as well as the analysis procedures related to linking scores from different KAP test forms.

IV.5.1. Within-year linking design. ELA and mathematics tests contain 55 items each: 25 items for Stage 1 and 15 items each for Stages 2 and 3. Out of 55 items, 20 items (about 36%) are designated within-year linking items. All within-year linking items are placed in Stage 1. The effect of content representation of linking items across forms is controlled by test blueprints: All blocks in the same stage are based on the same blueprint. This design ensures all forms have the same content coverage. Detailed descriptions of the test blueprints are provided in II: Test Design and Development.

IV.5.2. Cross-year linking design. To increase the number of linking items and maximize linking stability, cross-year linking utilizes both operational and field-test items from the 2015 administration. In other words, all items on the 2016 tests that have 2015 statistics are eligible for

cross-year linking. Table IV-32 summarizes the number of cross-year common items by subject and grade.

For ELA, about 92% of items are cross-year common items. The remaining items are ELA listening items, items whose scoring rules have changed, and items that have bad 2015 statistics. About 95% of mathematics items are cross-year common items. The remaining items are mathematics performance tasks and items with undesirable 2015 statistics.

Table IV-30. Cross-Year Common Item Counts for ELA and Mathematics

	ELA		Mathematics	
Grade	No. of items	No. of linking items	No. of items	No. of linking items
3	144	134	134	128
4	135	126	135	130
5	131	120	129	125
6	132	121	120	114
7	111	100	120	109
8	116	104	122	116
10	129	118	112	109

IV.5.3. Linking procedure. During concurrent calibration, the parameters of common items are fixed with 2015 values using flexMIRT (Cai, 2013; Houts & Cai, 2013). This procedure places the 2016 item parameters and consequent scaled scores directly onto the base scale (i.e., the 2015 scale).

IV.6. Multiple Versions of an Assessment

The KAP is administered online. All students who take the KAP must use the online administration portal. The only exceptions are the various accommodations provided to students with special needs, such as the Braille form.

IV.7. Technical Analysis and Ongoing Maintenance

KSDE intended to adopt new content standards and reduce student test burden. The adoption of new content standards requires the development of a new item pool. However, a healthy item pool that can cover all test construction needs takes years to build. Being in its second year after the adoption of new content standards, the item pool does not have a sufficient number of items to cover all content areas, thereby creating many challenges for test construction. For example, some item blocks did not yield a desirable information function and handicapped the path design. Because multistage test design is relatively new, the lack of models for validity studies, such as IRT model fit and dimensionality, becomes apparent. Extra effort has been expended in creating the best test and investigating the validity of this year's test. The same effort is also needed in developing and maintaining the quality of the KAP in future years.

V. Inclusion of All Students

KSDE complies with the Individuals with Disabilities Education Improvement Act (IDEA) and the Elementary and Secondary Education Act (ESEA), both of which require that all students, including students with disabilities, participate in assessments used for accountability purposes. One of the basic reform principles of ESEA is stronger accountability for educational achievement results for all students. Through this federal legislation, assessments that aim to increase accountability provide important information regarding (a) schools' success in including all students in standards-based education, (b) students' achievement of standards, and (c) improvements needed for specific groups of students. IDEA explicitly governs services provided to students with disabilities. Accountability at the individual level is provided through the Individualized Education Program (IEP) developed to address each student's unique needs.

Assessment accommodations are practices and procedures that provide equitable access during instruction and assessments for students with special needs. These accommodations may not alter the assessment's validity, score interpretation, reliability, or security. They are intended to reduce or eliminate the effects of a student's disability; however, they do not alter learning expectations. The accommodations provided to a student should be the same for classroom instruction, classroom assessments, and local educational agency and state assessments. It is critical to note that some accommodations that are appropriate for instructional uses may not be appropriate for use on standardized assessments. For example, a student with low vision will need accommodations to make a test accessible. However, in an ELA assessment, reading passages aloud to a student would change what is being measured and therefore is not a valid accommodation. Use of a magnifying tool or a large-print version of a test is an acceptable accommodation, though. It is very important for educators to become familiar with state policies regarding accommodations during assessments.

This chapter presents information about the KAP's inclusion of all students and accommodation usage. Much of this information is also available in other KSDE documents (e.g., *Tools and Accommodations for the Kansas Assessment Program (KAP) 2015–2016* and the *Kansas Assessment Examiner's Manual 2015–2016*). This chapter closes with a report of the frequency of use of specific accommodations.

V.1. Procedures for Including Students with Disabilities

KSDE is committed to including all students in the KAP assessments. The inclusion of students with disabilities is achieved by providing clear guidelines for educators to register their students with different needs. The *Examiner's Manual* describes step-by-step registration procedures for students who need accommodations. Additionally, educators are instructed to report students who are not assessed. Some notable exceptions that occur in Kansas include:

- students serving a long-term suspension,
- students who were truant more than two consecutive weeks at time of testing,
- students who had a catastrophic illness or accident,
- students who moved during testing, or
- students who were incarcerated.

V.2. Accommodations

A few basic rules apply to every available accommodation on the KAP. First and foremost, only accommodations that have been used regularly in instruction may be used on state assessments. Second, students with IEPs, 504 plans, or ELL plans, as well as students with Student Improvement Team plans, may use only the accommodations documented in their plans. Finally, for accommodations to be available during the KAP, teachers must submit accommodation requests through the student's PNP in the Educator Portal.

Some accommodations allowed for the KAP are handled by test administrators, but some are built-in features in the KITE system. Because features in KITE are activated according to students' needs, teachers are required to mark those needs in the PNP. Additionally, teachers need to report in advance if Braille is needed. Table V-2 shows available accommodations according to reporting requirement.

Table V-1. Available Nonreported and Reported KAP Accommodations

Nonreported	Reported
Allowable practice	Auditory background
Delivery of directions to student in ASL	Background color
Frequent breaks	Braille
Separate, quiet, or individual setting	Color overlay
Spanish translation	Foreground color
Student dictation of answers to scribe	Magnification
Student reading assessment aloud to self (via headset)	Invert color choice
Student response in American Sign Language (ASL)	Item-translation display
Student use of Braille writer or slate and stylus	Keyword-translation display
Student use of communication device	Large-print booklet
Student use of translation dictionary	Masking
Text-to-speech	Onscreen keyboard
Use of some other accommodation	Signing
	Speech (read aloud)
	Touch

V.3. Frequency of Accommodation Use

The PNPs submitted by teachers determine the availability of online test accommodations for individual students. Thus, the summary of PNP accommodation requests below also indicates the number of students for whom each accommodation is requested. Tables V-3 through V-5 summarize the PNPs by subject and grade; note that some students may receive multiple accommodations. All tables show that "Spoken (read aloud)" is the most commonly used accommodation.

Table V-2. Frequency of ELA Accommodation Requests by Grade

	Grade						
Accommodation	3	4	5	6	7	8	10
Auditory background	30	24	65	20	33	17	23
Background color	10	11	5	11	5	1	10
Braille	4	5	4	3	5	2	5
Color overlay	18	32	54	34	18	4	8
Foreground color	10	11	5	11	5	1	10
Magnification	28	19	13	14	18	10	15
Invert color choice	1	5	4	5	5	5	6
Item-translation display	10	20	16	21	31	30	23
Keyword-translation display	15	24	25	85	79	120	51
Large-print booklet	0	11	12	8	7	6	2
Masking	4	15	20	22	17	8	2
Onscreen keyboard	5	13	10	19	11	8	14
American Sign Language	11	14	9	6	10	13	11
Speech (read aloud)	3,512	3,781	2,995	3,377	3,277	2,094	1,711
Touch	0	2	2	0	0	0	1
Total	3,658	3,987	3,239	3,636	3,521	2,319	1,892

Table V-3. Frequency of Mathematics Accommodation Requests by Grade

	Grade						
Accommodation	3	4	5	6	7	8	10
Auditory background	30	24	65	20	34	18	24
Background color	11	11	5	11	5	1	10
Braille	4	5	4	4	5	2	4
Color overlay	19	32	54	34	18	4	9
Foreground color	11	11	5	11	5	1	10
Magnification	28	18	14	14	18	10	15
Invert color choice	1	5	4	5	5	5	6
Item-translation display	14	25	23	40	44	40	31
Keyword-translation display	24	37	35	110	103	134	67
Large-print booklet	0	11	12	8	7	6	1
Masking	4	15	20	21	17	8	3
Onscreen keyboard	5	13	10	18	11	8	15
American Sign Language	11	14	9	5	10	13	10
Speech (read aloud)	3,549	3,800	3,011	3,383	3,292	2,108	1,734
Touch	0	2	2	0	0	0	1
Total	3,711	4,023	3,273	3,684	3,574	2,358	1,940

Table V-4. Frequency of HGSS Accommodation Requests by Grade

Accommodation	Grade 6	Grade 8	Grade 11
Auditory background	20	18	16
Background color	12	1	1
Braille	4	2	1
Color overlay	35	4	2
Foreground color	12	1	1
Magnification	15	10	6
Invert color choice	6	5	1
Item-translation display	26	32	26
Keyword-translation display	89	122	42
Large-print booklet	9	6	1
Masking	22	8	3
Onscreen keyboard	19	8	1
American Sign Language	8	14	15
Speech (read aloud)	3,424	2,144	1,486
Touch	0	0	0
Total	3,701	2,375	1,602

VI. Academic Achievement Standards and Reporting

VI.1. State Adoption of Academic Achievement Standards for All Students

Performance level descriptors (PLDs) define the KAP academic achievement standards. While a test is developed according to content standards, students' performances are evaluated using the academic achievement standards. PLDs describe the expected academic performances at each performance level. When a performance level is assigned to a student, it means the student meets the minimum expected performances of the performance level. This score interpretation applies to all students who participated in the KAP assessment.

VI.2. Achievement Standard Setting

ELA and mathematics standard setting occurred in 2015. The procedures and outcomes can be found in the 2015 technical manual. CETE conducted HGSS standard setting in Lawrence, KS, on May 24 and May 25, 2016. This section focuses on HGSS standard setting. The main goal of the HGSS standard setting was to establish three cut scores that differentiate four performance levels for the assessment. The panelists' recommended cut scores and item weights were presented to the State Board for approval.

VI.2.1. Overview of the analytical judgment method.

The analytical judgment method (AJM) grew out of the need for a way to set cut scores that would appropriately separate examinees into ordered performance categories either on constructed-response items exclusively or in some combination with multiple-choice items (Plake, 1998; Plake & Hambleton, 2001). Different from the typically used threshold performance approach, AJM uses the adjacent borderline groups—the lower end of an upper category and the upper end of a lower category—to determine cut scores. For example, HGSS has four performance levels (Level 1 through Level 4). When applying AJM, each level was divided into *low, middle,* and *high* subcategories, resulting in 12 separate and distinguishable categories. After reviewing all items on the test, panelists independently recorded an expected total test score for each of the 12 subcategories. For the two borderline groups, the medians of panelist-assigned scores are used as cut scores for the adjacent levels.

VI.2.2. Panelist recruiting process. KSDE took several steps to recruit panelists that represent the variety of the Kansas educator population for the standard-setting workshop. In order to obtain a large and diverse pool of applicants, KSDE began recruitment efforts early in the year. Invitation was sent to all teachers and administrators in the current educator database and the invitation was extended to those educators' colleagues in case some educators were not in the database. Additional recruitment efforts were also made through relationships with school district and individual educators. When selecting panelists from the applicant pool, KSDE reviewed all applications and put emphasis on ethnic, gender, and geographic diversity.

VI.2.3. Performance level descriptors (PLDs). As mentioned earlier, PLDs describe the expected academic performance standards at each performance level. Thus, PLDs are the guiding performance standards when setting cut scores. The creation of HGSS PLDs started with CETE

content staff, who developed descriptors for the content that all students should know and be able to achieve at each performance level. These descriptors adhered to the cognitive alignment of the content standards, such as depth of knowledge, cognitive complexity, scope of skills, inquiry vs. process, etc. (see Appendix F). KSDE staff and Kansas educators reviewed and approved the grade-specific PLDs for all four levels prior to the standard-setting workshop.

VI.2.4. Standard-setting procedure. The standard-setting activities described in this section follow the event chronological order, as laid out by the meeting agenda (see Appendix G). Each grade had one panel and each panel had eight to eleven panelists. The workshop was steered by a lead facilitator and three table leads recruited from CETE. Both KSDE assessment personnel and CETE content team members were available at the workshop to address policy- or content-related questions. A description of the workshop structure follows.

On May 24, set cut scores by grade for the two machine-scorable test components:

- Complete the training session
- Complete the HGSS exam and reviewing items
- Complete the PLD workbook task
- Introduce of the possible total score point sheet
- Practice the rating task
- Set cut scores: Round 1
- Set cut scores: Round 2

On May 25, review the MDPT component and assign weights to each test component:

- Review MDPTs
- Assign weights

The benefits of structuring the meeting this way were twofold. First, it allowed those involved in the meeting to focus on rating the machine-scorable items on the first day and on assigning weights on the second day. Second, cut-score results from the first day were made available for the second-day weighting procedure.

VI.2.4.1 ***Training session.*** At the start of the meeting, panelists completed a participant survey form (see Appendix H) and signed a confidentiality form (see Appendix I). The survey collected panelist biographical data to contribute to the documentation of the procedural validity of the standard-setting process (Hambleton, Pitoniak, & Copella, 2012; Pitoniak & Morgan, 2012; Rosseel, 2012). Then, the lead facilitator conducted a large group training to address general topics that included an overview of the HGSS assessment and an introduction to the concept of cut scores. The small group training followed the large group training were given by table leads. In small group training, table leads emphasized the tasks to be performed and ensured panelists the assistances they would receive. Table leads also answered standard-setting related questions generated from panelists at their tables; however, policy-related questions were directed to KSDE staff.

Table VI-1 shows the demographic composite of panels by grade. Despite the efforts put into recruiting a diversified panelist group, the results were lower than desired.

Table VI-1. Summary of HGSS Panelists Demographic Information

		Grade		
		6	8	11
Demographics	Categories	% (n = 11)	% (n = 11)	% (n = 8)
Gender	Male	27.27	45.45	62.50
	Female	72.73	54.55	37.50
Race/ethnicity	Native American	0.00	0.00	0.00
	Asian/Pacific Islander	0.00	0.00	0.00
	Black	0.00	0.00	0.00
	Hispanic or Latino	0.00	9.09	0.00
	White	100.00	90.91	100.00
	Other	0.00	0.00	0.00
Teaching experience	1–3 years	18.18	9.09	0.00
	4–6 years	18.18	18.18	12.50
	7–12 years	27.27	18.18	37.50
	>12 years	36.36	45.54	50.00
Current assignment	Classroom teacher	90.91	90.91	100.00
	Educator (nonteacher)	9.09	0.00	0.00
	Other	0.00	9.09	0.00
Work setting	Urban	9.09	27.27	12.50
	Suburban	18.18	36.36	12.50
	Rural	72.72	27.27	75.00

VI.2.4.2 *Completing the HGSS exam.* In order to provide a frame of reference for considering student performances, the panelists took the HGSS test in a shorter timeframe than was used operationally. Following completion of the exam, panelists received an answer key so they could grade their own machine-scored items. Although panelist performances on the exam were not recorded, panelists were free to share their performance with other members of the panel during discussion.

VI.2.4.3 *Review of items.* Panelists then had an opportunity to review and discuss items that they found especially difficult or confusing, with an emphasis on characteristics of those items. Panelists were reminded that the purpose of this conversation was to discuss their perceived difficulty of items in the context of the entire HGSS exam, rather than an opportunity

to critique the items or the test. Any comments of a critical nature or editorial type beyond the scope of the standard-setting task were collected and shared with the appropriated person(s) in KAP. During this discussion, panelists were encouraged to refer to the PLDs.

VI.2.4.4 *PLD workbook task.* The PLDs provided the expected competencies of students in the state for specific grades. They are general because they represent a wide range of content knowledge and skills. Before proceeding to ratings, panelists interacted with PLD workbooks via a guided exercise. Specifically, panelists identified the PLDs for low, middle, and high subcategories for Level 2, Level 3, and Level 4 to achieve a common understanding of what students in those subcategories know and are able to learn. CETE prepared the PLD workbooks for this process. For each grade, there were three workbooks, one for each performance level. Each PLD statement was listed at the top of a box, along with three smaller, empty boxes for the panelists to write in statements for low, mid and high subcategories (See Figure VI-1). To help panelists conceptualize Level 1 High, a box for Level 1 High was provided on the left of each of the Level 2 PLD statement (See Figure VI-2).

Panelists were reminded that their task was not to edit or revise the PLDs but to use the PLDs to focus on the specific abilities of the three subcategories (low, middle, and high). By focusing on the subcategories, panelists identified the knowledge and abilities for each of the subcategories, helping them further understand PLDs and become familiar with the AJM. In addition, panelists were directed to make sure that their statements were item and content oriented rather than student oriented.

Grade-level groups discussed the ideas generated by individual panelists and used them to develop a group list. The table lead combined the lists for each performance level and distributed the notes to the panelists. These notes established a common understanding among panelists about the performance-level knowledge and skills of students in the subcategories. Later in the event, panelists were instructed to consider these notes when making their ratings.

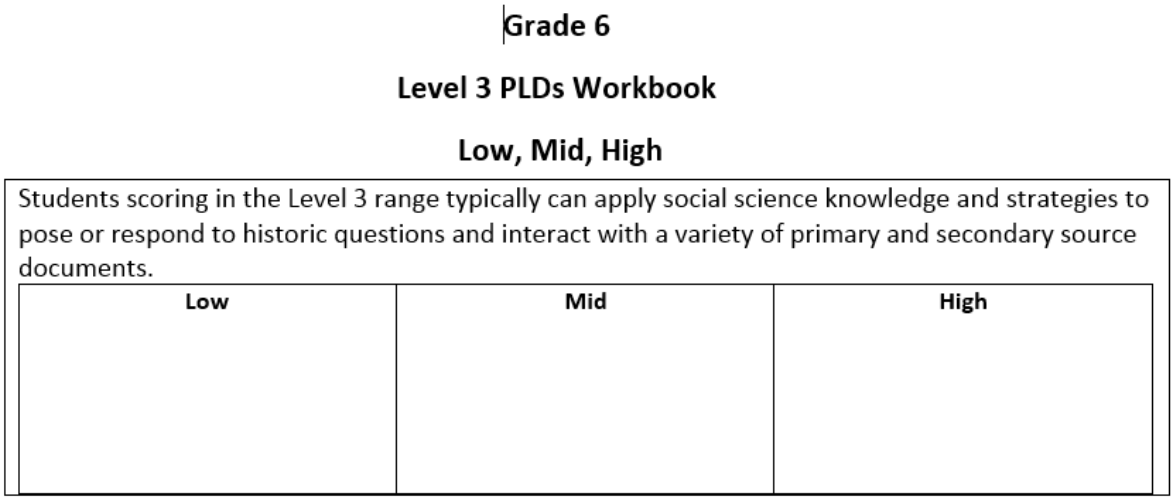


Figure VI-1. An example of a Level 3 PLD workbook.

Level 2 PLDs Workbook

Low, Mid, High

Level 1 High	Students scoring in the Level 2 range typically can apply social science knowledge to respond to basic historic questions and interact with a variety of familiar primary and secondary source documents.

Low	Mid	High

Figure VI-2. An example of a Level 2 PLD workbook.

VI.2.4.5 ***Introduction of the possible total score points sheet.*** After identified the PLDs for each subcategories, panelists started assigning scores for each of the 12 subcategories. A sheet of the possible total score points a student could receive on the test (see Table VI-2) was distributed to panelists to assist them to perform this task.

Table VI-2. Possible total score points

Possible Total Score Points												
Grade	**Part One**								**Part Two**			
6	0.00	0.33	0.50	0.67	0.83	1.00	1.17	1.33	0.00	0.33	0.67	1.00
	1.50	1.67	1.83	2.00	2.17	2.33	2.50	2.67	1.33	1.67	2.00	2.33
	2.83	3.00	3.17	3.33	3.50	3.67	3.83	4.00	2.67	3.00		
	4.17	4.33	4.50	4.67	4.83	5.00	5.17	5.33				
	5.50	5.67	5.83	6.00	6.17	6.33	6.50	6.67				
	7.00											
8	0.00	0.33	0.50	0.67	0.83	1.00	1.17	1.33	0.00	0.33	0.50	0.67
	1.50	1.67	1.83	2.00	2.17	2.33	2.50	2.67	0.83	1.00	1.17	1.33
	2.83	3.00	3.17	3.33	3.50	3.67	3.83	4.00	1.50	1.67	1.83	2.00
	4.17	4.33	4.50	4.67	5.00				2.17	2.33	2.50	2.67
									3.00			
11	0.00	0.33	0.67	1.00	1.33	1.67	2.00	2.33	0.00	0.25	0.50	0.75
	2.67	3.00	3.33	3.67	4.00	4.33	4.67	5.00	1.00	1.25	1.50	1.75
	5.33	5.67	6.00	6.33	6.67	7.00	7.33	7.67	2.00	2.25	2.50	2.75
	8.00								3.00			

VI.2.4.6 ***Practicing the rating task.*** After panelists were familiar with testing materials, and had assigned PLDs and total scores into low, middle, and high categories, they were trained on the AJM ratings. A variety of activities, such as presentation, rating practice, and discussion, were used to train panelists, and evaluate understanding prior to the collection of actual standard-setting ratings. Training materials included PowerPoints and handouts.

The rating practice started with the lead facilitator presenting the concept of a borderline examinee. Using the PLDs and expected total score points assigned to the subcategories, the panelists would rate the machine-scored items in the 12 categories. They were encouraged to move from low to high on the rating form and were reminded that any given cell number must be equal to or greater than the value in the lower category. Upon completion of the instructions, panelists practiced the rating method using five machine-scored items from the HGSS exam. After the training round of ratings, panelists shared their ratings with the group. A brief discussion followed to go over any discrepancies in ratings, with emphasis on relating the discussion to the PLDs. They were reminded that the standard-setting panel was designed to represent a variety of perspectives needed to inform decisions about cut scores and that consensus was not a goal.

Following the discussion, panelists completed the readiness form (Appendix K), where they indicated whether they felt ready to proceed to the operational task. After confirming that the panelists were ready to proceed, the operational standard-setting task began. Next, two rounds of ratings were collected. Discussion and feedback was provided between rounds.

VI.2.4.7 ***Setting cut score: Round 1.*** Round 1 ratings covered the entire set of machine-scored items on the HGSS exam. Table VI-3 presents the medians of Round 1 cut scores among panelists for grades 6, 8, and 11.

Table VI-3. Round 1 Medians of Cut Scores by Grade

	Grade 6 level			Grade 8 level			Grade 11 level		
Part	2	3	4	2	3	4	2	3	4
1	1.17	2.34	4.17	1.08	2.17	3.75	1.17	4.50	6.84
2	0.50	1.50	2.17	0.67	1.50	2.33	0.63	1.63	2.63

Round 1 individual ratings and the cut score summary were distributed among panelists to facilitate a group discussion of the rationale behind the ratings. The discussion started with items with the most discrepancies. As appropriate, panelists were encouraged to discuss the observed item difficulty and individual item characteristics, such as format and content. Panelists were also given the opportunity to discuss any items, particularly those they found difficult to rate. In addition, the lead facilitator chose several items with discrepant ratings for discussion. After sufficient discussion, panelists proceeded to Round 2 rating.

VI.2.4.8 ***Setting cut score: Round 2.*** Panelists were informed that the ratings collected during Round 2 would form part of the cut score recommendations given to KSDE after the standard-setting meetings. During Round 2, panelists were instructed to review each item to confirm their Round 1 ratings or to provide appropriate, new ratings based on the Round 1 discussion. Round 2 individual ratings and the cut score summary were then presented to panelists for another discussion. Table IV-4 summarizes the medians of Round 2 cut scores.

Table VI-4. Round 2 Medians of Cut Scores by Grade

	Grade 6 level			Grade 8 level			Grade 11 level		
Part	2	3	4	2	3	4	2	3	4
1	1.17	2.50	3.92	1.08	2.25	3.75	1.67	4.34	6.57
2	0.50	1.50	2.17	0.67	1.50	2.42	0.75	1.63	2.63

VI.2.4.9 *Review of MDPTs.* The MDPTs were scored from 1 to 4, and the scoring rubrics were drafted based on the PLDs (see Appendix F). Provided with sample student responses, panelists reviewed and discussed items that they found especially difficult or confusing, with an emphasis on characteristics of specific items.

VI.2.4.10 *Assigning weights.* The HGSS test includes two sets of machine-scored items and one set of MDPTs. During the weight-assigning procedure, panelists decided the appropriate proportion for each component's contribution to the final HGSS score. After determining the weights in their assigned grades, panelists participated in articulation to balance the section weights among grades. Table VI-5 presents the panelist recommendation of percentage weights for grades 6, 8, and 11 after articulation.

Table VI-5. Panelist Recommended Weights by Grade

	Grade 6 %	Grade 8 %	Grade 11 %
Part 1	35	45	20
Part 2	40	30	30
MDPT	25	25	50

VI.2.5. Final evaluation. Before leaving the workshops, panelists completed an evaluation form (see Appendix L). A summary of responses is presented in Table VI-6. The question sequence in Table VI-6 is organized by rating scale rather than the sequence used in the survey. Additionally, some of the questions in Table VI-6 were abbreviated for presentation purposes. Refer to Appendix L for the actual survey.

The evaluation results indicate panelists were comfortable with the process, understood its purpose and their role, and were able to follow the procedures. Group discussion and facilitation of the meetings were helpful in guiding them through the process. As for the recommended cut scores, panelists generally felt comfortable with the final recommendations after two rounds of ratings. Panelists also expressed their appreciation for the opportunity to participate in the standard-setting events.

Table VI-6. Summary of Evaluation Survey

Questions	Means		
Grade	**6**	**8**	**11**
Strongly Disagree (1) to Strongly Agree (6)			
Opening Session			
Adequate background of the assessment.	4.7	5.3	4.4
Clear understanding of the meeting purpose.	4.7	5.5	4.7
Appropriate context for my role in the meeting.	4.5	5.4	4.8
Addressed many of my questions and concerns.	3.7	5.0	4.4
Well organized.	4.7	5.5	4.8
Presenter clearly explained the procedures.	4.8	5.5	4.8
Taking the Test			
Taking the test helped me understand the assessment.	5.0	4.5	5.2
Practice			
The practice activity helped me understand my task.	4.5	4.9	5.0
After the practice, I understood my role in the event.	4.6	5.2	5.1
Staff effectively answered any questions that were asked during practice.	4.5	5.2	5.2
Round 1 Experience			
The rating form was easy to understand.	5.0	5.5	5.2
The expectations for Round 1 were made clear.	5.0	5.5	5.1
I made my ratings independently.	5.4	5.8	5.7
I understood the tasks I was to accomplish for Round 1.	5.2	5.7	5.2
I had the right amount of time to complete the tasks during Round 1.	5.5	5.6	5.6
Round 1 Results			
The Round 1 results were clear.	5.1	5.8	5.4
The Round 1 results were useful.	4.9	4.7	5.5
Round 2 Experience			
The expectations for Round 2 were made clear.	5.0	5.7	5.4
I made my ratings independently.	5.3	5.9	5.5
I understood the tasks I was to accomplish for Round 2.	5.1	5.9	5.5
I had the right amount of time to complete the tasks during Round 2.	5.1	5.9	5.6
Round 2 Results			
The Round 2 results were clear.	4.8	5.6	5.6
The Round 2 results were useful.	4.8	5.6	5.6
Group Discussion			
The group discussions aided my understanding of the issues.	4.7	5.5	5.7
The time provided for discussions was adequate.	5.1	5.5	5.3
Everyone had equal opportunity to contribute ideas and opinions.	4.8	5.5	5.3
The discussions about the Performance Level Descriptors (PLDs) were	4.7	5.5	5.3

Questions	Means		
Grade	**6**	**8**	**11**
Strongly Disagree (1) to Strongly Agree (6)			
helpful to me.			
The discussions about the Low, Middle, and High student were helpful to me.	4.7	5.6	5.2
MDPT			
The scoring rubric helped me understand the role of the MDPT in the HGSS assessment.	4.6	5.4	5.4
The sample responses helped me understand the role of the MDPT in the HGSS assessment.	5.0	5.3	5.4
Test Part Weights			
The final weight for the MDPT part is reasonable.	3.9	4.7	5.3
The final weight for machine-scorable Part 1 is reasonable.	4.0	5.4	5.2
The final weight for machine-scorable Part 2 is reasonable.	4.6	5.3	5.2
Results for Level 2 Cut Score			
The cut score is appropriate based on the PLDs and the Low, Middle, and High student activities.	5.0	5.5	5.2
The cut score for this achievement level is defensible due to panelists' adherence to procedures.	4.9	5.5	5.2
Results for Level 3 Cut Score			
The cut score is appropriate based on the PLDs and the Low, Middle, and High student activities.	5.0	5.4	4.8
The cut score for this achievement level is defensible due to panelists' adherence to procedures.	5.0	5.5	5.0
Results for Level 4 Cut Score			
The cut score is appropriate based on the PLDs and the Low, Middle, and High student activities.	4.9	5.4	5.1
The cut score for this achievement level is defensible due to panelists' adherence to procedures.	4.9	5.5	5.1
Too Little (1) to Too Much (3)			
Opening Session Time			
The amount of time used for the opening session	2.2	2.0	2.1
No Understanding (1) to Complete Understanding (4)			
Performance Level Descriptors (PLDs)			
Level 2	2.8	3.5	3.3
Level 3	2.8	3.5	3.3
Level 4	2.9	3.5	3.2
Low, Middle, High Student Activity			

Questions	Means		
Grade	6	8	11
Strongly Disagree (1) to Strongly Agree (6)			
Level 2 Low, Middle, and High	2.9	3.3	3.0
Level 3 Low, Middle, and High	2.9	3.3	3.1
Level 4 Low, Middle, and High	2.9	3.3	3.1
Not Important (1) to Very Important (4)			
Influential Factors for Round 1			
PLDs for Level 2, Level 3, and Level 4	3.2	3.5	3.8
Your perceptions of the Low, Middle, and High students	3.4	3.7	3.7
Your perceptions of the difficulty of the items	3.6	3.6	3.6
Your experience with students at your grade level	3.7	3.7	3.8
Your experience with the Kansas HGSS Standards	3.1	3.5	3.4
Influential Factors for Round 2			
PLDs for Level 2, Level 3, and Level 4	3.4	3.4	3.8
Your perceptions of the Low, Middle, and High students	3.3	3.6	3.8
Your perceptions of the difficulty of the items	3.6	3.8	3.6
Your experience with students at your grade level	3.3	3.8	3.8
Your experience with the Kansas HGSS Standards	3.1	3.3	3.5
Your Round 1 results	3.0	3.9	3.4
The Round 1 results of the other panelists	3.0	3.6	3.6
Group discussions about the Round 1 Results	2.7	3.8	3.8
Not Successful (1) to Very Successful (4)			
Successfulness of Agenda Items			
Opening session	3.2	3.5	3.6
Taking the operational test	3.5	3.8	3.9
Review of the PLDs	3.0	3.5	3.6
Discussions about the Low, Middle, and High student groups	3.2	3.7	3.6
Practice activities	3.4	3.6	3.6
Discussions after Round 1	3.4	3.9	3.7
Not Helpful (1) to Very Helpful (4)			
CETE Staff			
Orientation Leader	3.6	3.9	3.8
Facilitator	3.3	3.9	3.7
Table Leaders	3.5	3.9	3.9
Content Specialists	3.8	3.8	3.8

VI.2.6. Setting the final cuts. Round 2 ratings and weights, as well as impact data, were provided to KSDE for review and approval. The information was then presented to the School Board for final approval. The final cuts for three sections of each grade are shown in Table VI-7, and the weights are shown in Table VI-8.

Table VI-7. Final Cut Score Summary by Grade

	Grade 6			Grade 8			Grade 11		
	Level			Level			Level		
	2	3	4	2	3	4	2	3	4
Part 1	1.17	2.50	3.92	1.08	2.25	3.75	1.67	4.34	6.57
Part 2	0.50	1.50	2.17	0.67	1.50	2.42	0.75	1.63	2.63
MDPT	2	3	4	2	3	4	2	3	4

Table VI-8. Final Percentage Weights Summary by Grade

	Grade 6 %	Grade 8 %	Grade 11 %
Part 1	35	45	20
Part 2	40	30	30
MDPT	25	25	50

VI.3. Challenging and Aligned Academic Achievement Standards

The KAP grade-level academic achievement standards are drafted to align with the state content standards: the KCCRS for ELA, mathematics, science, and HGSS. CETE content experts worked alongside KSDE staff to define the PLDs. The iterative process ended when both sides agreed that the expected performances adhered to KCCRS content standards, as well as to cognitive demands, and the overall expectation properly reflected the rigor of the KCCRS. Then, the PLDs were presented to Kansas educators for review and approval. As described in the HGSS standard-setting section, PLDs are the basis of the cut scores.

VI.4. Reporting

For each tested subject, the KAP provides score reports to students, schools, and districts separately (see Appendix M). The content of these reports includes overall performance and performance by content standards. These statistics are presented using various graphs, colors, and symbols so they are easy to read. In order to assist readers in interpreting the information in the reports, descriptions of what students should be able to do in each subject area are presented with the statistics. As stated by Peterson, Kolen, and Hoover (1989), add score interpretations in score reports can minimize misinterpretations and unwarranted inferences. It is as important to help readers understand the meaning of the statistics as it is to report the values.

Although these reports are intended for different groups (e.g., students, schools, and districts), the content of these reports is uniform. Presentation and text are adjusted by according to group, but the symbols and interpretation of those symbols are consistent across reports. This design eases educators' reviewing burden and helps them explain score reports to parents.

VI.4.1. Group masking. When group *n* counts are very small, individual students may be identified through demographic information, even on roll-up summary reports. Various types of suppression logic are used to protect individual identities. One way is to report student score results by percentage ranges instead of the actual observed percentages. For example, if only one student in a group of five students is in Level 4, the group's actual percentage is 20%. In a roll-up summary, however, the report gives a range of percentages instead (e.g., 0%–40%). The other way is to suppress reports when the number of students on the group is small.

VI.4.2. Student reports. A sample of an ELA student report is presented in Appendix M. In the report, a student's performance level is placed immediately after student identifiers so that it is the first information presented. Next are the student's scaled score and comparisons with students in the same school, district, and state (i.e., the score meters), as well as a brief summary of the PLDs that describe what this student should be able to do. Score meters report the medians of school, district, and state performances. The median is used because it is more robust to outliers than the mean in describing the central tendency of a group.

A student's overall-score performance level represents a student's performance on all sections of the test. In mathematics and HGSS, the overall-score performance level is the same as the performance level displayed on the meter. In ELA, the overall score level combines the student's score level on the reading, writing, and listening section, as well as the on-demand writing task section score (MDPT). Students must complete both sections of the ELA test to receive an overall score level. The scaled score and performance level shown on the meter of the ELA student report include scores from reading, writing, and listening. The third section of the report is an overall description of PLDs for each performance level.

The first section of the second page reports the student's performance by content standards. This information indicates strengths and weaknesses on different claim or target. Each claim/target represents a group of test items that assess related skills. Some items of a test are counted in multiple categories. Subscore information is not available for HGSS reports.

The bottom of the second page shows the standard error of scaled scores and standard errors of school, district, and state median scores. The standard error reported on student scores is the CSEM derived from the IRT scaling model. It indicates how much a student score might vary if the student took many equivalent versions of the test. Standard errors of group scores (school, district, and state) account for sampling error but not for measurement error.

The median of the standard error is computed using equation VI-1. It is equivalent to standard error for the mean but multiplied by an extension factor of 1.253 to account for the additional sampling variability of the median.

$$SE_{median(x)} = 1.253 * \frac{S_{\bar{x}}}{\sqrt{N}}, \qquad \text{(IV-1)}$$

where $SE_{median(x)}$ is the standard error of the median of the group scores, $S_{\bar{x}}$ is the standard

deviation of the group's observed scores, and *N* is the number of students in the group.

VI.4.2.1 ***Timeline for delivering student reports.*** The KAP testing window ended on May 12, 2016. In July 2016, student reports were available for all students who took the KAP ELA and mathematics tests. The standard-setting event and the required approval of cut scores delayed the delivery of HGSS student reports to educators from summer to early fall 2016.

VI.4.3. School and district reports. While student reports focus on individual student performance, school and district reports focus on group-level performances. Information provided in the school and district reports aggregates student performances at the given level (see Appendix M).

School reports provide summary information of the same subject by grades. On the first page, bar graphs show a school's median scaled scores of three grades, along with scores of the school's district and state overall performances. District and state median scaled scores are reference for schools to interpret their standings. Standard errors are given at the bottom of the first page. The second page shows the percentage of students in each of the four performance levels; again, district and state results are provided for reference. The floating bar graphs use four different colors to represent the different performance levels, allowing readers to distinguish performance-level outcomes instantly. The next section of the school report presents the school's performance by claim/target: student performances by content standards and a summary of students' relative strengths and weaknesses in the different content standards.

District reports use the same layout and provide the same information as school reports; however, only state data are provided as the reference group.

VI.4.4. Interpretive guides. Besides adding descriptions to score reports, two score interpretive guides, *2016 Educator Guide—Understand the Kansas Assessment Program Score Report* and *2016 Parent Guide—Understanding the Kansas Program Score Report*, are available for educators and parents to download from the KAP website.

http://ksassessments.org/sites/default/files/documents/educator_guide.pdf
http://ksassessments.org/sites/default/files/documents/parent_guide.pdf

VI.4.5. Letters from the Commissioner of Education. The letters to Kansas educators and parents from Dr. Randy Watson, Kansas Commissioner of Education, are an important part of the interpretive guides. Copies of these two letters are provided in Appendix N.

References

American Psychological Association, American Educational Research Association [AERA], & National Council on Measurement in Education (2014). *Standards for educational and psychological testing*. Washington, DC: AERA.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397–472). Reading, MA: Addison Wesley.

Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika, 52,* 345–370.

Brennan, R. L. (2004). *Manual for BB-CLASS: A computer program that uses the Beta-binomial model for classification consistency and accuracy, Version 1.1 (Center for Advanced Studies in Measurement and Assessment Research Report No. 9)*. Iowa City: University of Iowa.

Cai, L. (2013). *flexMIRT version 2: Flexible multilevel multidimensional item analysis and test scoring* [Computer software]. Chapel Hill, NC: Vector Psychometric Group.

Common Core State Standards Initiative. (2010). *Common core state standards for English language arts and literacy in history/social studies, science, and technical subjects.* Retrieved from http://www.corestandards.org/assets/CCSSI_ELA%20Standards.pdf.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Orlando, FL: Holt, Rinehart and Winston, Inc.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16,* 297–334.

Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational measurement*, pp. 443–507.

Forte, E. (2013, June). *Next generation alignment approaches: Needs and promising directions.* Paper presented at the National Conference on Student Assessment, Baltimore, MD.

Forte, E. (2016). *Evaluating alignment in large-scale standards-based assessment systems*. Washington, DC: Technical Issues in Large Scale Assessment SCASS of CCSSO.

Forte, E., Nebelsick-Gullett, L., Deters, L., Buchanan, E., Herrera, B., Morris, J, & Phlegar, J. (2016). *Kansas Assessment Program alignment evaluation report 2015-2016*. edCount LLC.

Green, B. F., Bock, R. D., Humphreys, L. G., Linn, R. L., & Reckase, M. D. (1984). Technical guidelines for assessing computerized adaptive tests. *Journal of Educational Measurement, 21*(4), 347–360.

Hambleton, R. K., Pitoniak, M. J., & Copella, J. M. (2012). Essential steps in setting performance standards on educational tests and strategies for assessing the reliability of results. In G. J. Cizek (Ed.), *Setting performance standards: Foundations, methods, and innovations* (pp. 47–76). New York, NY: Routledge.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston, MA: Kluwer-Nijhoff.

Houts, C. R., & Cai, L. (2013). *flexMIRT user's manual version 2: Flexible multilevel multidimensional item analysis and test scoring*. Chapel Hill, NC: Vector Psychometric Group.

Jodoin, M. G., & Gierl, M. J. (2001). Evaluating Type I error and power rate using an effect size measure with the logistic regression procedure for DIF detection. *Applied Measurement in Education, 14*(4), 329–349.

Kansas State Department of Education. (2010a). *Kansas college and career ready standards: English language arts and literacy in history/social studies, science, and technical subjects.* Retrieved from http://community.ksde.org/LinkClick.aspx?fileticket=tzz1aDOC0v8%3D&tabid=5559&mid=13575.

Kansas State Department of Education. (2010b). *Kansas college and career ready standards: Mathematics grades K–12 with Kansas 15%.* Retrieved from http://community.ksde.org/LinkClick.aspx?fileticket=iX3kWvXtgRY%3d&tabid=5276&mid=13067.

Kansas State Department of Education. (2013). *Kansas standards for history, government, and social studies*. Retrieved from http://www.ksde.org/LinkClick.aspx?fileticket=zNGRyc6vESw%3D&tabid=472&portalid=0&mid=1587.

Kansas State Department of Education. (2015a). *Kansas assessment examiner's manual 2015–2016*. Retrieved from http://www.tregoeagles.com/Pages/District/testMaterials/Kansas%20Assessment%20Examiners%20Manual%202015-2016-2.pdf.

Kansas State Department of Education. (2015b). *Tools and accommodations for the Kansas assessment program (KAP)*. 2015–2016. Website updated to 2016–2017 manual.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices* (2nd ed.). New York, NY: Springer-Verlag. Livingston, S. A.

Koretz, D. (2008). *Measuring up: What educational testing really tells us* (pp. 7–14). Cambridge, MA: Harvard University Press.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159–174.

Livingston, S. A., & Lewis, C. (1995). Estimating the consistency and accuracy of classifications based on test scores. Journal of Educational Measurement, 32, 179-197.

Messick, S. (1989). Validity. In R.L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York, NY: Macmillan.

National Research Council of the National Academies. (2012). *A framework for K–12 science education: Practice, crosscutting concepts, and core ideas.* Washington, DC: The National Academies Press. Retrieved from https://www.nap.edu/read/13165/chapter/1.

Next Generation Science Standards [NGSS]. (2013). *The next generation science standards: Executive summary.* Retrieved from NGSS http://www.nextgenscience.org/sites/default/files/Final%20Release%20NGSS%20Front

%20Matter%20-%206.17.13%20Update_0.pdf.

Petersen, N. S., Kolen, M. J., & Hoover, H. D. (1989). Scaling, norming, and equating. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 221–262). New York, NY: American Council on Education/Macmillan.

Pitoniak, M. J., & Morgan, D. L. (2012). Setting and validating cut scores for tests. In C. Secolsky & D. B. Denison (Eds.), *Handbook on measurement, assessment, and evaluation in higher education* (pp. 343–366). New York, NY: Routledge.

Plake, B. S. (1998). Setting performance standards for professional licensure and certification. *Applied Measurement in Education, 11,* 65–80. http://dx.doi.org/10.1207/s15324818ame1101_4

Plake, B. S., & Hambleton, R. K. (2001). The analytical judgment method for setting standards on complex performance assessments. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 283–312). Mahwah, NJ: Erlbaum.

Press, W. H., Flannery B. P., Teukolsky S. A., & Vetterling W. T. (1989). *Numerical recipes*. New York, NY: Cambridge University Press.

Rijmen, F. (2010). Formal relations and an empirical comparison between the bi-factor, the testlet, and a second-order multidimensional IRT model. *Journal of Educational Measurement, 47*(3), 361–372. http://www.jstor.org/stable/20778960

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software, 48*(2), 1–36.

Samejima, F. (1969). *Estimation of latent ability using a response pattern of graded scores* (Psychometric Monograph No. 17). Richmond, VA: Psychometric Society.

School performance accreditation system; curriculum standards; student assessments; school site councils, Kan. Stat. Ann. §72-6479 (2015). Retrieved from http://kslegislature.org/li/b2015_16/statute/072_000_0000_chapter/072_064_0000_article/072_064_0079_section/072_064_0079_k/.

U.S. Department of Education (2015). Peer Review of State Assessment Systems: non-regulatory guidance for states for meeting requirements of the Elementary and Secondary Education Act of 1965, as amended. Retrieved from https://www2.ed.gov/policy/elsec/guid/assessguid15.pdf.

Webb, N. L. (1997). *Criteria for alignment of expectations and assessments in mathematics and science education* (Council of Chief State School Officers [CCSSO] and National Institute for Science Education Research Monograph No. 6). Washington, DC: CCSSO.

Yen, W. (1981). Using simulation results to choose a latent trait model. *Applied Psychological Measurement,* 5, 245–262.

Appendix A: Test Administration and Security Training

Tickets and Security

- New process for "tickets"
- Student uses his/her designated username and password for the entire testing season
- Daily access codes will be posted in the Educator Portal
- Plan for your secure transmission of this information each day
- Details in the Examiner's Manual

ETHICS

Test Item Security

Teacher/test proctor **may not store or save on computers or personal storage devices** any test items; test items may not be shared via email or other file sharing systems or reproduced by any means.

Still happening so please reinforce!

District Test Coordinator Responsibilities

- List of responsibilities included with fact sheet
- Review carefully to be certain all expectations have been met

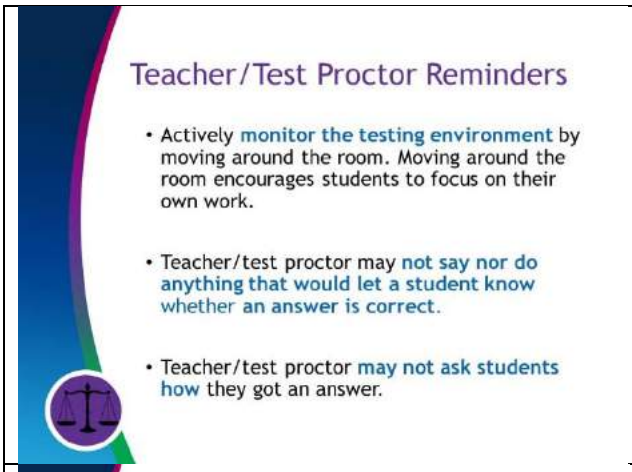
Reporting Item Issues

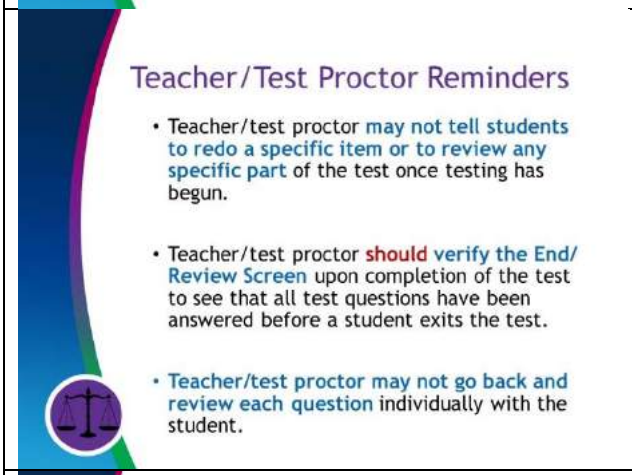
Problems reported to District Test Coordinator who contacts KSDE

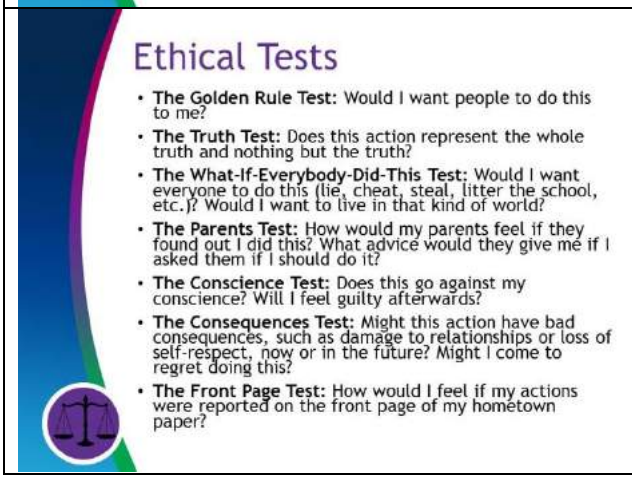
Process outlined on the Test Security Fact Sheet

Teacher/Test Proctor Responsibilities

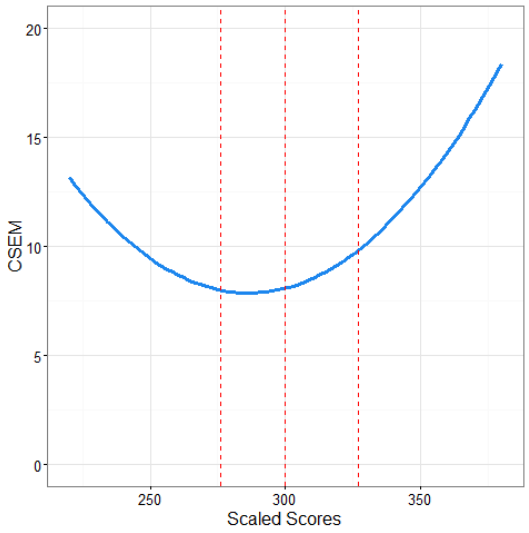
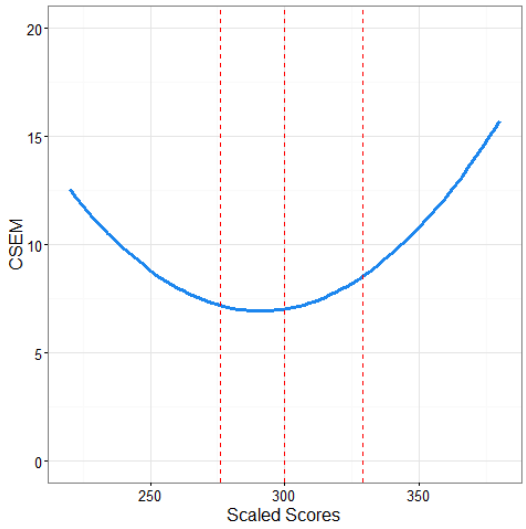
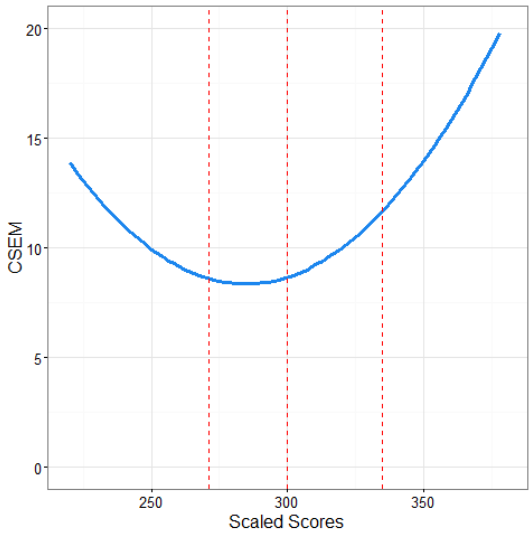
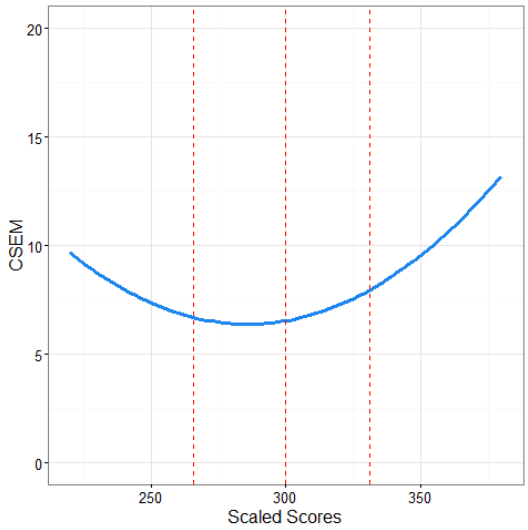
- List of responsibilities included with fact sheet
- Must also review the Examiner's Manual and abide by all directions
- Fact sheet also includes list for Building Test Coordinators

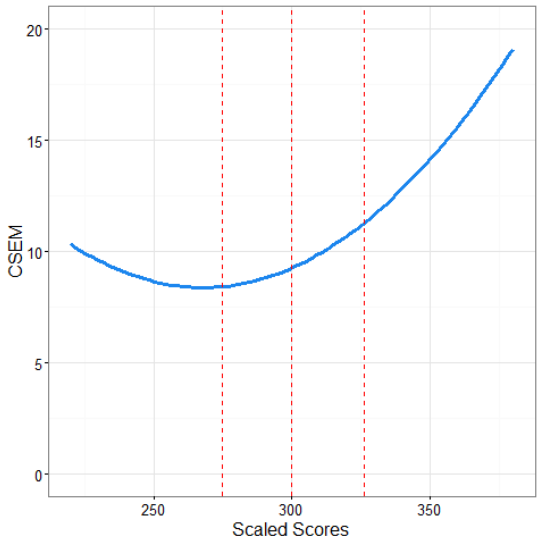
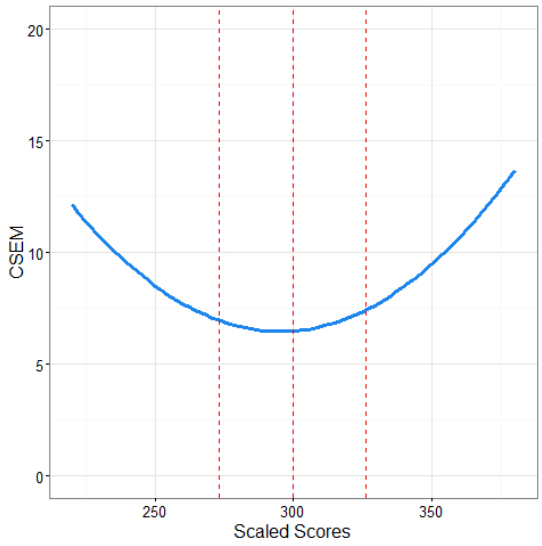
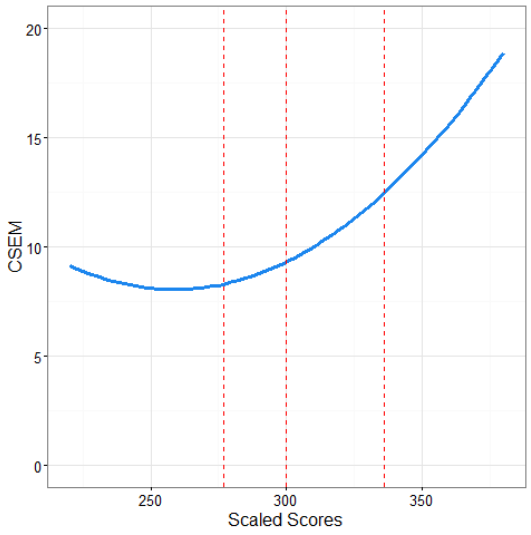
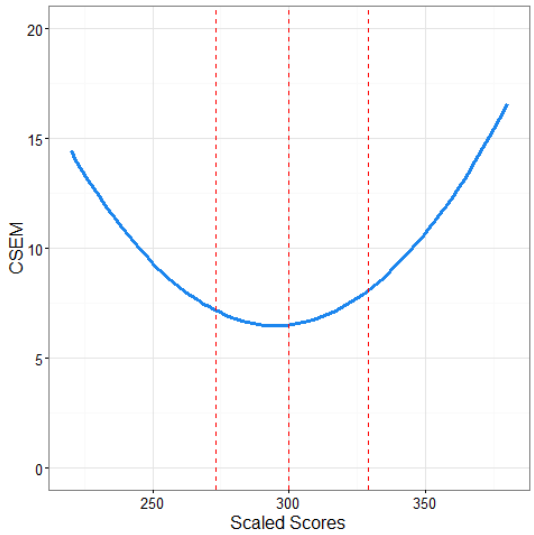

Teacher/Test Proctor Reminders
• Actively monitor the testing environment by moving around the room. Moving around the room encourages students to focus on their own work.
• Teacher/test proctor may not say nor do anything that would let a student know whether an answer is correct.
• Teacher/test proctor may not ask students how they got an answer.

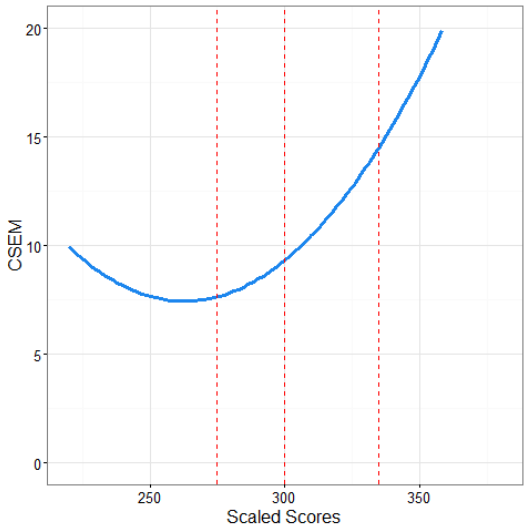
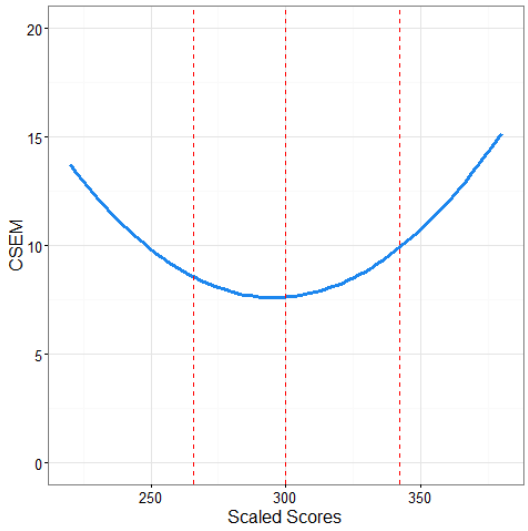
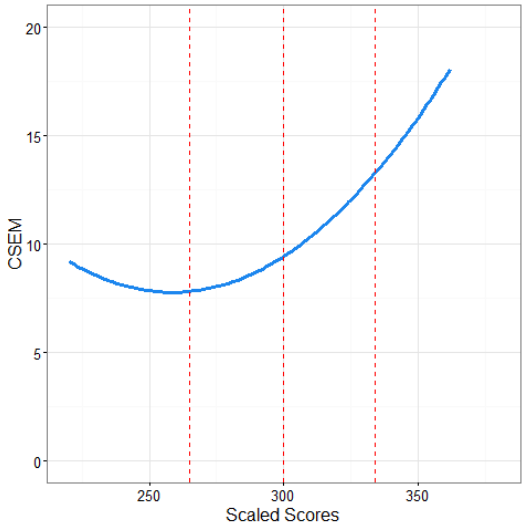
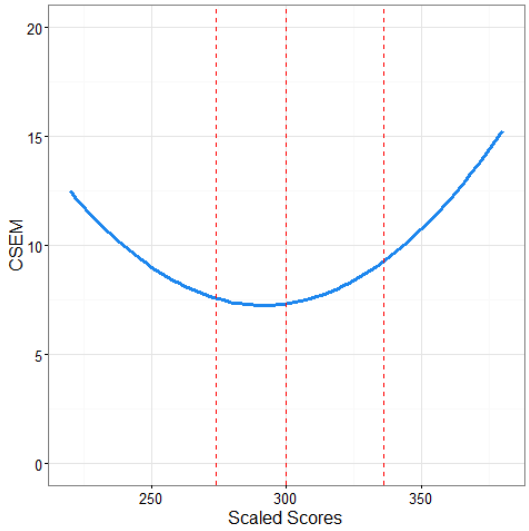

Teacher/Test Proctor Reminders
• Teacher/test proctor may not tell students to redo a specific item or to review any specific part of the test once testing has begun.
• Teacher/test proctor should verify the End/ Review Screen upon completion of the test to see that all test questions have been answered before a student exits the test.
• Teacher/test proctor may not go back and review each question individually with the student.

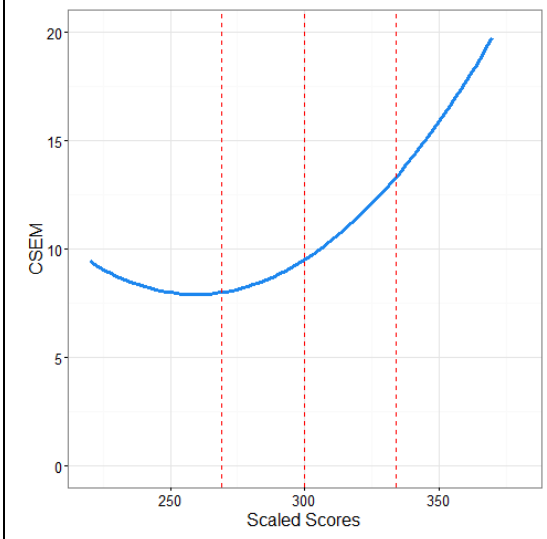
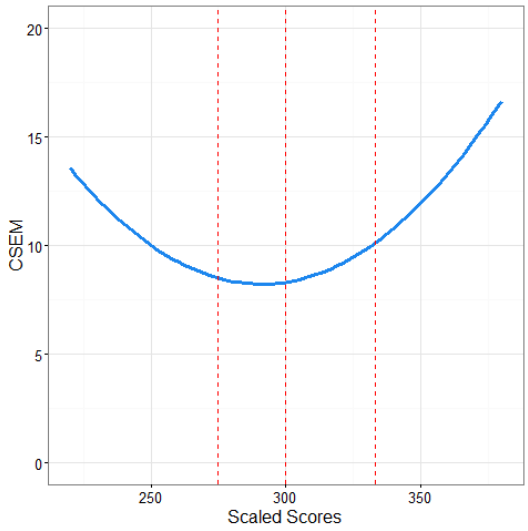

Ethical Tests
• The Golden Rule Test: Would I want people to do this to me?
• The Truth Test: Does this action represent the whole truth and nothing but the truth?
• The What-If-Everybody-Did-This Test: Would I want everyone to do this (lie, cheat, steal, litter the school, etc.)? Would I want to live in that kind of world?
• The Parents Test: How would my parents feel if they found out I did this? What advice would they give me if I asked them if I should do it?
• The Conscience Test: Does this go against my conscience? Will I feel guilty afterwards?
• The Consequences Test: Might this action have bad consequences, such as damage to relationships or loss of self-respect, now or in the future? Might I come to regret doing this?
• The Front Page Test: How would I feel if my actions were reported on the front page of my hometown paper?

Appendix B: Conditional Error of Measurement (CSEM)

Grade	ELA CSEM	Math CSEM
3		
4		

5	 CSEM Scaled Scores 0 5 10 15 20 250 300 350	 CSEM Scaled Scores 0 5 10 15 20 250 300 350
6	 CSEM Scaled Scores 0 5 10 15 20 250 300 350	 CSEM Scaled Scores 0 5 10 15 20 250 300 350

7	 CSEM Scaled Scores	 CSEM Scaled Scores
8	 CSEM Scaled Scores	 CSEM Scaled Scores

10	 CSEM 20 15 10 5 0 250 300 350 Scaled Scores	 CSEM 20 15 10 5 0 250 300 350 Scaled Scores

Appendix C: Subgroup Reliability and Performance

For all tables in Appendix C: NHPI = Native Hawaiian and Pacific Islander; SWD = students with disabilities; ELL = English language learners.

Table C-1

Grade 3 Subgroup Reliability and Performance for ELA

Subgroups	Grade (*N*)	Group (*n*)	%	Reliability	Scaled score	
					M	*SD*
Race	35,838					
Black		2,750	7.7%	.93	281.8	25.0
American Indian		996	2.8%	.93	287.0	23.3
Asian		1,091	3.0%	.91	307.7	29.3
NHPI		85	0.2%	.92	290.5	30.3
White		30,916	86.3%	.92	300.4	27.7
Hispanic	38,282					
Yes		7,510	19.6%	.93	287.6	24.7
No		30,772	80.4%	.92	301.3	28.1
SWD	38,282					
Yes		4,734	12.4%	.93	279.4	25.3
No		33,548	87.6%	.92	301.4	27.3
ELL	38,282					
Yes		4,978	13.0%	.93	283.9	23.6
No		33,304	87.0%	.92	300.9	28.0

Table C-2
Grade 4 Subgroup Reliability and Performance for ELA

Subgroups	Grade (*N*)	Group (*n*)	%	Reliability	Scaled score	
					M	*SD*
Race	34,956					
Black		2,681	7.7%	.92	283.9	26.8
American Indian		1,091	3.1%	.92	290.7	25.5
Asian		1,070	3.1%	.89	311.3	32.2
NHPI		84	0.2%	.92	295.8	28.5
White		30,030	85.9%	.91	305.1	28.9
Hispanic	37,374					
Yes		7,309	19.6%	.92	291.0	27.0
No		30,065	80.4%	.91	305.9	29.1
SWD	37,374					
Yes		4,695	12.6%	.92	280.0	27.2
No		32,679	87.4%	.91	306.3	28.1
ELL	37,374					
Yes		4,785	12.8%	.92	286.5	25.8
No		32,589	87.2%	.91	305.4	29.0

Table C-3
Grade 5 Subgroup Reliability and Performance for ELA

Subgroups	Grade (*N*)	Group (*n*)	%	Reliability	Scaled score	
					M	*SD*
Race	34,406					
Black		2,621	7.6%	.93	282.2	27.2
American Indian		1,273	3.7%	.92	286.2	26.5
Asian		1,040	3.0%	.90	308.8	30.9
NHPI		85	0.3%	.91	295.6	28.0
White		29,387	85.4%	.91	300.0	29.0
Hispanic	36,741					
Yes		7,073	19.3%	.92	286.5	26.8
No		29,668	80.8%	.91	301.0	29.2
SWD	36,741					
Yes		4,468	12.2%	.93	271.9	26.9
No		32,273	87.8%	.91	301.8	27.7
ELL	36,741					
Yes		4,682	12.7%	.93	282.9	26.3
No		32,059	87.3%	.91	300.4	29.1

Table C-4
Grade 6 Subgroup Reliability and Performance for ELA

Subgroups	Grade (*N*)	Group (*n*)	%	Reliability	Scaled score	
					M	*SD*
Race	34,334					
Black		2,517	7.3%	.93	273.8	26.0
American Indian		1,428	4.2%	.93	278.7	25.4
Asian		1,095	3.2%	.90	302.0	32.2
NHPI		99	0.3%	.92	275.0	28.9
White		29,195	85.0%	.91	295.4	27.7
Hispanic	36,661					
Yes		7,003	19.1%	.92	280.9	26.5
No		29,658	80.9%	.91	295.9	28.1
SWD	36,661					
Yes		4,282	11.7%	.93	266.0	24.8
No		32,379	88.3%	.91	296.6	26.9
ELL	36,661					
Yes		4,544	12.4%	.93	275.6	25.7
No		32,117	87.6%	.91	295.5	27.9

Table C-5
Grade 7 Subgroup Reliability and Performance for ELA

Subgroups	Grade (*N*)	Group (*n*)	%	Reliability	Scaled score	
					M	*SD*
Race	34,390					
Black		2,587	7.5%	.93	275.1	25.7
American Indian		1,491	4.3%	.92	280.4	24.8
Asian		1,057	3.1%	.87	302.4	29.7
NHPI		90	0.3%	.92	284.9	26.8
White		29,165	84.8%	.90	294.8	27.3
Hispanic	36,524					
Yes		6,786	18.6%	.92	281.5	26.2
No		29,738	81.4%	.90	295.2	27.6
SWD	36,524					
Yes		4,136	11.3%	.93	265.1	23.3
No		32,388	88.7%	.90	296.2	26.4
ELL	36,524					
Yes		4,266	11.7%	.93	276.1	24.7
No		32,258	88.3%	.90	294.9	27.5

Table C-6
Grade 8 Subgroup Reliability and Performance for ELA

Subgroups	Grade (*N*)	Group (*n*)	%	Reliability	Scaled score *M*	Scaled score *SD*
Race	33,997					
Black		2,565	7.5%	.93	269.9	25.6
American Indian		1,488	4.4%	.93	276.0	24.6
Asian		1,031	3.0%	.89	296.3	31.9
NHPI		84	0.3%	.93	277.1	23.3
White		28,829	84.8%	.91	288.6	28.3
Hispanic	36,133					
Yes		6,573	18.2%	.93	276.0	25.6
No		29,560	81.8%	.91	289.0	28.6
SWD	36,133					
Yes		4,113	11.4%	.94	256.9	22.7
No		32,020	88.6%	.91	290.5	26.9
ELL	36,133					
Yes		3,996	11.1%	.93	270.4	23.6
No		32,137	88.9%	.91	288.7	28.4

Table C-7
Grade 10 Subgroup Reliability and Performance for ELA

Subgroups	Grade (*N*)	Group (*n*)	%	Reliability	Scaled score *M*	Scaled score *SD*
Race	33,728					
Black		2,352	7.0%	.93	266.8	27.0
American Indian		1,491	4.4%	.93	272.2	26.4
Asian		1,080	3.2%	.91	288.6	33.4
NHPI		94	0.3%	.92	279.7	33.1
White		28,711	85.1%	.92	287.6	29.2
Hispanic	35,645					
Yes		6,274	17.6%	.93	272.3	27.3
No		29,371	82.4%	.92	288.0	29.4
SWD	35,645					
Yes		3,619	10.2%	.94	256.1	23.8
No		32,026	89.9%	.92	288.6	28.4
ELL	35,645					
Yes		3,437	9.6%	.94	263.1	24.5
No		32,208	90.4%	.92	287.6	29.2

Table C-8
Grade 3 Subgroup Reliability and Performance for Mathematics

Subgroups	Grade (*N*)	Group (*n*)	%	Reliability	Scaled score	
					M	*SD*
Race	35,822	2,750	7.7%			
Black		995	2.8%	.95	286.7	23.5
American Indian		1,108	3.1%	.94	292.2	23.1
Asian		86	0.2%	.92	317.6	31.5
NHPI		30,883	86.2%	.94	293.7	25.4
White				.93	306.1	27.5
Hispanic	38,263	7,520	19.7%			
Yes		30,743	80.4%	.94	293.1	23.9
No				.93	307.1	28.0
SWD	38,263	4,726	12.4%			
Yes		33,537	87.7%	.94	285.4	26.9
No				.93	307.0	26.9
ELL	38,263	5,032	13.2%			
Yes		33,231	86.9%	.94	290.8	24.0
No		2,750	7.7%	.93	306.4	27.8

Table C-9
Grade 4 Subgroup Reliability and Performance for Mathematics

Subgroups	Grade (*N*)	Group (*n*)	%	Reliability	Scaled score	
					M	*SD*
Race	34,957					
Black		2,678	7.7%	.96	273.8	22.9
American Indian		1,095	3.1%	.96	282.2	23.2
Asian		1,088	3.1%	.94	305.6	33.5
NHPI		84	0.2%	.95	285.5	27.0
White		30,012	85.9%	.95	295.3	27.8
Hispanic	37,373					
Yes		7,338	19.6%	.96	281.6	24.2
No		30,035	80.4%	.95	296.1	28.4
SWD	37,373					
Yes		4,687	12.5%	.95	272.9	25.3
No		32,686	87.5%	.95	296.2	27.4
ELL	37,373					
Yes		4,862	13.0%	.96	279.0	23.8
No		32,511	87.0%	.95	295.4	28.2

Table C-10
Grade 5 Subgroup Reliability and Performance for Mathematics

Subgroups	Grade (*N*)	Group (*n*)	%	Reliability	Scaled score	
					M	*SD*
Race	34,416					
Black		2,615	7.6%	.95	276.1	21.5
American Indian		1,283	3.7%	.95	281.1	22.5
Asian		1,052	3.1%	.94	309.5	31.9
NHPI		85	0.3%	.95	287.8	25.1
White		29,381	85.4%	.95	293.9	27.2
Hispanic	36,742					
Yes		7,117	19.4%	.95	281.0	22.9
No		29,625	80.6%	.95	294.9	27.7
SWD	36,742					
Yes		4,449	12.1%	.94	271.2	22.9
No		32,293	87.9%	.95	295.1	26.7
ELL	36,742					
Yes		4,755	12.9%	.95	279.4	22.8
No		31,987	87.1%	.95	294.1	27.5

Table C-11
Grade 6 Subgroup Reliability and Performance for Mathematics

Subgroups	Grade (*N*)	Group (*n*)	%	Reliability	Scaled score	
					M	*SD*
Race	34,272					
Black		2,520	7.4%	.94	274.8	20.8
American Indian		1,427	4.2%	.94	279.7	20.7
Asian		1,099	3.2%	.92	308.5	34.8
NHPI		98	0.3%	.94	278.9	21.4
White		29,128	85.0%	.94	294.1	26.9
Hispanic	36,591					
Yes		6,994	19.1%	.94	281.1	22.1
No		29,597	80.9%	.94	294.8	27.7
SWD	36,591					
Yes		4,271	11.7%	.94	269.2	19.4
No		32,320	88.3%	.94	295.2	26.6
ELL	36,591					
Yes		4,574	12.5%	.94	278.5	21.5
No		32,017	87.5%	.94	294.1	27.4

Table C-12
Grade 7 Subgroup Reliability and Performance for Mathematics

Subgroups	Grade (*N*)	Group (*n*)	%	Reliability	Scaled score	
					M	*SD*
Race	34,385					
Black		2,578	7.5%	.93	271.5	22.0
American Indian		1,488	4.3%	.93	276.2	21.7
Asian		1,075	3.1%	.92	307.8	35.8
NHPI		88	0.3%	.93	283.0	28.5
White		29,156	84.8%	.93	291.4	28.0
Hispanic	36,515					
Yes		6,796	18.6%	.93	277.7	23.8
No		29,719	81.4%	.93	292.1	28.7
SWD	36,515					
Yes		4,130	11.3%	.92	264.7	20.9
No		32,385	88.7%	.93	292.5	27.7
ELL	36,515					
Yes		4,325	11.8%	.93	274.7	22.6
No		32,190	88.2%	.93	291.3	28.6

Table C-13
Grade 8 Subgroup Reliability and Performance for Mathematics

Subgroups	Grade (*N*)	Group (*n*)	%	Reliability	Scaled score	
					M	*SD*
Race	33,972					
Black		2,563	7.5%	.94	268.0	21.5
American Indian		1,495	4.4%	.94	275.3	23.0
Asian		1,038	3.1%	.92	305.5	38.0
NHPI		86	0.3%	.94	274.3	22.6
White		28,790	84.8%	.94	287.2	28.1
Hispanic	36,103					
Yes		6,584	18.2%	.94	275.1	23.3
No		29,519	81.8%	.94	287.9	28.9
SWD	36,103					
Yes		4,104	11.4%	.93	261.4	19.7
No		31,999	88.6%	.94	288.6	27.9
ELL	36,103					
Yes		4,058	11.2%	.94	272.1	22.1
No		32,045	88.8%	.94	287.2	28.7

Table C-14
Grade 10 Subgroup Reliability and Performance for Mathematics

Subgroups	Grade (*N*)	Group (*n*)	%	Reliability	Scaled score	
					M	*SD*
Race	33,972					
Black		2,350	7.0%	.91	269.6	22.4
American Indian		1,489	4.4%	.92	273.4	21.4
Asian		1,090	3.2%	.90	300.5	37.3
NHPI		95	0.3%	.91	278.3	31.4
White		28,678	85.1%	.92	286.5	28.0
Hispanic	35,619					
Yes		6,282	17.6%	.92	273.6	23.0
No		29,337	82.4%	.92	287.4	28.7
SWD	35,619					
Yes		3,613	10.1%	.91	263.3	17.8
No		32,006	89.9%	.92	287.4	28.1
ELL	35,619					
Yes		3,475	9.8%	.92	269.3	21.0
No		32,144	90.2%	.92	286.7	28.4

Table C-15
Grade 6 Subgroup Reliability and Performance for HGSS

Subgroups	Grade (*N*)	Group (*n*)	%	Reliability	Scaled score	
					M	*SD*
Race	34,159					
Black		2,495	7.3%	.57	282.3	23.0
American Indian		1,425	4.2%	.56	286.6	22.9
Asian		1,093	3.2%	.61	305.3	25.5
NHPI		99	0.3%	.58	290.4	23.7
White		29,047	85.0%	.58	298.8	24.0
Hispanic	36,479					
Yes		6,960	19.1%	.57	288.6	23.5
No		29,519	80.9%	.58	299.1	24.3
SWD	36,479					
Yes		4,260	11.7%	.55	277.8	21.6
No		32,219	88.3%	.57	299.6	23.7
ELL	36,479					
Yes		4,530	12.4%	.56	285.2	23.0
No		31,949	87.6%	.58	298.8	24.2

Table C-16
Grade 8 Subgroup Reliability and Performance for HGSS

Subgroups	Grade (*N*)	Group (*n*)	%	Reliability	Scaled score	
					M	*SD*
Race	33,760					
Black		2,521	7.5%	.48	280.8	21.4
American Indian		1,484	4.4%	.48	284.4	21.4
Asian		1,027	3.0%	.61	301.4	26.0
NHPI		82	0.2%	.53	284.6	23.3
White		28,646	84.9%	.54	294.7	24.0
Hispanic	35,885					
Yes		6,488	18.1%	.51	284.9	22.3
No		29,397	81.9%	.55	295.1	24.2
SWD	35,885					
Yes		4,072	11.3%	.43	275.7	19.8
No		31,813	88.7%	.54	295.5	23.8
ELL	35,885					
Yes		3,970	11.1%	.46	281.1	20.8
No		31,915	88.9%	.55	294.8	24.1

Table C-17
Grade 11 Subgroup Reliability and Performance for HGSS

Subgroups	Grade (*N*)	Group (*n*)	%	Reliability	Scaled score	
					M	*SD*
Race	31,473					
Black		2,160	6.9%	.37	266.7	21.8
American Indian		1,334	4.2%	.32	271.4	22.0
Asian		948	3.0%	.43	284.4	26.9
NHPI		88	0.3%	.36	276.2	24.3
White		26,943	85.6%	.35	282.7	24.3
Hispanic	33,326					
Yes		5,333	16.0%	.35	272.2	22.5
No		27,993	84.0%	.36	282.5	24.5
SWD	33,326					
Yes		3,125	9.4%	.33	260.5	19.2
No		30,201	90.6%	.34	283.0	24.0
ELL	33,326					
Yes		2,597	7.8%	.31	266.2	20.0
No		30,729	92.2%	.36	282.1	24.5

Appendix D: Path Reliability

Table D-1
Grade 3 Path Reliability for ELA

Path	Stage 1	Stage 2	Stage 3	*N*	%	Reliability
				Total 38,208		
1	Medium	Easy	Very easy	11,220	29.4%	.94
2	Medium	Easy	Medium easy	1,268	3.3%	.94
3	Medium	Easy	Medium hard	0		
4	Medium	Medium	Very easy	989	2.6%	.94
5	Medium	Medium	Medium easy	16,430	43.0%	.93
6	Medium	Medium	Medium hard	6,442	16.9%	.89
7	Medium	Medium	Very hard	1,631	4.3%	.83
8	Medium	Hard	Medium easy	0		
9	Medium	Hard	Medium hard	9	0.0%	.86
10	Medium	Hard	Very hard	219	0.6%	.79

Table D-2
Grade 4 Path Reliability for ELA

Path	Stage 1	Stage 2	Stage 3	*N*	%	Reliability
				Total 37,258		
1	Medium	Easy	Very easy	3,194	8.6%	.92
2	Medium	Easy	Medium easy	3,765	10.1%	.94
3	Medium	Easy	Medium hard	0		
4	Medium	Medium	Very easy	53	0.1%	.93
5	Medium	Medium	Medium easy	21,891	58.8%	.93
6	Medium	Medium	Medium hard	3,413	9.2%	.88
7	Medium	Medium	Very hard	0		
8	Medium	Hard	Medium easy	147	0.4%	.90
9	Medium	Hard	Medium hard	4,406	11.8%	.83
10	Medium	Hard	Very hard	389	1.0%	.71

Table D-3
Grade 5 Path Reliability for ELA

Path	Stage 1	Stage 2	Stage 3	*N*	%	Reliability
				Total 36,648		
1	Medium	Easy	Very easy	3,150	8.6%	.93
2	Medium	Easy	Medium easy	13,725	37.5%	.94
3	Medium	Easy	Medium hard	11,099	30.3%	.91
4	Medium	Medium	Very easy	0		
5	Medium	Medium	Medium easy	3	0.0%	.92
6	Medium	Medium	Medium hard	186	0.5%	.91
7	Medium	Medium	Very hard	3,696	10.1%	.88
8	Medium	Hard	Medium easy	0		
9	Medium	Hard	Medium hard	4	0.0%	.92
10	Medium	Hard	Very hard	4,785	13.1%	.85

Table D-4
Grade 6 Path Reliability for ELA

Path	Stage 1	Stage 2	Stage 3	*N*	%	Reliability
				Total 36,471		
1	Medium	Easy	Very easy	10,967	30.1%	.94
2	Medium	Easy	Medium easy	28	0.1%	.92
3	Medium	Easy	Medium hard	0		
4	Medium	Medium	Very easy	4,942	13.6%	.93
5	Medium	Medium	Medium easy	6,411	17.6%	.92
6	Medium	Medium	Medium hard	6,065	16.6%	.91
7	Medium	Medium	Very hard	0		
8	Medium	Hard	Medium easy	106	0.3%	.91
9	Medium	Hard	Medium hard	7,351	20.2%	.87
10	Medium	Hard	Very hard	601	1.7%	.75

Table D-5
Grade 7 Path Reliability for ELA

Path	Stage 1	Stage 2	Stage 3	*N*	%	Reliability
				Total 36,351		
1	Medium	Easy	Very easy	1,578	4.3%	.90
2	Medium	Easy	Medium easy	25,606	70.4%	.93
3	Medium	Easy	Medium hard	0		
4	Medium	Medium	Very easy	0		
5	Medium	Medium	Medium easy	0		
6	Medium	Medium	Medium hard	0		
7	Medium	Medium	Very hard	0		
8	Medium	Hard	Medium easy	4,978	13.7%	.87
9	Medium	Hard	Medium hard	2,204	6.1%	.81
10	Medium	Hard	Very hard	1,985	5.5%	.71

Table D-6
Grade 8 Path Reliability for ELA

Path	Stage 1	Stage 2	Stage 3	*N*	%	Reliability
				Total 35,900		
1	Medium	Easy	Very easy	4,902	13.7%	.94
2	Medium	Easy	Medium easy	4,138	11.5%	.94
3	Medium	Easy	Medium hard	0		
4	Medium	Medium	Very easy	10	0.0%	.94
5	Medium	Medium	Medium easy	12,429	34.6%	.93
6	Medium	Medium	Medium hard	0		
7	Medium	Medium	Very hard	9,069	25.3%	.90
8	Medium	Hard	Medium easy	1	0.0%	.88
9	Medium	Hard	Medium hard	0		
10	Medium	Hard	Very hard	5,351	14.9%	.84

Table D-7
Grade 10 Path Reliability for ELA

Path	Stage 1	Stage 2	Stage 3	*N*	%	Reliability
				Total 35,133		
1	Medium	Easy	Very easy	3,491	9.9%	.94
2	Medium	Easy	Medium easy	1	0.0%	.93
3	Medium	Easy	Medium hard	0		
4	Medium	Medium	Very easy	12,724	36.2%	.94
5	Medium	Medium	Medium easy	6,237	17.8%	.93
6	Medium	Medium	Medium hard	11,093	31.6%	.89
7	Medium	Medium	Very hard	746	2.1%	.82
8	Medium	Hard	Medium easy	0		
9	Medium	Hard	Medium hard	154	0.4%	.86
10	Medium	Hard	Very hard	687	2.0%	.79

Table D-8
Grade 3 Path Reliability for Mathematics

Path	Stage 1	Stage 2	Stage 3	*N*	%	Reliability
				38,176		
1	Medium	Easy	Very easy	1,935	5.1%	.94
2	Medium	Easy	Medium easy	3,484	9.1%	.95
3	Medium	Easy	Medium hard	0		
4	Medium	Medium	Very easy	1	0.0%	.95
5	Medium	Medium	Medium easy	11,577	30.3%	.95
6	Medium	Medium	Medium hard	12,119	31.8%	.94
7	Medium	Medium	Very hard	41	0.1%	.88
8	Medium	Hard	Medium easy	0		
9	Medium	Hard	Medium hard	4,383	11.5%	.93
10	Medium	Hard	Very hard	4,636	12.1%	.87

Table D-9
Grade 4 Path Reliability for Mathematics

Path	Stage 1	Stage 2	Stage 3	*N*	%	Reliability
				37,243		
1	Medium	Easy	Very easy	16,976	45.6%	.96
2	Medium	Easy	Medium easy	928	2.5%	.96
3	Medium	Easy	Medium hard	2	0.0%	.95
4	Medium	Medium	Very easy	969	2.6%	.96
5	Medium	Medium	Medium easy	5,109	13.7%	.96
6	Medium	Medium	Medium hard	4,459	12.0%	.95
7	Medium	Medium	Very hard	817	2.2%	.95
8	Medium	Hard	Medium easy	13	0.0%	.95
9	Medium	Hard	Medium hard	1,021	2.7%	.94
10	Medium	Hard	Very hard	6,949	18.7%	.93

Table D-10
Grade 5 Path Reliability for Mathematics

Path	Stage 1	Stage 2	Stage 3	*N*	%	Reliability
				36,660		
1	Medium	Easy	Very easy	5,740	15.7%	.94
2	Medium	Easy	Medium easy	0		
3	Medium	Easy	Medium hard	0		
4	Medium	Medium	Very easy	11,938	32.6%	.95
5	Medium	Medium	Medium easy	6,001	16.4%	.96
6	Medium	Medium	Medium hard	4,058	11.1%	.95
7	Medium	Medium	Very hard	565	1.5%	.95
8	Medium	Hard	Medium easy	4	0.0%	.95
9	Medium	Hard	Medium hard	794	2.2%	.95
10	Medium	Hard	Very hard	7,560	20.6%	.93

Table D-11
Grade 6 Path Reliability for Mathematics

Path	Stage 1	Stage 2	Stage 3	*N*	%	Reliability
				36,399		
1	Medium	Easy	Very easy	10,024	27.5%	.94
2	Medium	Easy	Medium easy	915	2.5%	.95
3	Medium	Easy	Medium hard	0		
4	Medium	Medium	Very easy	843	2.3%	.95
5	Medium	Medium	Medium easy	6,223	17.1%	.95
6	Medium	Medium	Medium hard	9,972	27.4%	.95
7	Medium	Medium	Very hard	717	2.0%	.94
8	Medium	Hard	Medium easy	0		
9	Medium	Hard	Medium hard	814	2.2%	.95
10	Medium	Hard	Very hard	6,891	18.9%	.91

Table D-12
Grade 7 Path Reliability for Mathematics

Path	Stage 1	Stage 2	Stage 3	*N*	%	Reliability
				36,264		
1	Medium	Easy	Very easy	10,622	29.3%	.92
2	Medium	Easy	Medium easy	0		
3	Medium	Easy	Medium hard	5	0.0%	.94
4	Medium	Medium	Very easy	7,135	19.7%	.94
5	Medium	Medium	Medium easy	0		
6	Medium	Medium	Medium hard	10,464	28.9%	.94
7	Medium	Medium	Very hard	0		
8	Medium	Hard	Medium easy	0		
9	Medium	Hard	Medium hard	6,528	18.0%	.93
10	Medium	Hard	Very hard	1,510	4.2%	.86

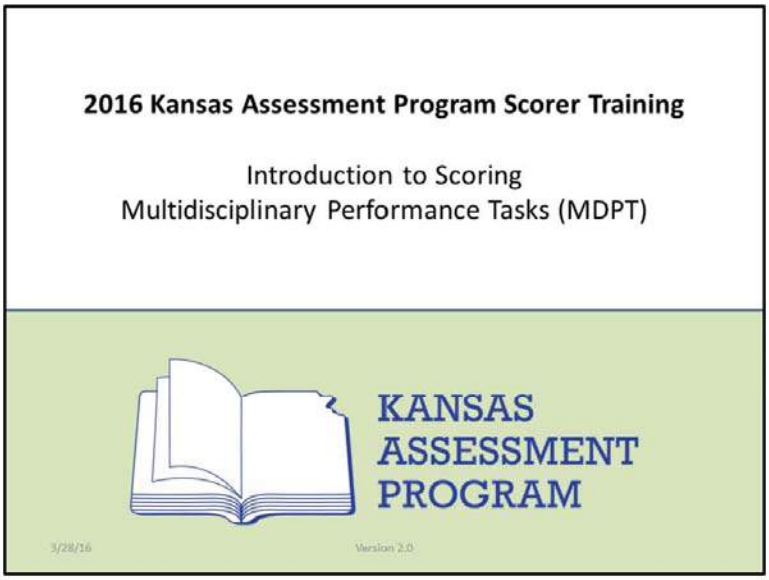
Table D-13
Grade 8 Path Reliability for Mathematics

Path	Stage 1	Stage 2	Stage 3	*N*	%	Reliability
				35,867		
1	Medium	Easy	Very easy	22,711	63.3%	.94
2	Medium	Easy	Medium easy	713	2.0%	.95
3	Medium	Easy	Medium hard	0		
4	Medium	Medium	Very easy	1,088	3.0%	.95
5	Medium	Medium	Medium easy	3,604	10.1%	.95
6	Medium	Medium	Medium hard	1	0.0%	.92
7	Medium	Medium	Very hard	0		
8	Medium	Hard	Medium easy	3,130	8.7%	.94
9	Medium	Hard	Medium hard	2,020	5.6%	.93
10	Medium	Hard	Very hard	2,600	7.3%	.88

Table D-14
Grade 10 Path Reliability for Mathematics

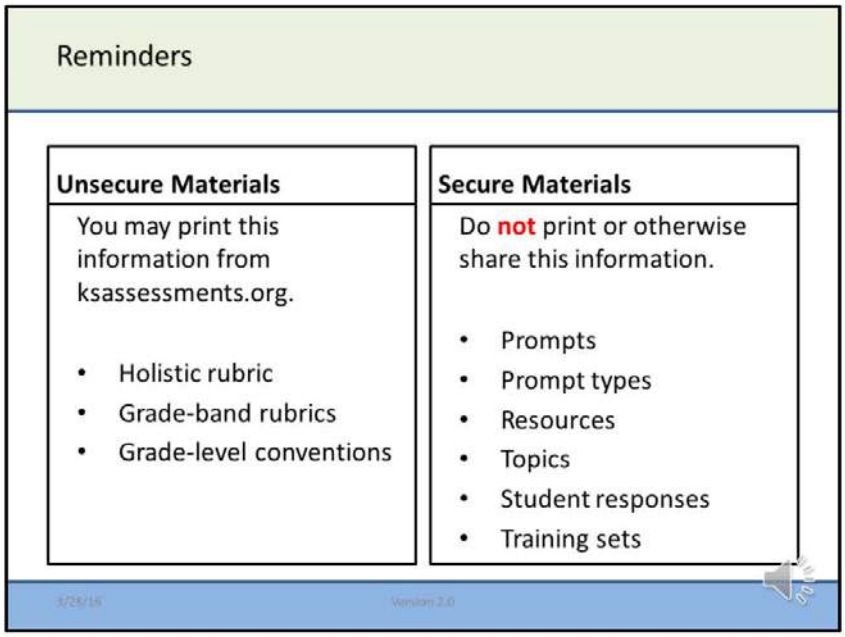
Path	Stage 1	Stage 2	Stage 3	*N*	%	Reliability
				35,124		
1	Medium	Easy	Very easy	119	0.3%	.85
2	Medium	Easy	Medium easy	23,583	67.1%	.92
3	Medium	Easy	Medium hard	0		
4	Medium	Medium	Very easy	0		
5	Medium	Medium	Medium easy	8,234	23.4%	.92
6	Medium	Medium	Medium hard	0		
7	Medium	Medium	Very hard	359	1.0%	.88
8	Medium	Hard	Medium easy	368	1.1%	.90
9	Medium	Hard	Medium hard	0		
10	Medium	Hard	Very hard	2,461	7.0%	.85

Appendix E: Rater Scoring Training



Audio:

None



Audio:

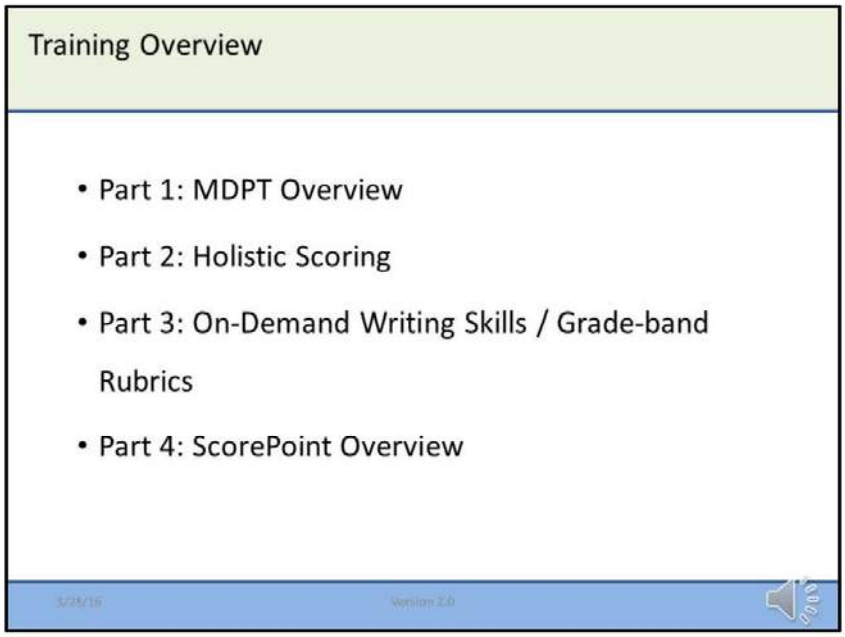
If you have not already printed paper copies of the holistic rubric, the grade-band rubrics for the grades you plan to score, and the grade-level conventions document, we recommend that you do so now.

These documents are available on CETE's Kansas Assessments webpage.

It will be helpful to have these documents in front of you as you complete this training and as you score.

Also, please remember that you will be working with secure materials. It is critical that everyone remember that the MDPTs must be kept secure and that normal test security protocols apply.

For example, please do **not** email any information that describes a particular task in any way, discuss specific tasks or student responses with anyone outside of the scoring environment, allow anyone else to use your username and password, take pictures of tasks or responses with a camera, or do anything else that could detract from the validity of the assessment.



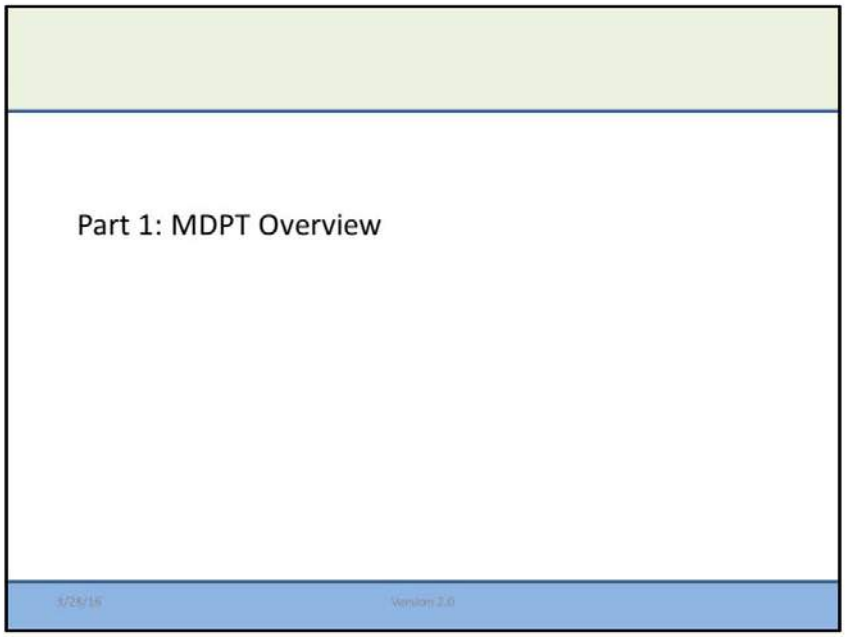
Audio:

This training will prepare you to score the responses to the Multidisciplinary Performance Tasks!

Please note that the training materials and scoring procedures have changed from last year, so it is important to review all four parts of this training **prior** to logging in to ScorePoint.

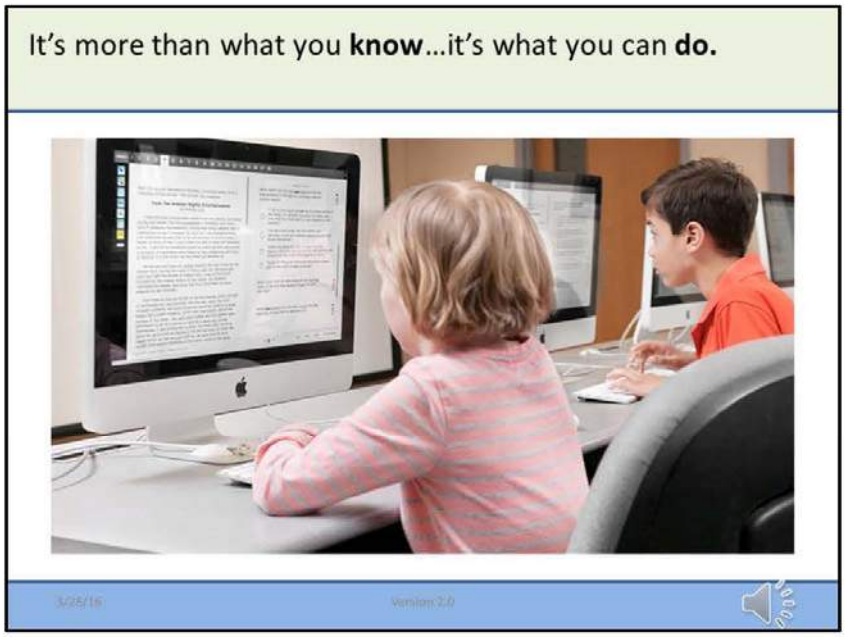
We will cover background information about the MDPT and provide an overview of the holistic rubric that you will use to score.

In addition, this training will review the on-demand writing skills outlined on the grade-band rubrics and prepare you to use ScorePoint to score practice responses and then student responses.



Audio:

None

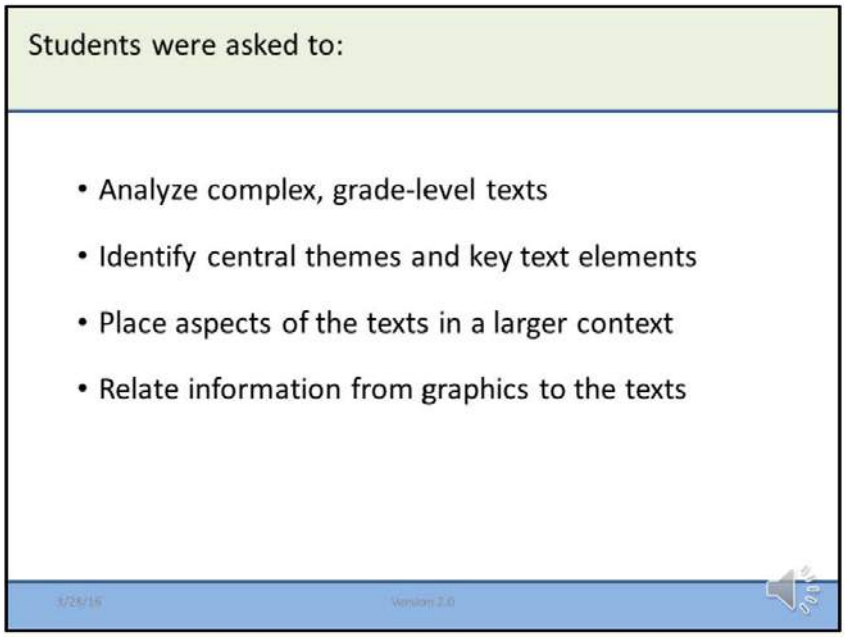


Audio:

In addition to content-based assessments with multiple-choice and technology-enhanced items, two types of performance tasks were administered this year: Multidisciplinary Performance Tasks, or MDPTs, and Mathematics Performance Tasks.

This training will focus solely on scoring the MDPT responses.

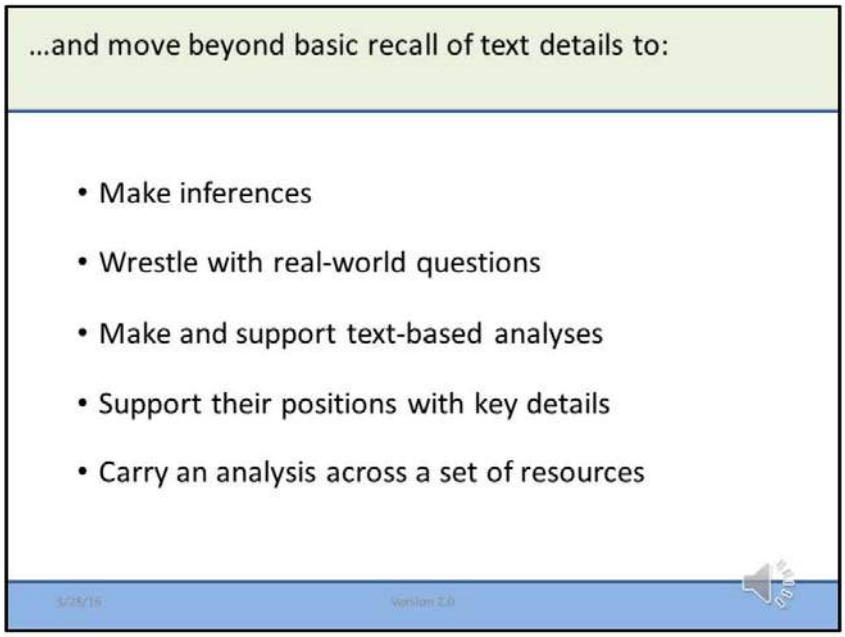
Prior to scoring, it is helpful to understand a little bit about how the MDPTs are structured.



Audio:

The MDPTs asked students to:

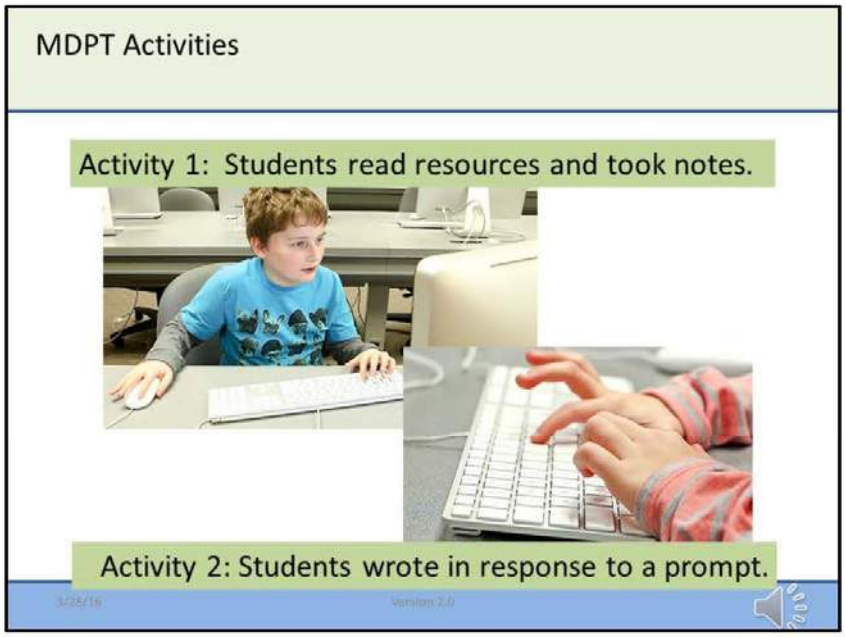
- Analyze complex, grade-level texts
- Identify central themes and key text elements
- Place aspects of the texts in their larger context, and
- Relate information from graphics to the texts...



Audio:

...and move beyond basic recall of text details to:

- Infer how specific portions of text relate to the larger text
- Wrestle with meaningful, real-world questions
- Make and support text-based analyses
- Support text-based analyses with key details, and
- Carry an analysis beyond one text, relating details to an overarching idea from several resources.



Audio:

Because the MDPTs are on-demand writing tasks, students are **not** expected to submit a polished piece of writing that follows a comprehensive, multi-day writing process, which often includes brainstorming, drafting, peer editing, and publishing.

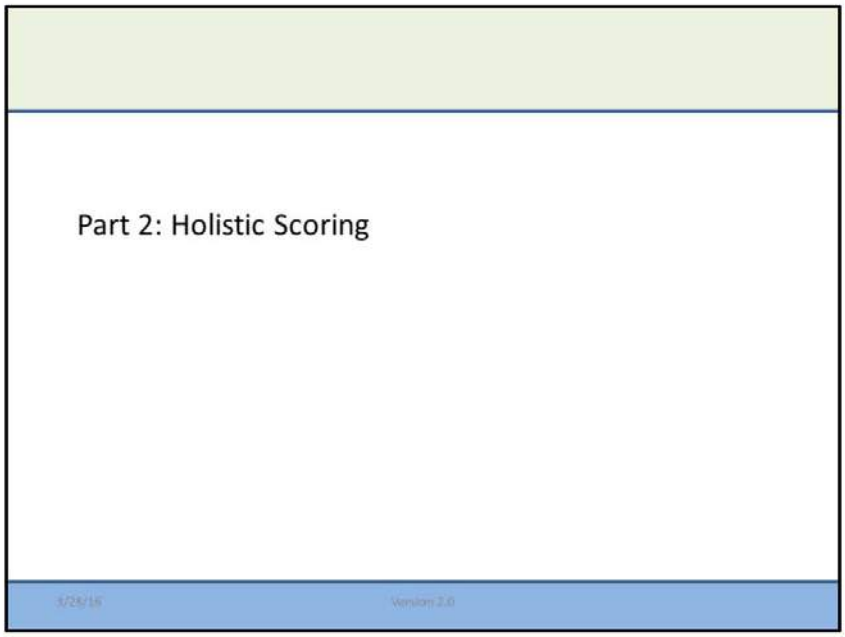
For the summative assessment, each student completed one MDPT. Each MDPT included two activities.

During Activity 1, students had time to read and analyze a set of related resources and take notes.
The resources included texts and graphics such as timelines and charts. The texts may have been literary or informational and may have been related to any topic such as culture, science, or history.

During Activity 2, students wrote in response to a prompt. The prompt may have asked for a narrative, informative/explanatory, opinion, or argumentative response. Students had access to their notes and to the resources while they were writing.

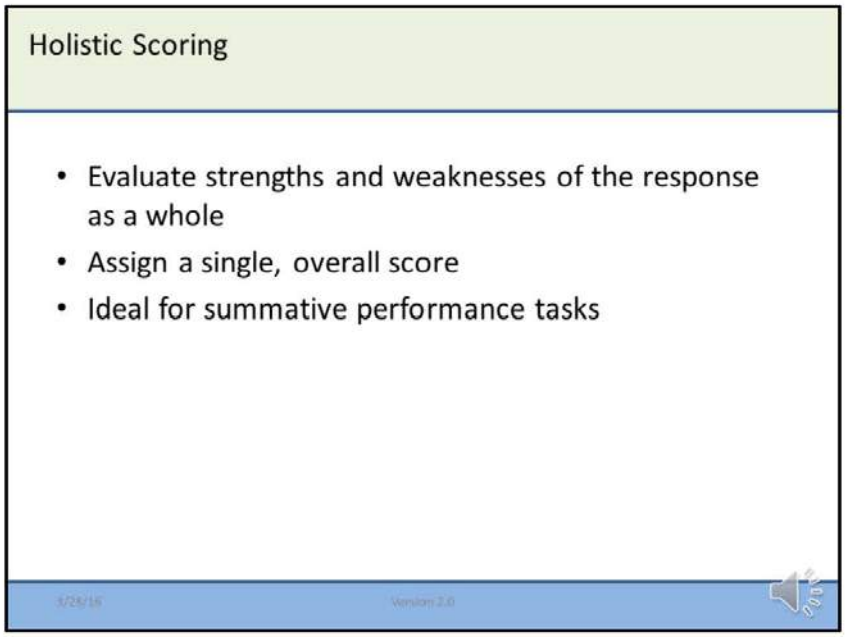
Next we will review holistic scoring and the holistic rubric that you will use to score.

Please continue to Part 2 of this training.



Audio:

None



Audio:

For the field test in 2015, each response received a score for each dimension on the grade-band rubric.

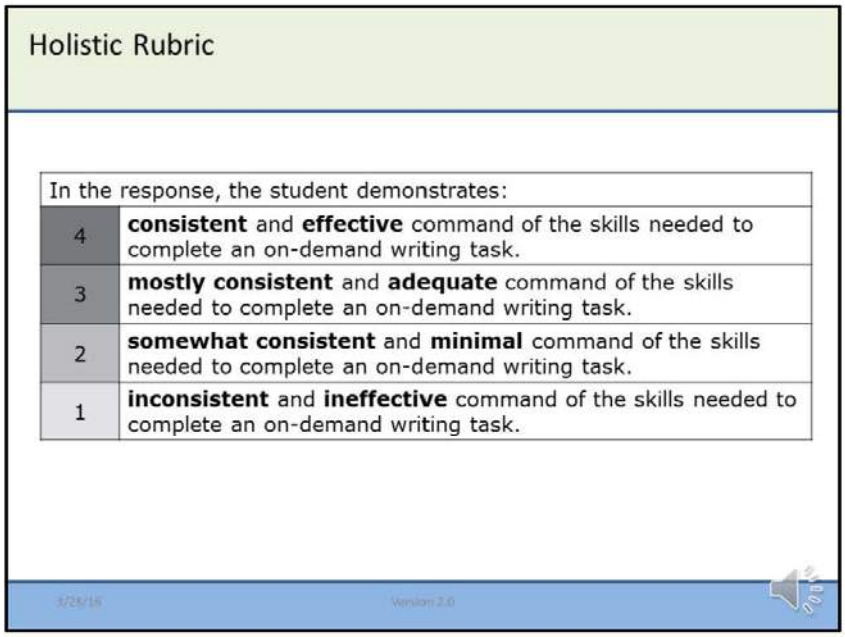
The data showed that there was little distinction among the individual dimension scores.

Stronger responses generally received higher scores across all dimensions, and weaker responses generally received lower scores across all dimensions.

Based on the field test results, KSDE's Technical Advisory Committee recommended providing a single, holistic score for each MDPT response moving forward.

Holistic scoring requires the scorer to evaluate the strengths and weaknesses of a response as a whole and then assign a **single** score based on the scorer's overall judgment of the response.

Holistic scoring is also an efficient scoring method for summative performance tasks as the rubric can be applied consistently by raters.



Audio:

This is the holistic rubric that will be used to score the summative MDPT responses. In the next part of this training, we will review the skills that are needed to complete this on-demand writing task.

Please note that this rubric was created specifically for scoring MDPT responses and may not be applicable to other writing tasks or assignments.

Take a few minutes to review **this** rubric. You may want to make note of the differences among the levels on your rubric.

As you can see, Level 4 responses are the strongest, and Level 1 responses are the weakest; however, Level 4 responses may not be perfect and might not demonstrate strong command of **all** skills.

A thorough understanding of the on-demand writing skills outlined in the grade-band rubrics is necessary in order to provide a holistic score, so now we will review those rubrics.

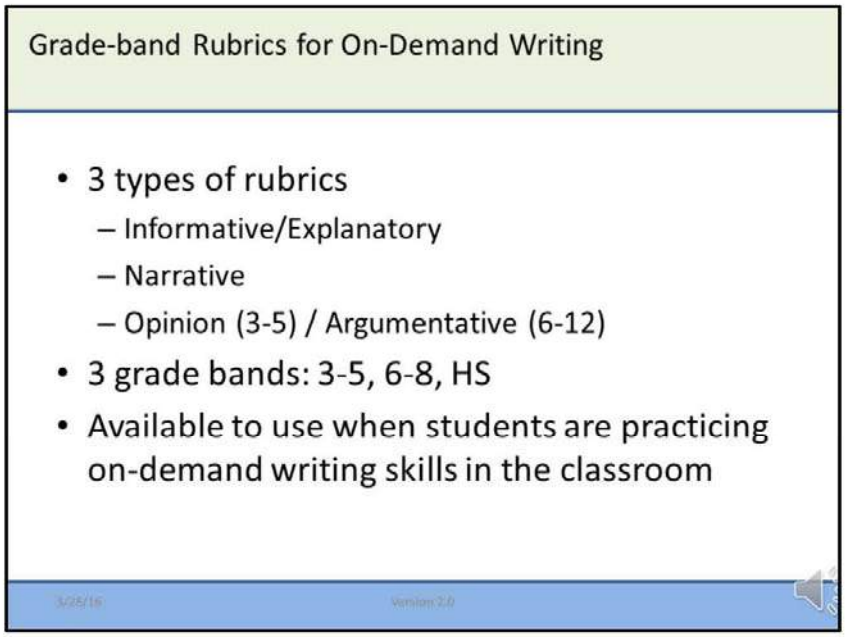
Please continue to Part 3 of this training.

Part 3: On-Demand Writing Skills / Grade-band Rubrics

3/28/16 Version 2.0

Audio:

None



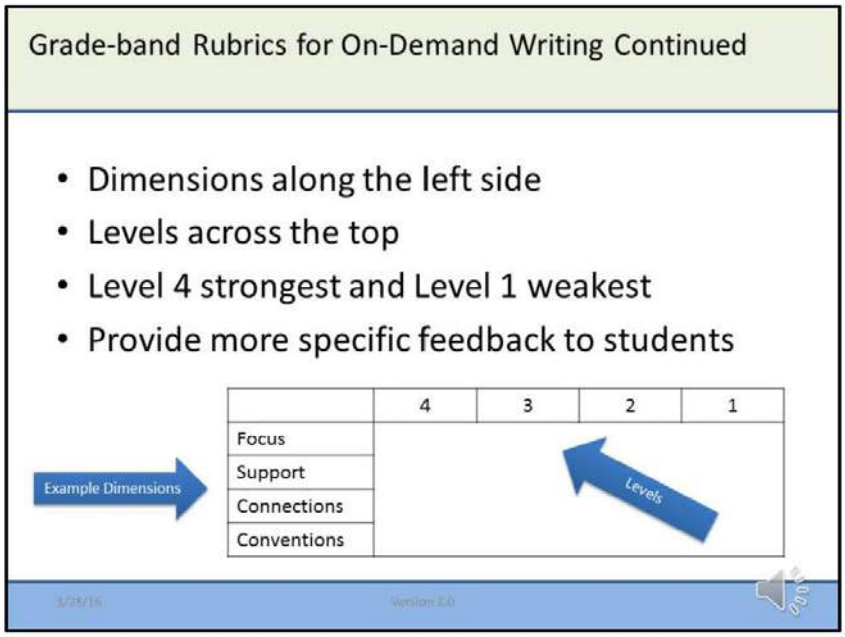
Audio:

The holistic rubric is the same across grade bands and prompt types; however, the grade-band rubrics vary slightly by grade band and prompt type because they are more descriptive.

The differences in the rubrics are derived from the types of writing as well as the progression of skill development defined in the standards.

The grade-band rubrics were developed by KSDE in collaboration with teachers with the MDPTs in mind.

These rubrics are available for teachers to use in the classroom if the rubric matches the students' writing task.



Audio:

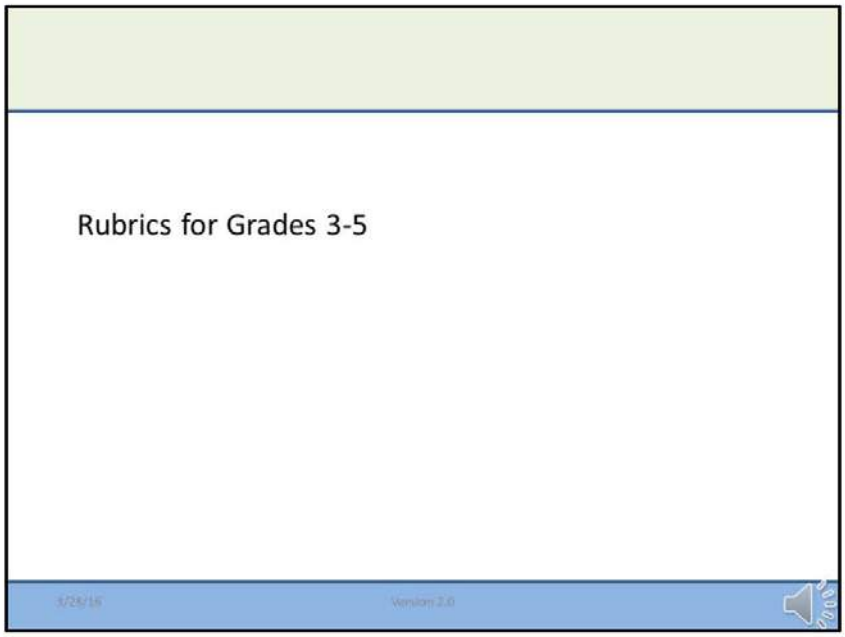
The grade-band rubrics are all set up the same way.

The "dimensions," such as "Focus" and "Conventions" are along the left side.

The levels, from 4 to 1, are across the top.

Level 4 represents stronger command of the skills within a dimension and Level 1 represents weaker command of the skills within a dimension.

Now, please move to the section of this training for the grade band or grade bands for which you will be scoring.



Audio:

In this section of the training, we will identify the on-demand writing skills that are outlined in the rubrics for grades 3-5.

Grades 3-5
Informative/Explanatory Rubric

3-5 Inf./Expl.	Student's Response…			
PL:	4	3	2	1
Focus	☐ Clearly states and maintains a controlling idea that directly addresses the resources and prompt	☐ Adequately states and maintains a controlling idea that mostly addresses the resources and prompt	☐ States a controlling idea somewhat related to the resources and prompt	☐ Does not state a controlling idea related to the resources and prompt
Support	☐ Effectively uses relevant and accurate facts, definitions, and details throughout the work	☐ Adequately uses relevant and accurate facts, definitions, and details in the work	☐ Uses some relevant and accurate facts, definitions and details in the work	☐ Does not use relevant and accurate facts, definitions, or details in the work
Connections	☐ Consistently uses grade-appropriate strategies to clarify relationships between and among ideas	☐ Adequately uses grade-appropriate strategies to clarify relationships between and among ideas	☐ Inconsistently uses grade-appropriate strategies to clarify relationships between and among ideas	☐ Shows little or no attempt to clarify relationships between and among ideas
Conventions	☐ Is readable with most grade-level conventions used correctly and may use them creatively to enhance the message; minor mistakes do not impede the reader's ability to understand the writer's meaning	☐ Is readable with most grade-level conventions used correctly; mistakes do not affect the reader's ability to understand the writer's meaning	☐ Is readable but some errors in grade-level conventions negatively impact the reader's ability to understand the writer's meaning	☐ Is nearly unreadable due to pervasive errors in grade-level conventions

3/28/16 Version 2.0

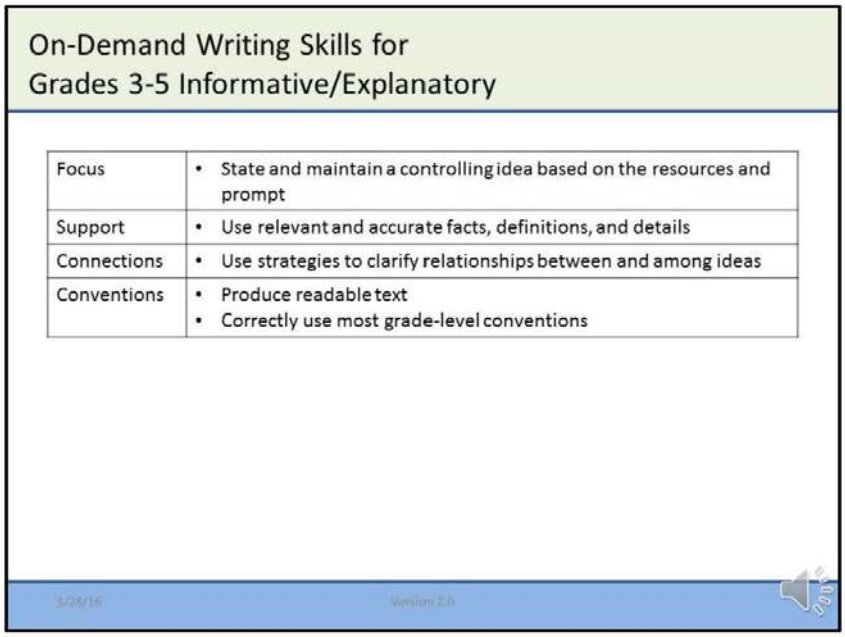
Audio:

Take a few minutes to read through the informative/explanatory rubric.

Make note of the anchor skill or skills listed for each dimension.

For example, being able to "state and maintain a controlling idea based on the resources and prompt" is the expected skill in the "Focus" dimension.

16



Audio:

Compare your notes to the skills we have listed.

For an informative/explanatory on-demand writing task, a student is expected to:

- State and maintain a controlling idea based on the resources and prompt,
- Use relevant and accurate facts, definitions, and details,
- Produce readable text, and
- Correctly use most grade-level conventions.

In all grades, the "Conventions" dimension is the same. For an overview of convention-related skills by grade, please refer to the "Grade-Level Conventions" document.

Grades 3-5
Opinion Rubric

3-5 Opinion	Student's Response...			
PL:	4	3	2	1
Focus/Opinion	☐ States and maintains a clear opinion that directly addresses the resources and prompt	☐ States and mostly maintains a clear opinion related to the resources and prompt	☐ States an opinion that is somewhat clear, and somewhat related to the resources and prompt	☐ Does not state a clear opinion, or stated opinion is unrelated to resources or prompt
Evidence	☐ Uses relevant and accurate details/evidence from one or more resources to support opinion	☐ Uses mostly relevant and accurate details/ evidence from one or more resources to support opinion	☐ Uses some relevant and accurate details/evidence from one or more resources to support opinion	☐ Does not use relevant and accurate details or evidence from resources to support opinion
Connections	☐ Consistently uses grade-appropriate strategies to clarify relationships between and among ideas	☐ Adequately uses grade-appropriate strategies to clarify relationships between and among ideas	☐ Inconsistently uses grade-appropriate strategies to clarify relationships between and among ideas	☐ Shows little or no attempt to clarify relationships between and among ideas
Conventions	☐ Is readable with most grade-level conventions used correctly and may use them creatively to enhance the message; minor mistakes do not impede the reader's ability to understand the writer's meaning	☐ Is readable with most grade-level conventions used correctly; mistakes do not affect the reader's ability to understand the writer's meaning	☐ Is readable but some errors in grade-level conventions negatively impact the reader's ability to understand the writer's meaning	☐ Is nearly unreadable due to pervasive errors in grade-level conventions

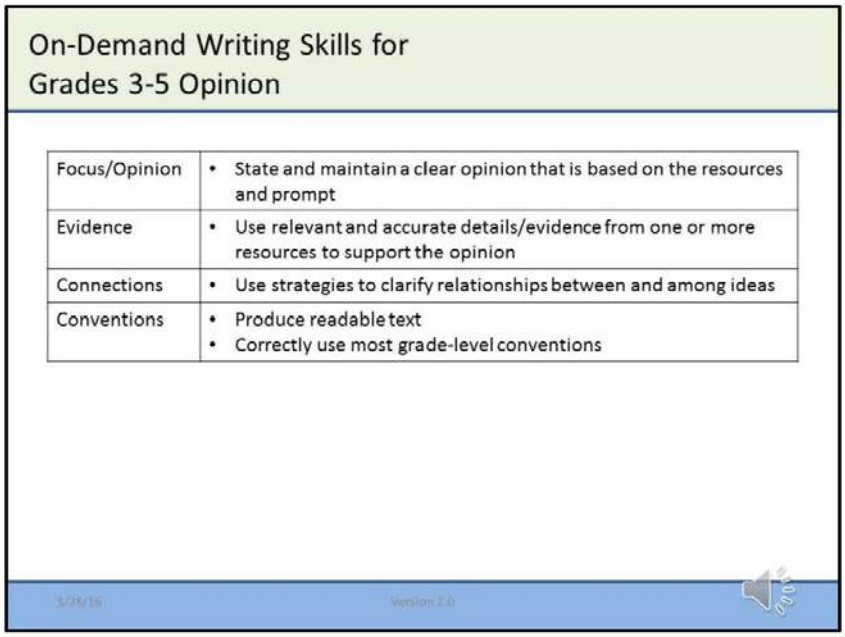
3/28/16 Version 2.0

Audio:

Take a few minutes to read through the opinion rubric.

Make note of the anchor skill or skills listed for each dimension.

18

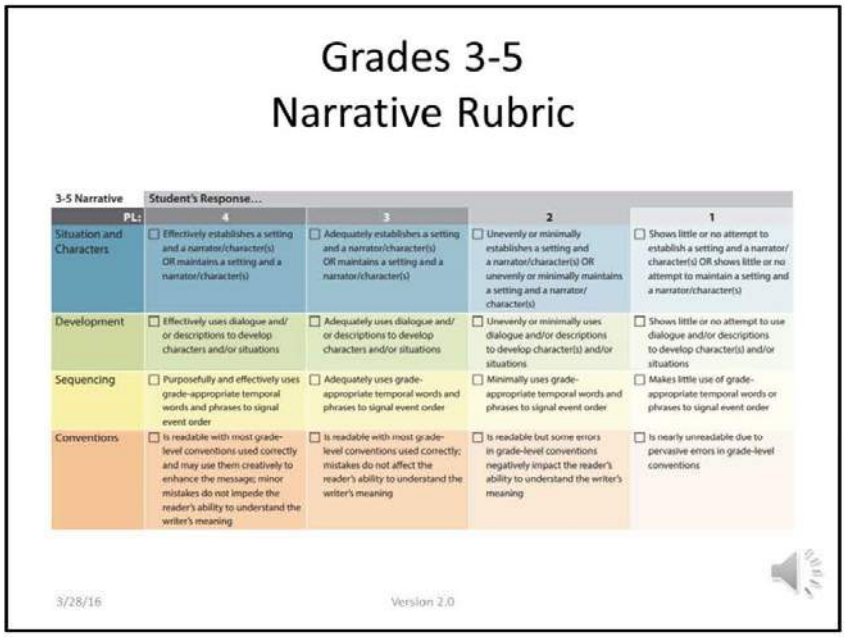


Audio:

Compare your notes to the skills we have listed.

For an opinion on-demand writing task, a student is expected to:

- State and maintain a clear opinion that is based on the resources and prompt,
- Use relevant and accurate details/evidence from one or more resources to support the opinion,
- Use strategies to clarify relationships between and among ideas,
- Produce readable text, and
- Correctly use most grade-level conventions.

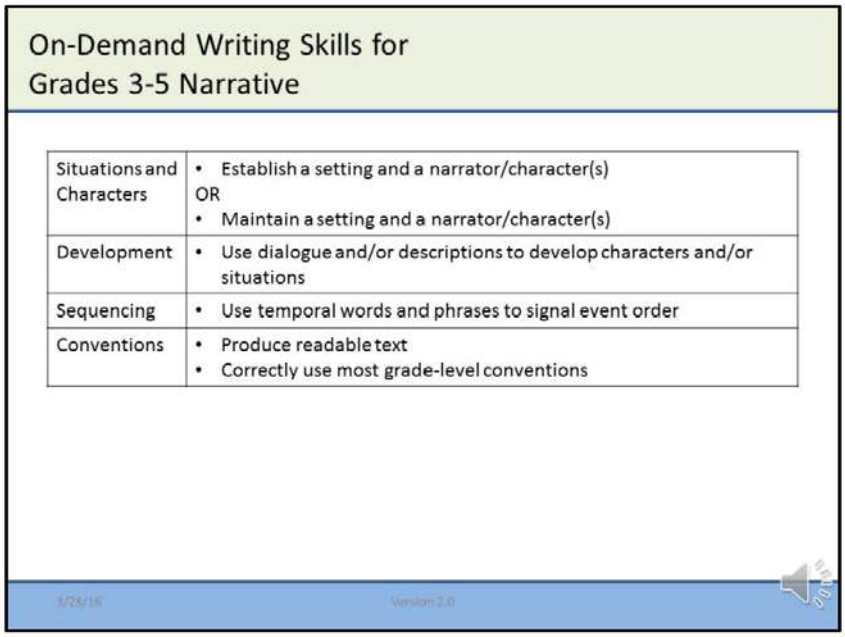


Audio:

Take a few minutes to read through the narrative rubric.

Make note of the anchor skill or skills listed for each dimension.

20

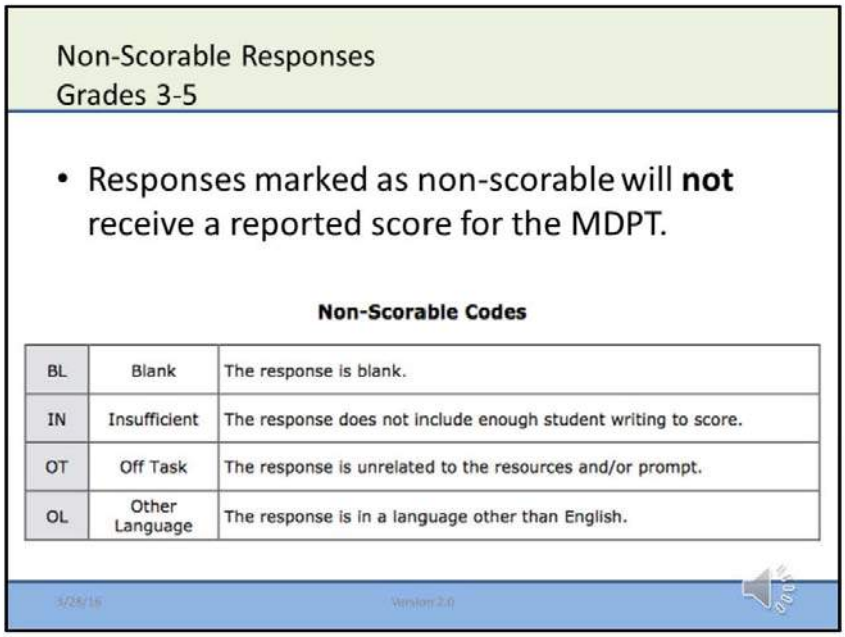


Audio:

Compare your notes to the skills we have listed.

For a narrative on-demand writing task, a student is expected to:

- Establish a setting and a narrator/character(s), OR
- Maintain a setting and a narrator/character(s),
- Use dialogue and/or descriptions to develop characters and/or situations,
- Use temporal words and phrases to signal event order,
- Produce readable text, and
- Correctly use most grade-level conventions.



Audio:

It is possible that you may read a response that is non-scorable.

Please note that students with responses marked as non-scorable will **not** receive a reported score for the MDPT.

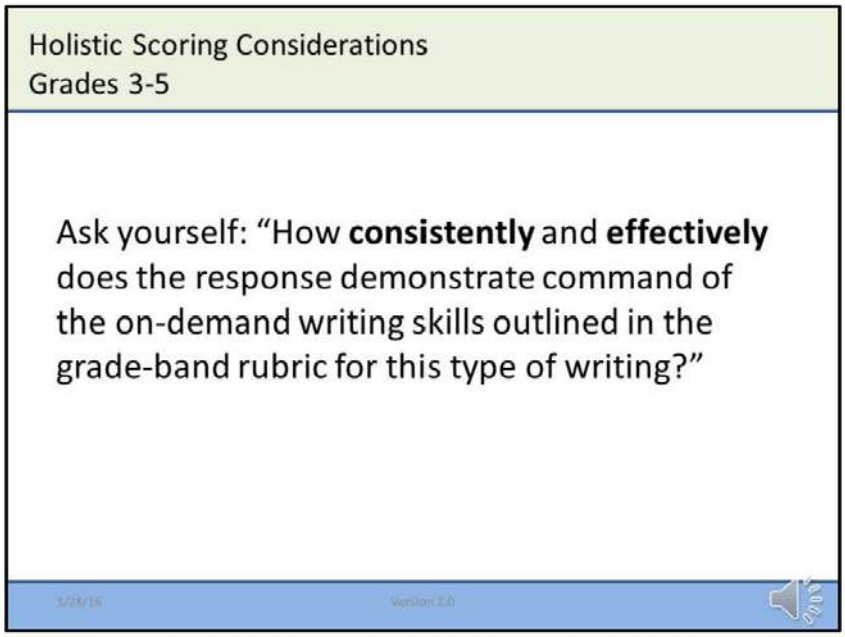
Blank responses and responses written in a language other than English cannot be scored.

A response may be considered insufficient and non-scorable if there are not enough of the student's original words and ideas to score.

A response may be considered off task and non-scorable if the student wrote about something unrelated to the prompt and/or resources.

If you think that a response is non-scorable, we recommend that you visit with a peer within the scoring environment to further analyze the response.

In Part 4, you will learn how to mark non-scorable responses in ScorePoint.



Audio:

Refer to the levels and descriptions in the grade-band rubric as needed; however, calculating a holistic score based on a formula is unnecessary and **not** recommended.

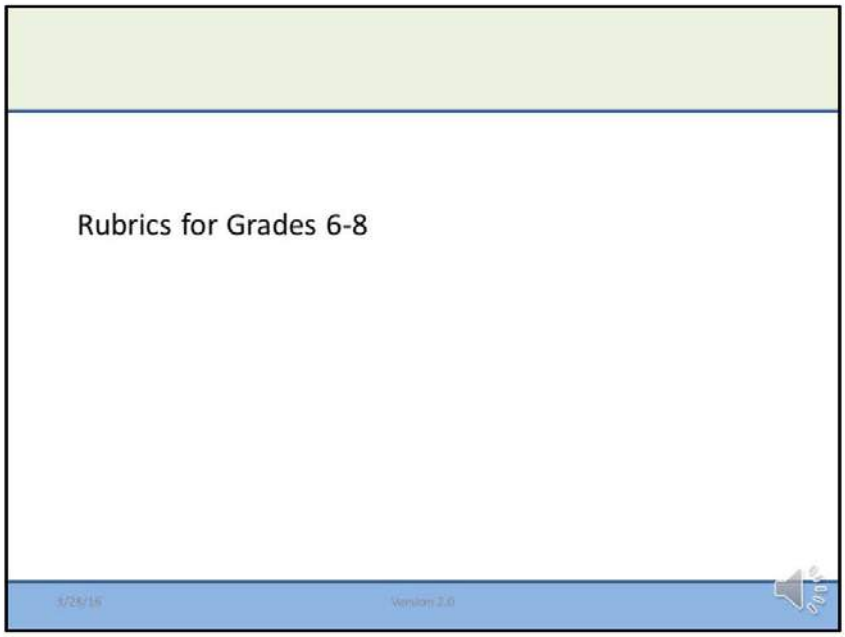
As you read each student response, ask yourself: "How consistently and effectively does the response demonstrate command of the on-demand writing skills outlined in the grade-band rubric for this type of writing?"

Remember to consider the work as a whole as you assign a single score based on your overall judgment of the response.

Prior to scoring student responses, you will score several sample responses and see how experts evaluated them.

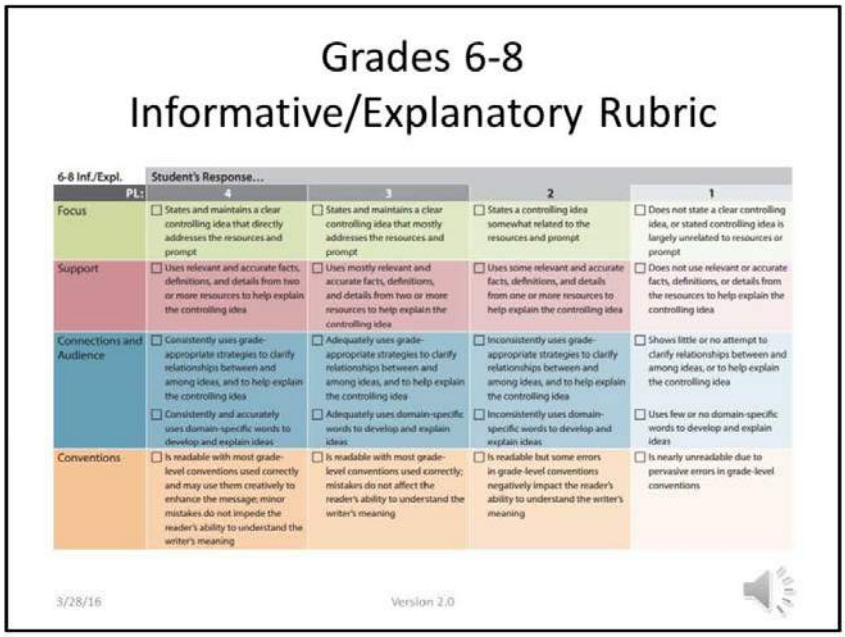
This completes the grade-band training for grades 3-5. If you will also be scoring within another grade band, please complete that section prior to moving on to Part 4 of this training.

Otherwise, please move on to Part 4.



Audio:

In this section of the training, we will identify the on-demand writing skills that are outlined in the rubrics for grades 6-8.



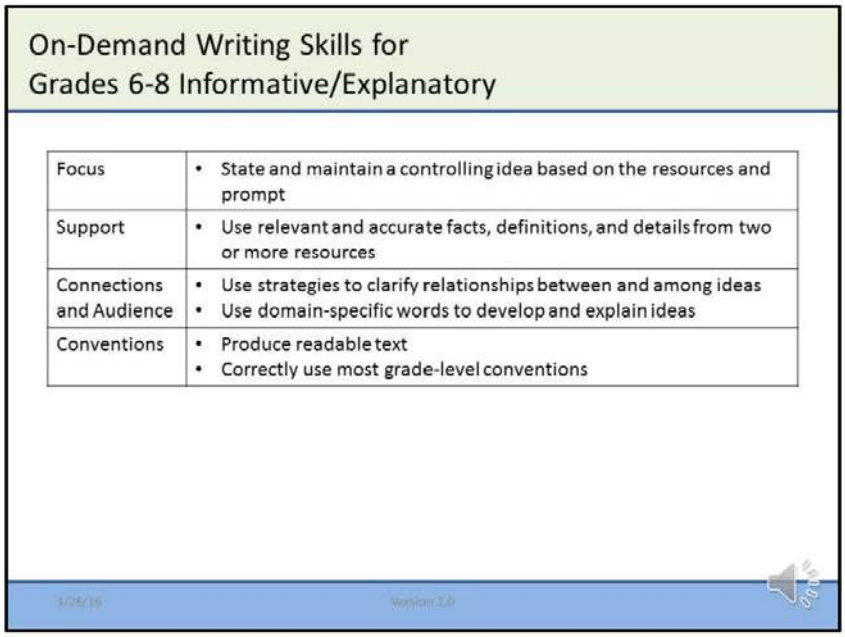
Audio:

Take a few minutes to read through the informative/explanatory rubric.

Make note of the anchor skill or skills listed for each dimension.

For example, being able to "state and maintain a controlling idea based on the resources and prompt" is the expected skill in the "Focus" dimension.

25



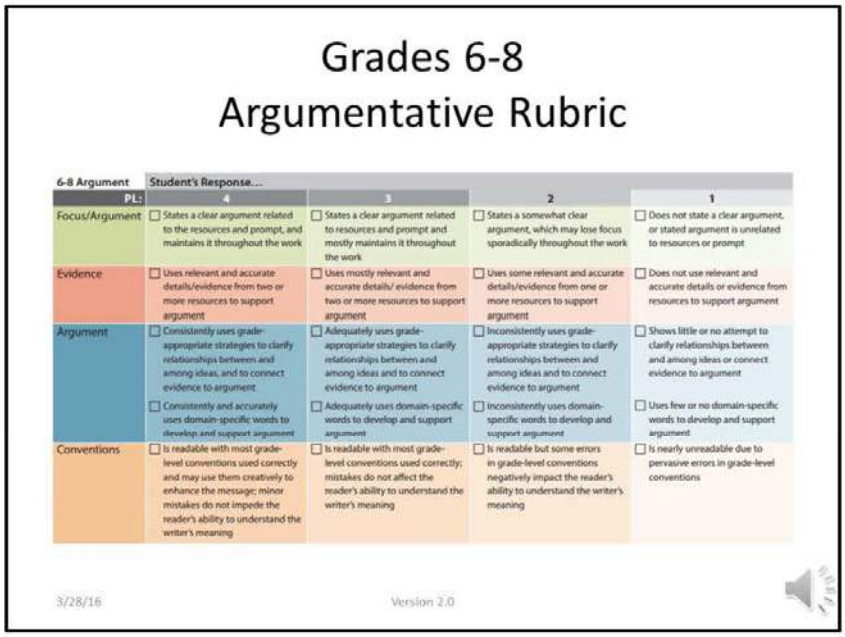
Audio:

Compare your notes to the skills we have listed.

For an informative/explanatory on-demand writing task, a student is expected to:

- State and maintain a controlling idea based on the resources and prompt,
- Use relevant and accurate facts, definitions, and details from two or more resources,
- Use strategies to clarify relationships between and among ideas,
- Use domain-specific words to develop and explain ideas,
- Produce readable text, and
- Correctly use most grade-level conventions.

In all grades, the "Conventions" dimension is the same. For an overview of convention-related skills by grade, please refer to the "Grade-Level Conventions" document.

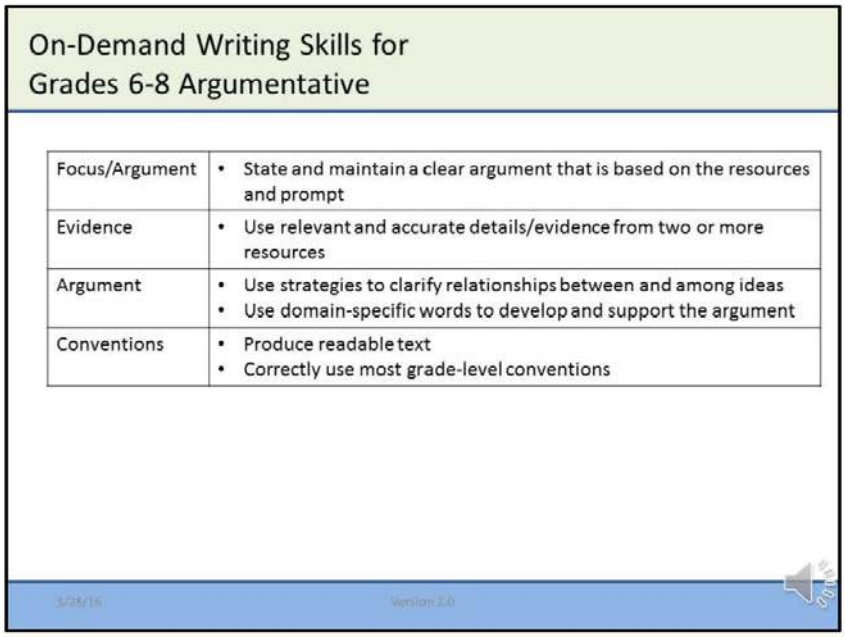


Audio:

Take a few minutes to read through the argumentative rubric.

Make note of the anchor skill or skills listed for each dimension.

27



Audio:

Compare your notes to the skills we have listed.

For an opinion on-demand writing task, a student is expected to:

- State and maintain a clear argument that is based on the resources and prompt,
- Use relevant and accurate details/evidence from two or more resources,
- Use strategies to clarify relationships between and among ideas,
- Use domain-specific words to develop and support the argument,
- Produce readable text, and
- Correctly use most grade-level conventions.

Grades 6-8 Narrative Rubric

6-8 Narrative	Student's Response...			
PL:	4	3	2	1
Storyline and Characters	☐ Effectively establishes a setting and a narrator/character(s) OR maintains a setting and a narrator/character(s)	☐ Adequately establishes a setting and a narrator/character(s) OR maintains a setting and a narrator/character(s)	☐ Unevenly or minimally establishes a setting and a narrator/character(s) OR unevenly or minimally maintains a setting and a narrator/ character(s)	☐ Shows little or no attempt to establish a setting and a narrator/ character(s) OR shows little or no attempt to maintain a setting and a narrator/character(s)
Development	☐ Effectively uses dialogue and/ or descriptions to develop characters and/or situations ☐ Effectively uses grade-appropriate sensory details to convey experiences and events	☐ Adequately uses dialogue and/ or descriptions to develop characters and/or situations ☐ Adequately uses grade-appropriate sensory details to convey experiences and events	☐ Unevenly or minimally uses dialogue and/or descriptions to develop character(s) and/or situations ☐ Unevenly uses grade-appropriate sensory details to convey experiences and events	☐ Shows little or no attempt to use dialogue and/or descriptions to develop character(s) and/or situations ☐ Uses few or no grade-appropriate sensory details to convey experiences and events
Sequencing	☐ Purposefully uses grade-appropriate temporal words and phrases to signal event order	☐ Uses some grade-appropriate temporal words and phrases to signal event order	☐ Uses few grade-appropriate temporal words and phrases to signal event order	☐ Does not use any grade-appropriate temporal words or phrases to signal event order
Introduction and Conclusion	☐ Includes an effective and grade-appropriate introduction and conclusion	☐ Includes an adequate and grade-appropriate introduction and conclusion	☐ Might include a grade-appropriate introduction or conclusion, but one or both are weak.	☐ Does not include an introduction or a conclusion.
Conventions	☐ Is readable with most grade-level conventions used correctly and may use them creatively to enhance the message; minor mistakes do not impede the reader's ability to understand the writer's meaning	☐ Is readable with most grade-level conventions used correctly; mistakes do not affect the reader's ability to understand the writer's meaning	☐ Is readable but some errors in grade-level conventions negatively impact the reader's ability to understand the writer's meaning	☐ Is nearly unreadable due to pervasive errors in grade-level conventions

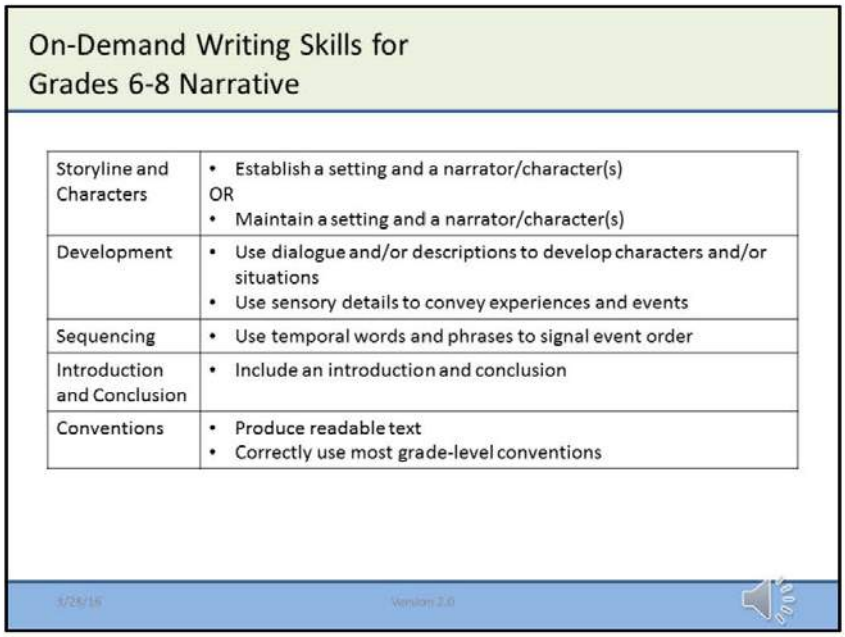
3/28/16 Version 2.0

Audio:

Take a few minutes to read through the narrative rubric.

Make note of the anchor skill or skills listed for each dimension.

29

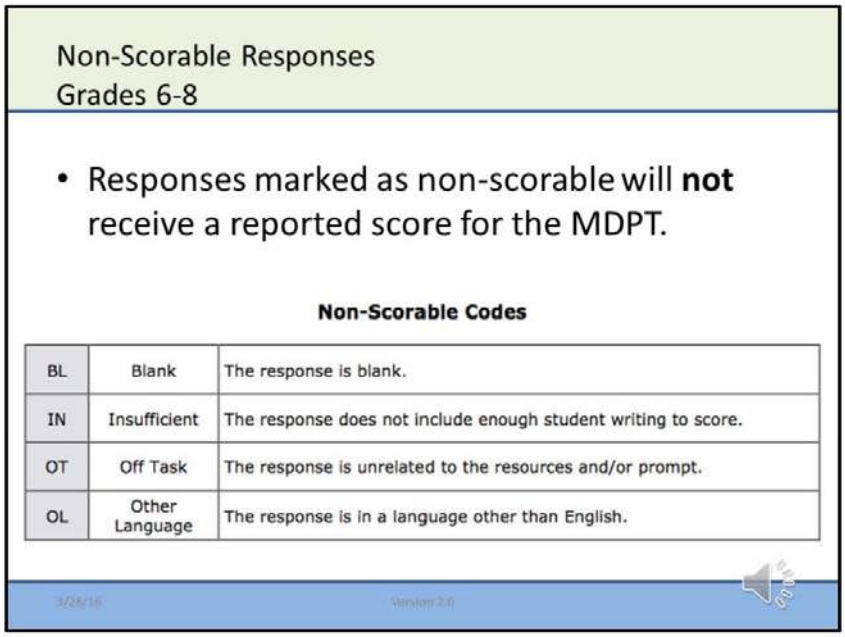


Audio:

Compare your notes to the skills we have listed.

For a narrative on-demand writing task, a student is expected to:

- Establish a setting and a narrator/character(s), OR
- Maintain a setting and a narrator/character(s),
- Use dialogue and/or descriptions to develop characters and/or situations,
- Use sensory details to convey experiences and events,
- Use temporal words and phrases to signal event order,
- Include an introduction and conclusion,
- Produce readable text, and
- Correctly use most grade-level conventions.



Audio:

It is possible that you may read a response that is non-scorable.

Please note that students with responses marked as non-scorable will **not** receive a reported score for the MDPT.

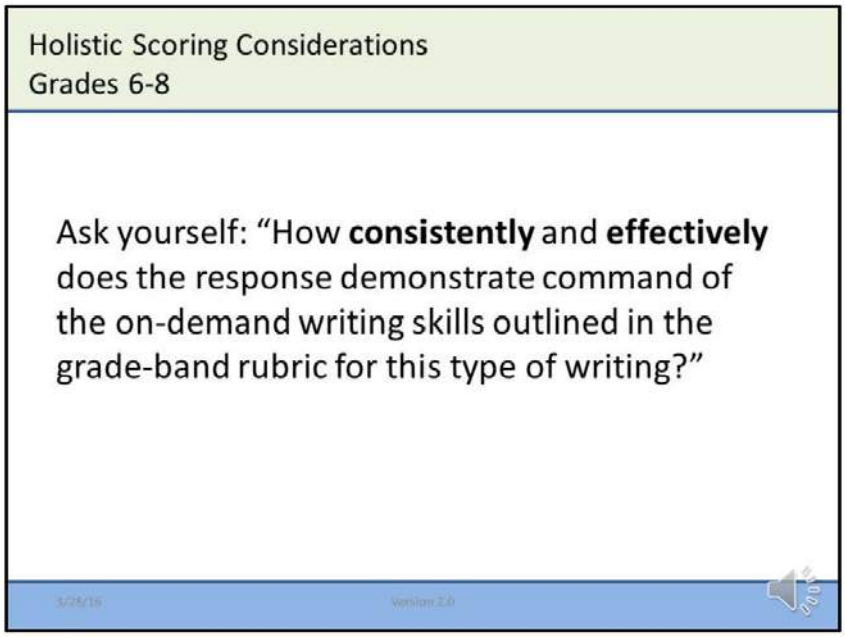
Blank responses and responses written in a language other than English cannot be scored.

A response may be considered insufficient and non-scorable if there are not enough of the student's original words and ideas to score.

A response may be considered off task and non-scorable if the student wrote about something unrelated to the prompt and/or resources.

If you think that a response is non-scorable, we recommend that you visit with a peer within the scoring environment to further analyze the response.

In Part 4, you will learn how to mark non-scorable responses in ScorePoint.



Audio:

Refer to the levels and descriptions in the grade-band rubric as needed; however, calculating a holistic score based on a formula is unnecessary and **not** recommended.

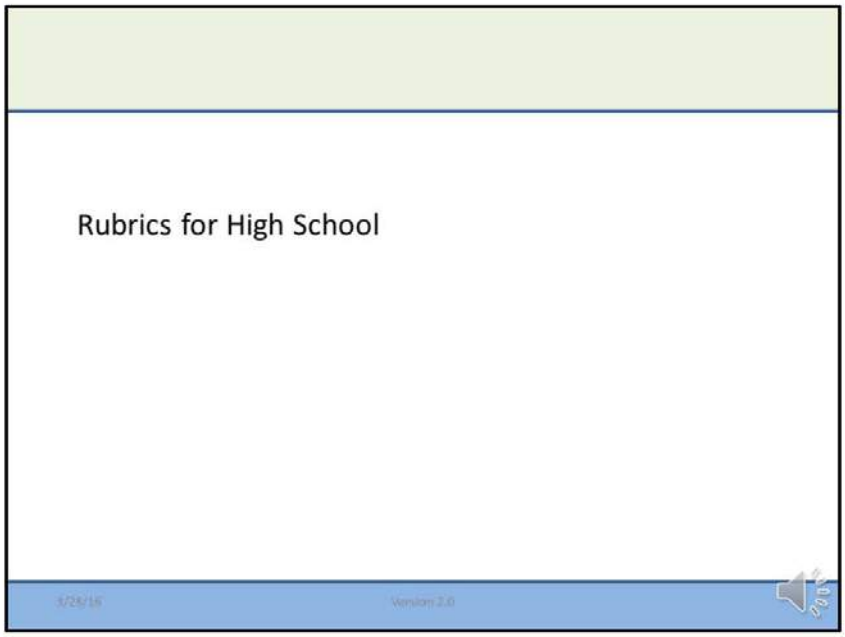
As you read each student response, ask yourself: "How consistently and effectively does the response demonstrate command of the on-demand writing skills outlined in the grade-band rubric for this type of writing?"

Remember to consider the work as a whole as you assign a single score based on your overall judgment of the response.

Prior to scoring student responses, you will score several sample responses and see how experts evaluated them.

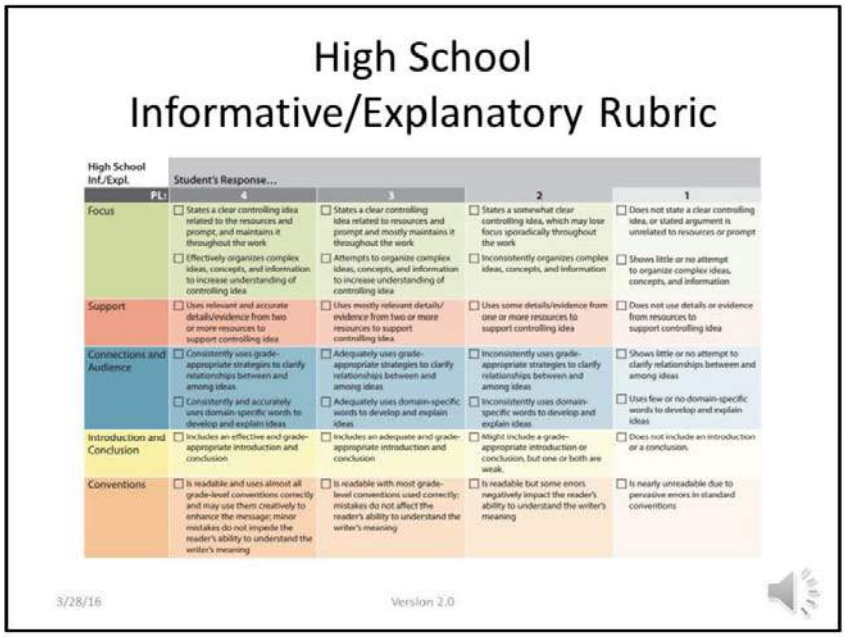
This completes the grade-band training for grades 6-8. If you will also be scoring within another grade band, please complete that section prior to moving on to Part 4 of this training.

Otherwise, please move on to Part 4.



Audio:

In this section of the training, we will identify the on-demand writing skills that are outlined in the rubrics for high school.

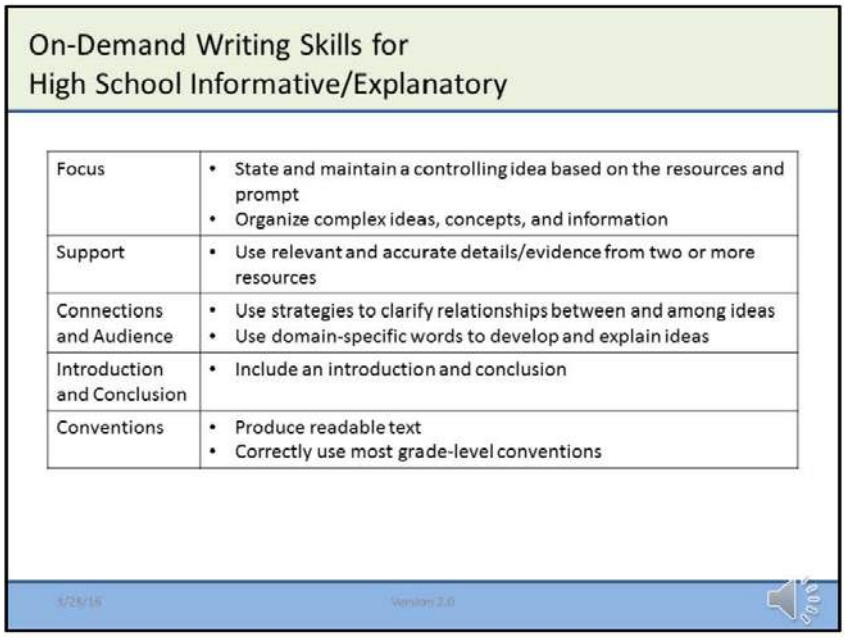


Audio:

Take a few minutes to read through the informative/explanatory rubric.

Make note of the anchor skill or skills listed for each dimension.

For example, being able to "state and maintain a controlling idea based on the resources and prompt" is one of the expected skills in the "Focus" dimension.



Audio:

Compare your notes to the skills we have listed. For an informative/explanatory on-demand writing task, a student is expected to:

- State and maintain a controlling idea based on the resources and prompt,
- Organize complex ideas, concepts, and information,
- Use relevant and accurate details/evidence from two or more resources,
- Use strategies to clarify relationships between and among ideas,
- Use domain-specific words to develop and explain ideas,
- Include an introduction and conclusion,
- Produce readable text, and
- Correctly use most grade-level conventions.

In all grades, the "Conventions" dimension is the same. For an overview of convention-related skills by grade, please refer to the "Grade-Level Conventions" document.

High School Argumentative Rubric

High School Argument	Student's Response...			
PL:	4	3	2	1
Focus/Argument	☐ States a clear argument related to the resources and prompt and maintains it throughout the work ☐ Effectively distinguishes main argument from alternate or opposing arguments	☐ States a clear argument related to resources and prompt and mostly maintains it throughout the work ☐ Attempts to distinguish main argument from alternate or opposing arguments	☐ States a somewhat clear argument, which may lose focus from time to time throughout the work ☐ Recognizes alternate or opposing arguments, but does not adequately distinguish them from the main argument	☐ Does not state a clear argument, or stated argument is unrelated to resources or prompt ☐ Does not recognize or distinguish main argument from alternate or opposing arguments
Evidence	☐ Uses relevant and accurate details/evidence from two or more resources to support argument	☐ Uses mostly relevant and accurate details/ evidence from two or more resources to support argument	☐ Uses some relevant and accurate details/evidence from one or more resources to support argument	☐ Does not use relevant and accurate details or evidence from resources to support argument
Argument	☐ Consistently uses grade-appropriate strategies to clarify relationships between and among ideas, and to connect evidence to argument ☐ Consistently and accurately uses domain-specific words to develop and support argument	☐ Adequately uses grade-appropriate strategies to clarify relationships between and among ideas and to connect evidence to argument ☐ Adequately uses domain-specific words to develop and support argument	☐ Inconsistently uses grade-appropriate strategies to clarify relationships between and among ideas and to connect evidence to argument ☐ Inconsistently uses domain-specific words to develop and support argument	☐ Shows little or no attempt to clarify relationships between and among ideas or connect evidence to argument ☐ Uses few or no domain-specific words to develop and support argument
Introduction and Conclusion	☐ Includes an effective and grade-appropriate introduction and conclusion	☐ Includes an adequate and grade-appropriate introduction and conclusion	☐ Might include a grade-appropriate introduction or conclusion, but one or both are weak.	☐ Does not include an introduction or a conclusion.
Conventions	☐ Is readable and uses almost all grade-level conventions correctly and may use them creatively to enhance the message; minor mistakes do not impede the reader's ability to understand the writer's meaning	☐ Is readable with most grade-level conventions used correctly; mistakes do not affect the reader's ability to understand the writer's meaning	☐ Is readable but some errors negatively impact the reader's ability to understand the writer's meaning	☐ Is nearly unreadable due to pervasive errors in standard conventions

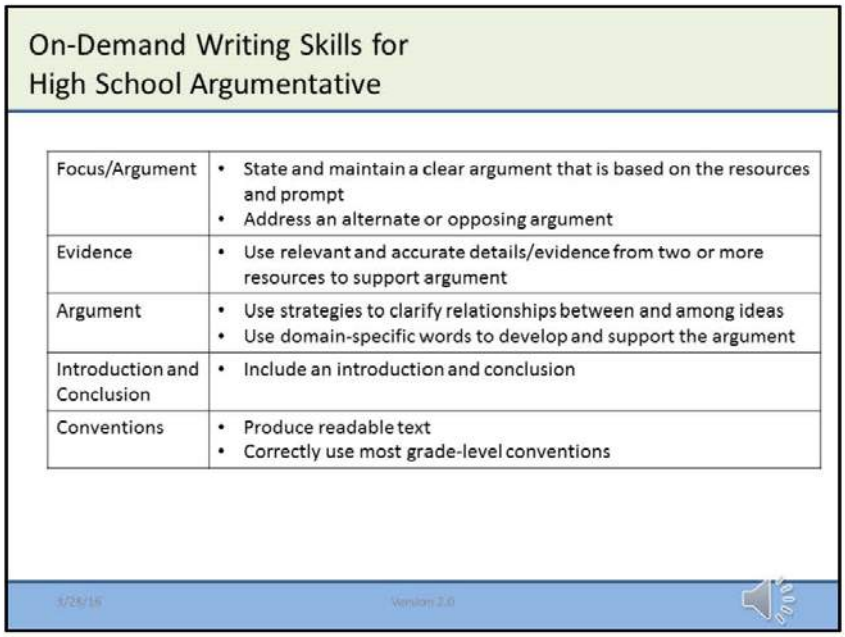
3/28/16 Version 2.0

Audio:

Take a few minutes to read through the argumentative rubric.

Make note of the anchor skill or skills listed for each dimension.

36

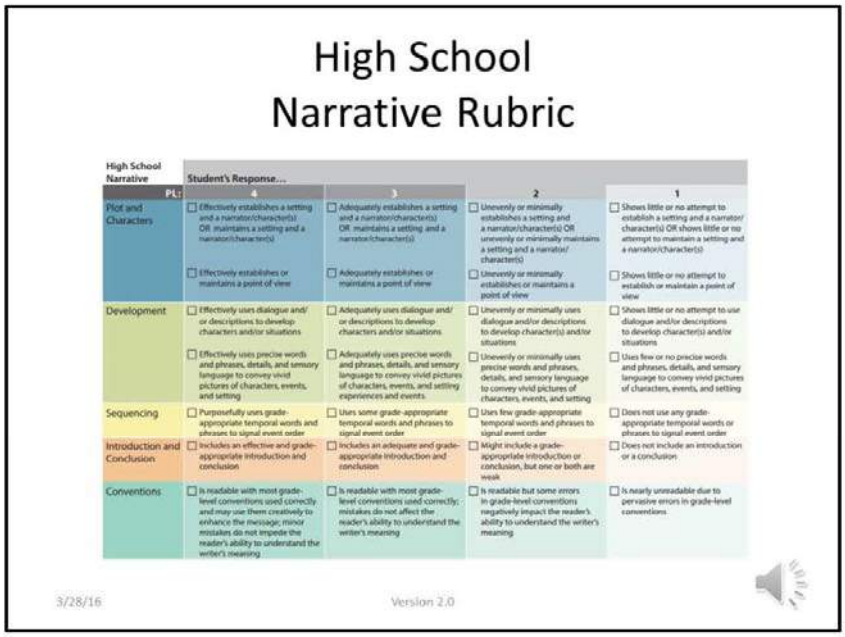


Audio:

Compare your notes to the skills we have listed.

For an opinion on-demand writing task, a student is expected to:

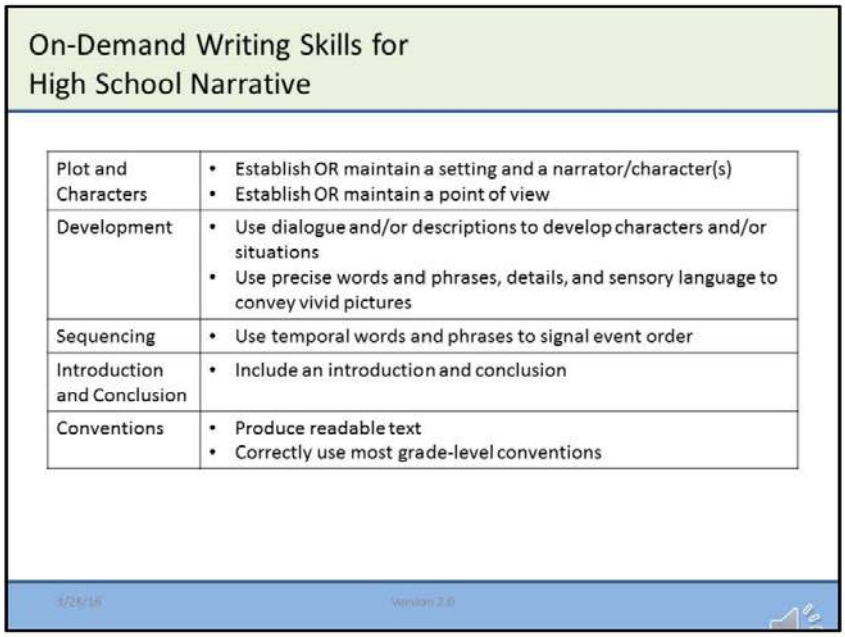
- State and maintain a clear argument that is based on the resources and prompt,
- Address an alternate or opposing argument,
- Use relevant and accurate details/evidence from two or more resources to support argument,
- Use strategies to clarify relationships between and among ideas,
- Use domain-specific words to develop and support the argument,
- Include an introduction and conclusion,
- Produce readable text, and
- Correctly use most grade-level conventions.



Audio:

Take a few minutes to read through the narrative rubric.

Make note of the anchor skill or skills listed for each dimension.

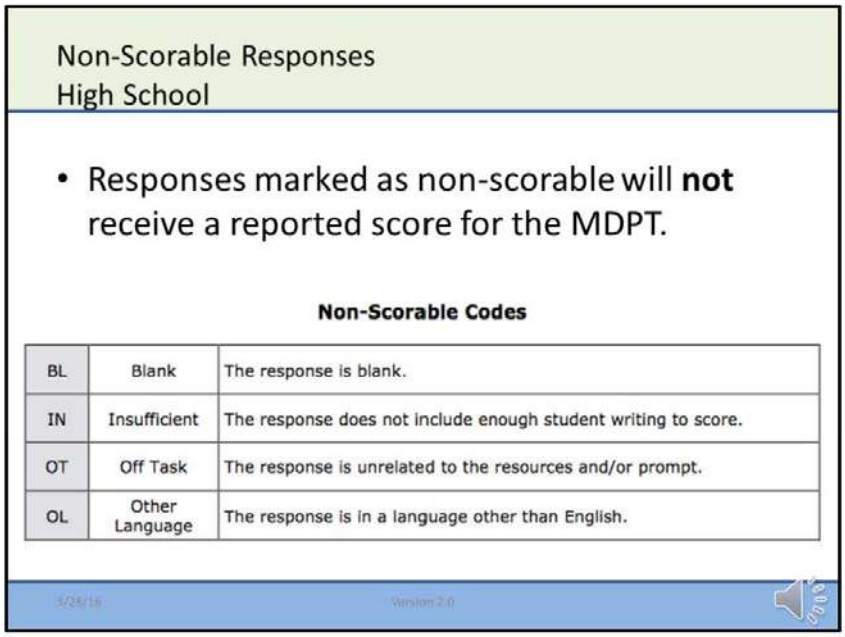


Audio:

Compare your notes to the skills we have listed.

For a narrative on-demand writing task, a student is expected to:

- Establish OR maintain a setting and a narrator/character(s),
- Establish OR maintain a point of view,
- Use dialogue and/or descriptions to develop characters and/or situations,
- Use precise words and phrases, details, and sensory language to convey vivid pictures,
- Use temporal words and phrases to signal event order,
- Produce readable text, and
- Correctly use most grade-level conventions.



Audio:

It is possible that you may read a response that is non-scorable.

Please note that students with responses marked as non-scorable will **not** receive a reported score for the MDPT.

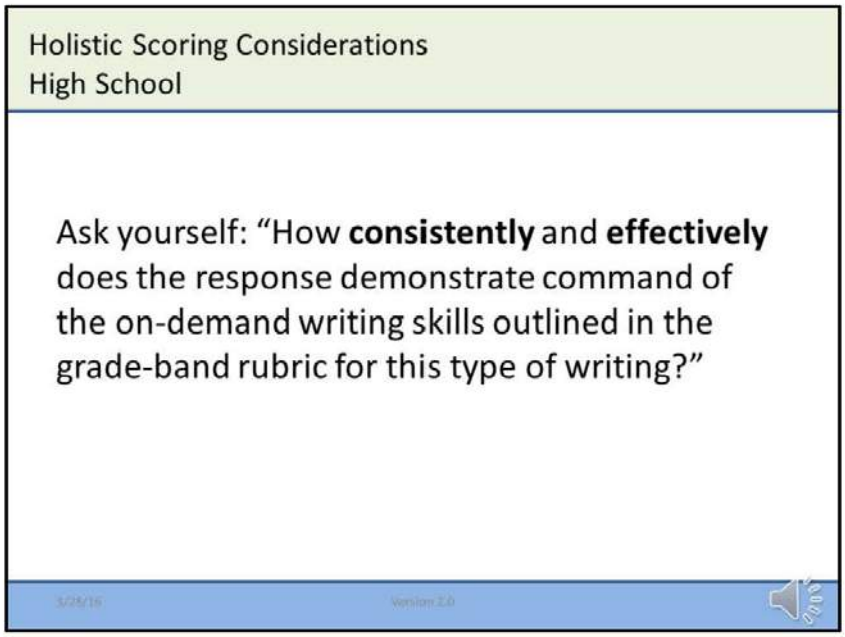
Blank responses and responses written in a language other than English cannot be scored.

A response may be considered insufficient and non-scorable if there are not enough of the student's original words and ideas to score.

A response may be considered off task and non-scorable if the student wrote about something unrelated to the prompt and/or resources.

If you think that a response is non-scorable, we recommend that you visit with a peer within the scoring environment to further analyze the response.

In Part 4, you will learn how to mark non-scorable responses in ScorePoint.



Audio:

Refer to the levels and descriptions in the grade-band rubric as needed; however, calculating a holistic score based on a formula is unnecessary and **not** recommended.

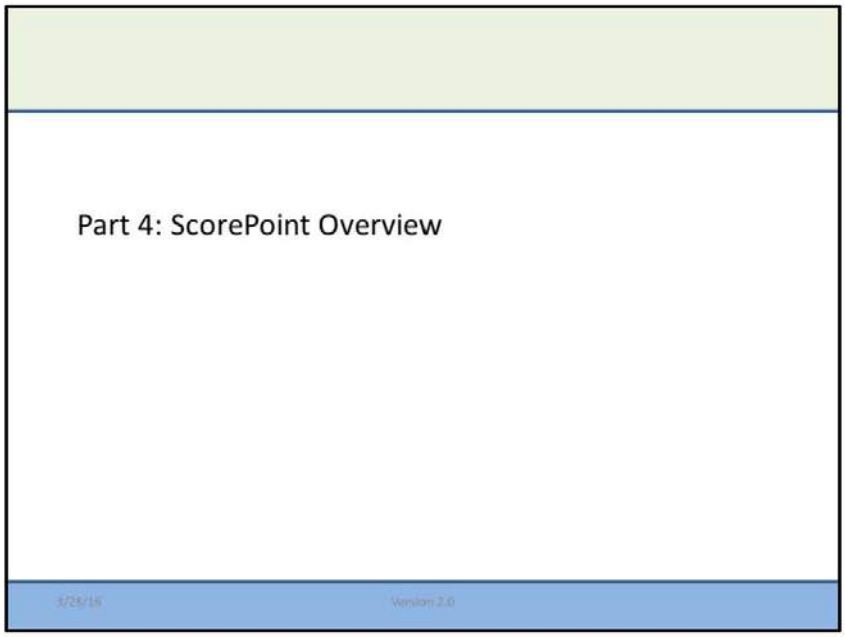
As you read each student response, ask yourself: "How consistently and effectively does the response demonstrate command of the on-demand writing skills outlined in the grade-band rubric for this type of writing?"

Remember to consider the work as a whole as you assign a single score based on your overall judgment of the response.

Prior to scoring student responses, you will score several sample responses and see how experts evaluated them.

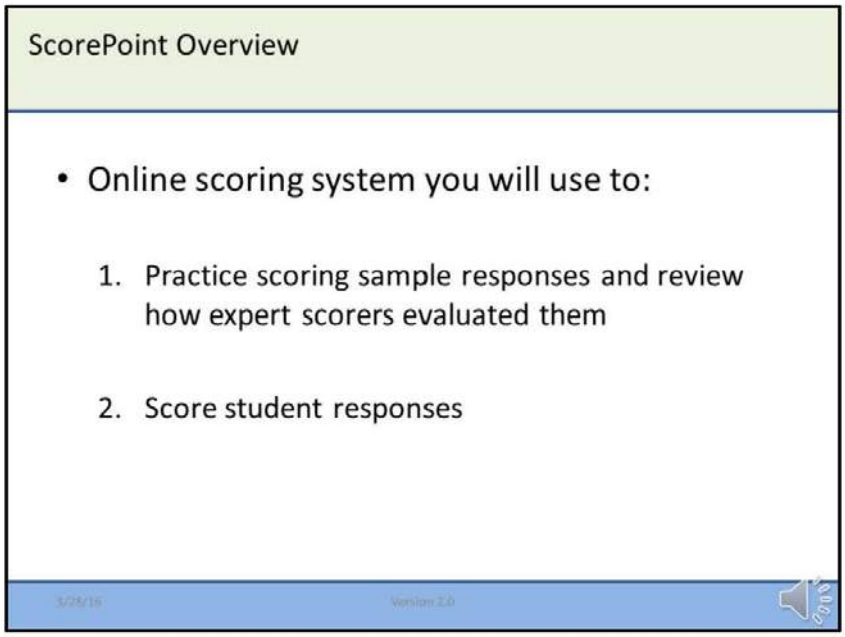
This completes the grade-band training for high school. If you will also be scoring within another grade band, please complete that section prior to moving on to Part 4 of this training.

Otherwise, please move on to Part 4.



Audio:

None



Audio:

ScorePoint is the online scoring system that you will use to score the MDPT responses.

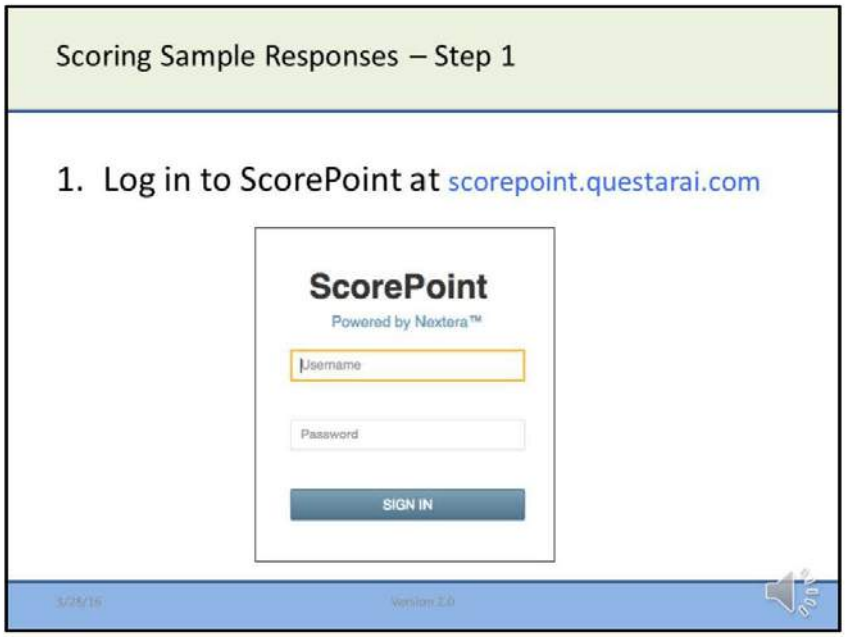
First you will use ScorePoint to practice scoring sample responses and review how expert scorers evaluated them.

Then you will use ScorePoint to score student responses.

This section of the training will prepare you for navigating ScorePoint for these two purposes.

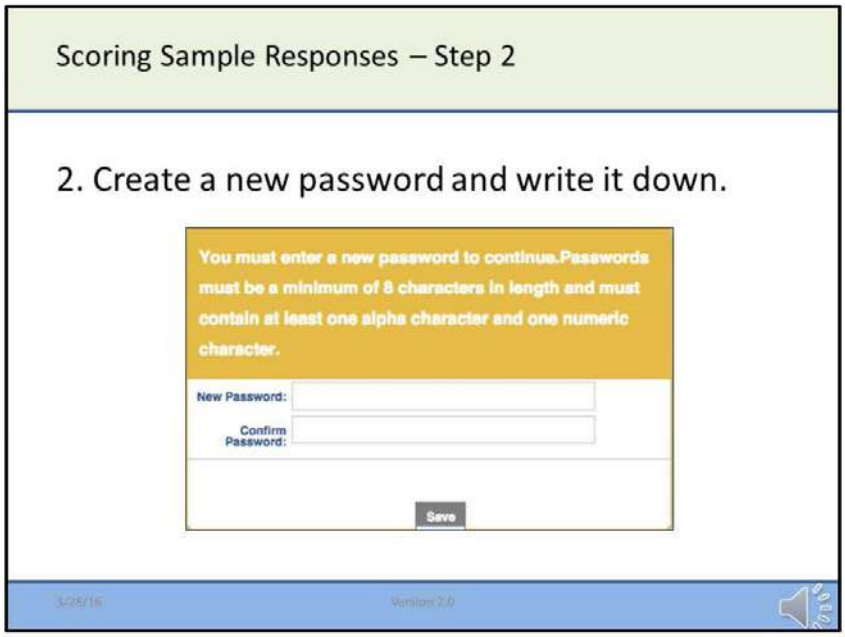
Please complete this section prior to logging in to ScorePoint.

Refer back to this training as needed.



Audio:

Log in to ScorePoint using the username and temporary password you received from your school or district test coordinator.



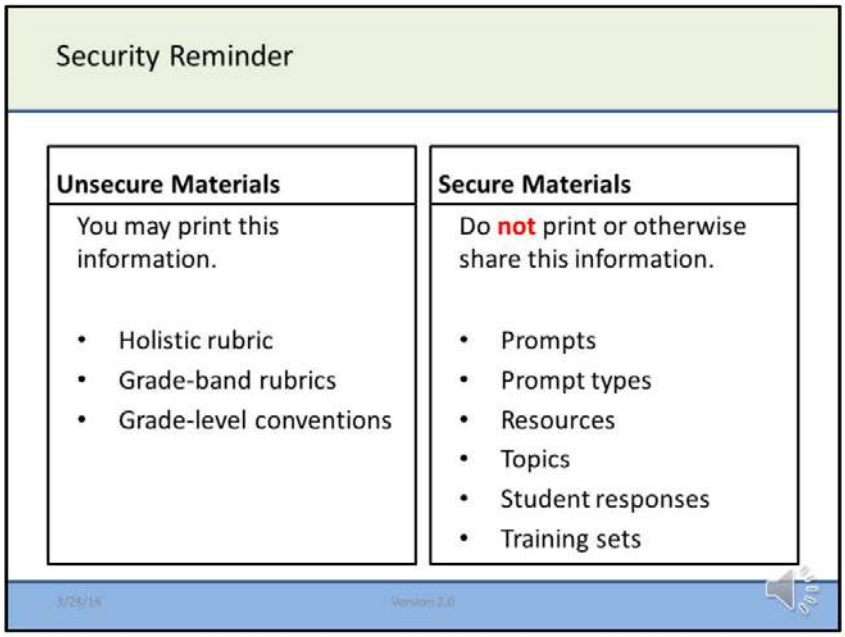
Audio:

You will be prompted to create and save a new password.

Please make note of your new password because your password cannot be reset locally.

If you forget your password, please contact kap@ku.edu.

It may take up to 24 hours to receive a new password.

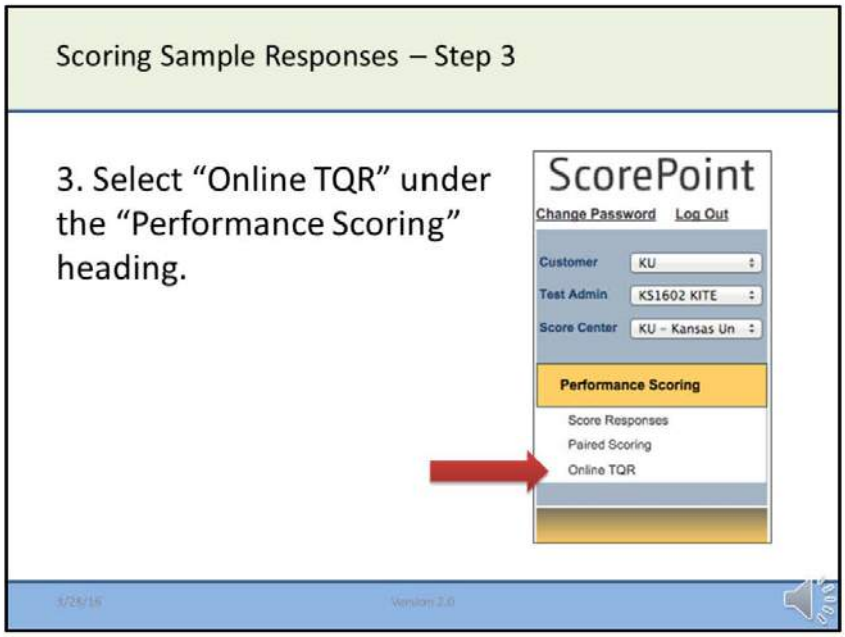


Audio:

As you enter ScorePoint and begin to review responses, please remember that you will be working with secure materials.

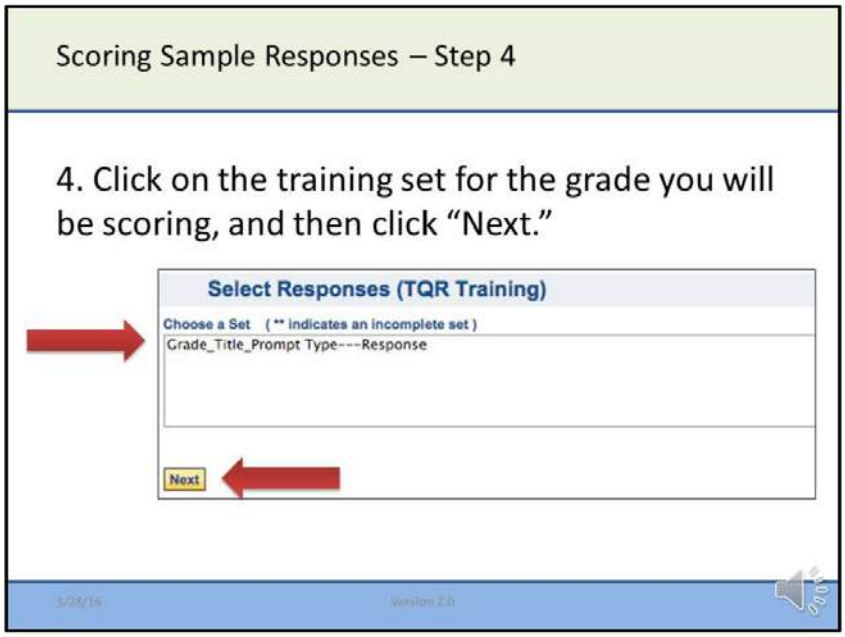
It is critical that the MDPTs are kept secure and that normal test security protocols apply.

For example, please do **not** email any information that describes a particular task in any way, discuss specific tasks or responses with anyone outside of the scoring environment, allow anyone else to use your username and password, take pictures of tasks or responses with a camera, or do anything else that could detract from the validity of the assessment.



Audio:

On the left side of the screen, select "Online TQR" under the "Performance Scoring" heading.

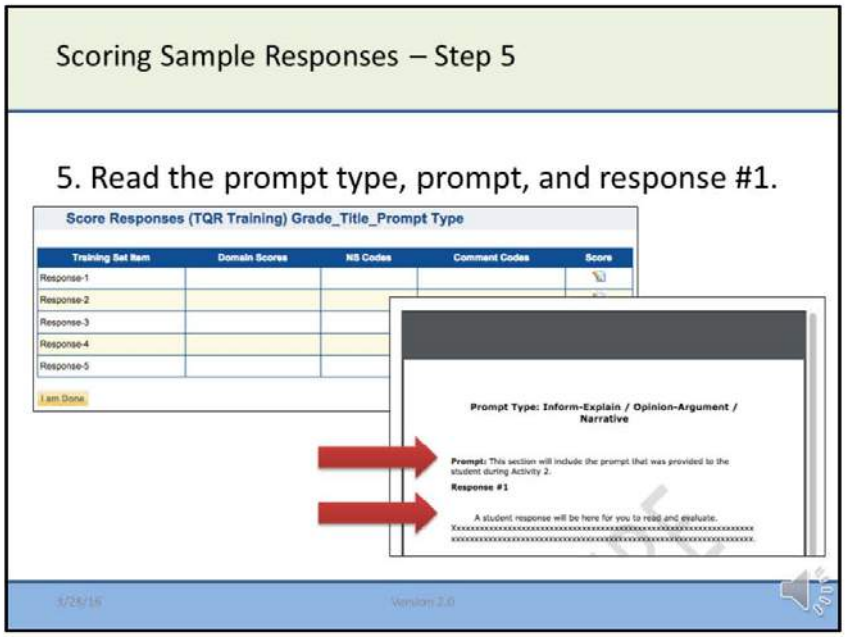


Audio:

Complete the grade-level training sets for every grade that your school or district test coordinator assigned to you.

For example, if you are assigned to score grade 3 and grade 5, complete the training set that begins with "Grade 3" and the training set that begins with "Grade 5."

Click on the set, and then click "Next."



Audio:

The training set scoring table lists the number of sample responses you will score and review.

You will score a sample response and then review the expert score and rationale for that sample response before moving on to the next response.

Below the training set scoring table, you will see an image of a PDF.

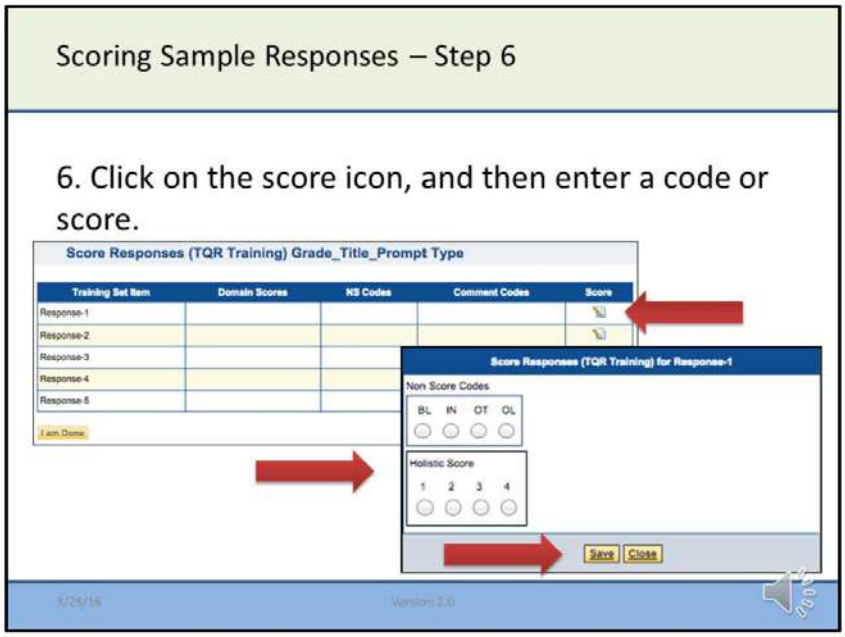
On the first page, read the prompt type and prompt.

Remember, it will be helpful to have paper copies of the rubrics and non-scorable codes in front of you during training and scoring.

Please note that when you are scoring summative student responses in ScorePoint, you can view these materials as well as the resources the student used to write the response.

Then read response #1 with both rubrics and the non-scorable codes in mind.

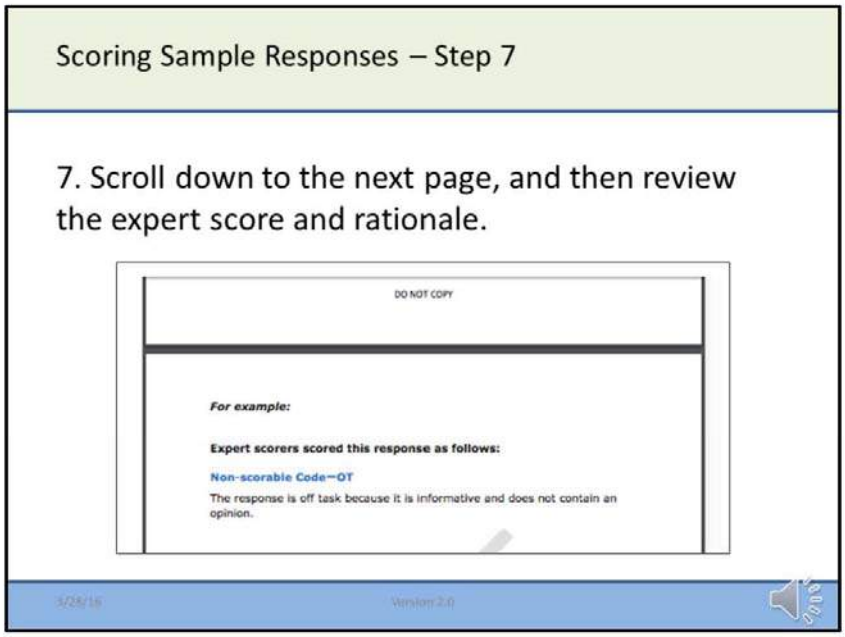
Consider collaborating with others or taking a few notes as you work through your thinking.



Audio:

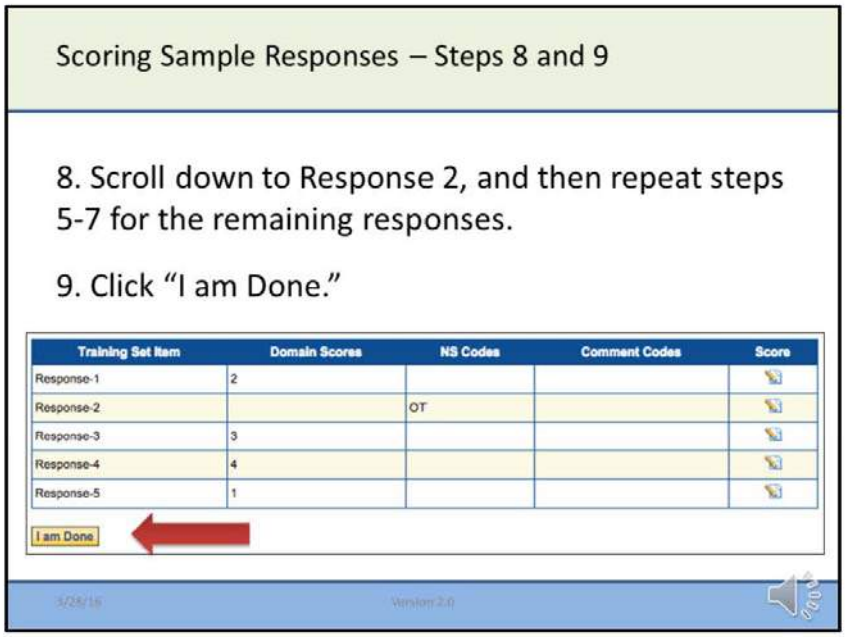
Click on the icon in the "Score" column for Response 1.

Select a code or score in the pop-up window, and then click "Save."



Audio:

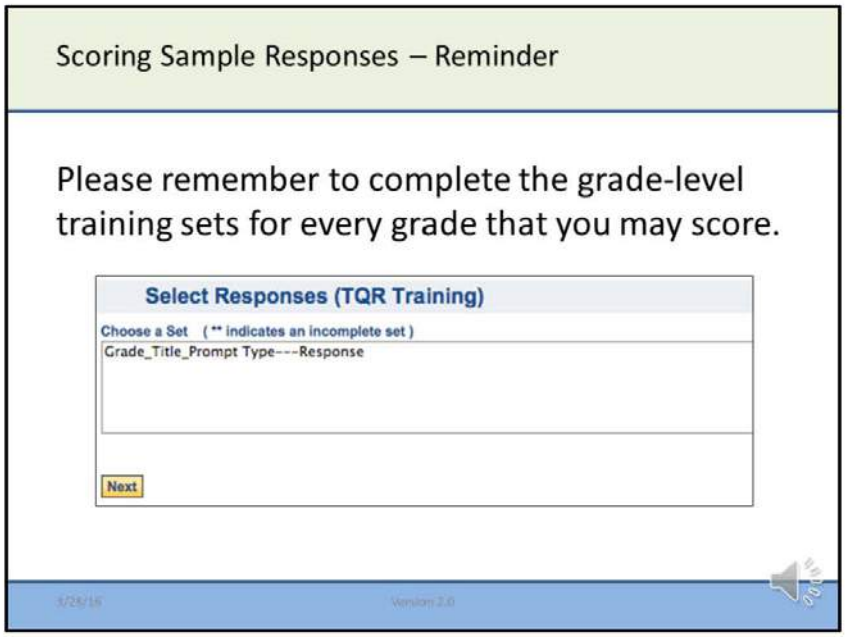
Scroll down to the next page, and then review the expert score and rationale.



Audio:

Scroll down to Response 2, and then repeat steps 5-7 for the remaining responses listed in the Scoring Set Item Table.

When you have scored and reviewed all of the responses for this grade, click "I am Done."

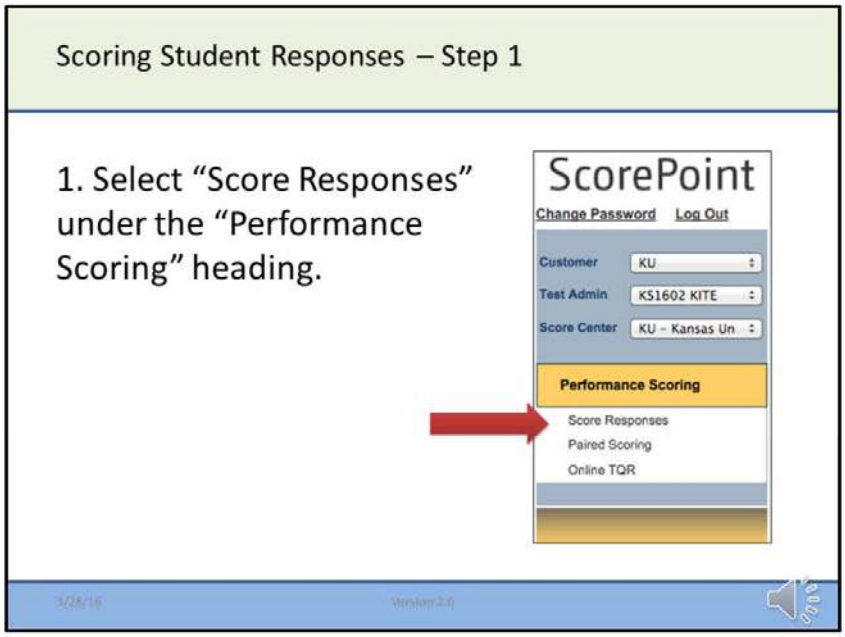


Audio:

You will return to the list of available training sets.

Please remember that prior to scoring summative responses you need to complete the grade-level training sets for every grade that you may score.

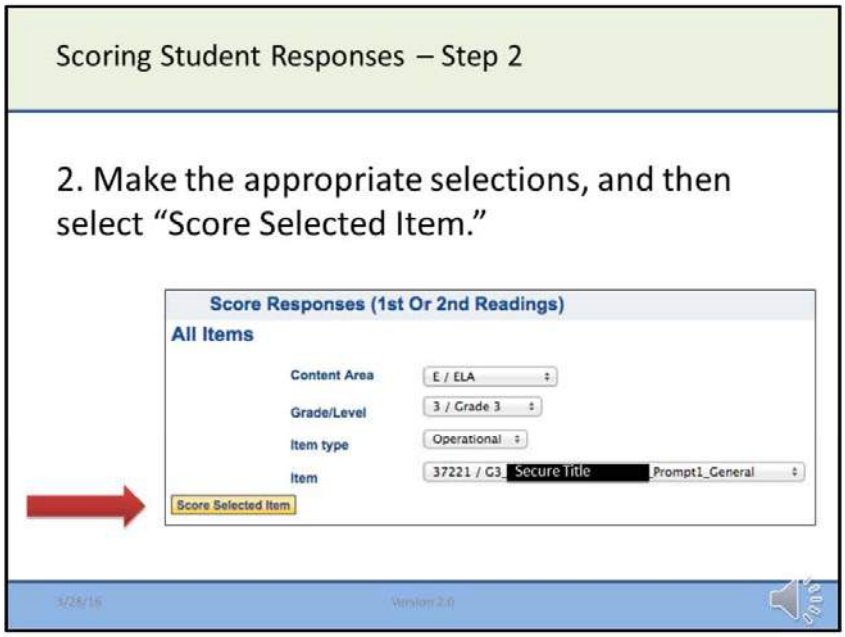
This will ensure that you are fully prepared for summative scoring for those grades.



Audio:

Now we will go over how to use ScorePoint to score summative student responses.

On the left side of the screen, select "Score Responses" under the "Performance Scoring" heading.



Audio:

For "Content Area," select "E / ELA."

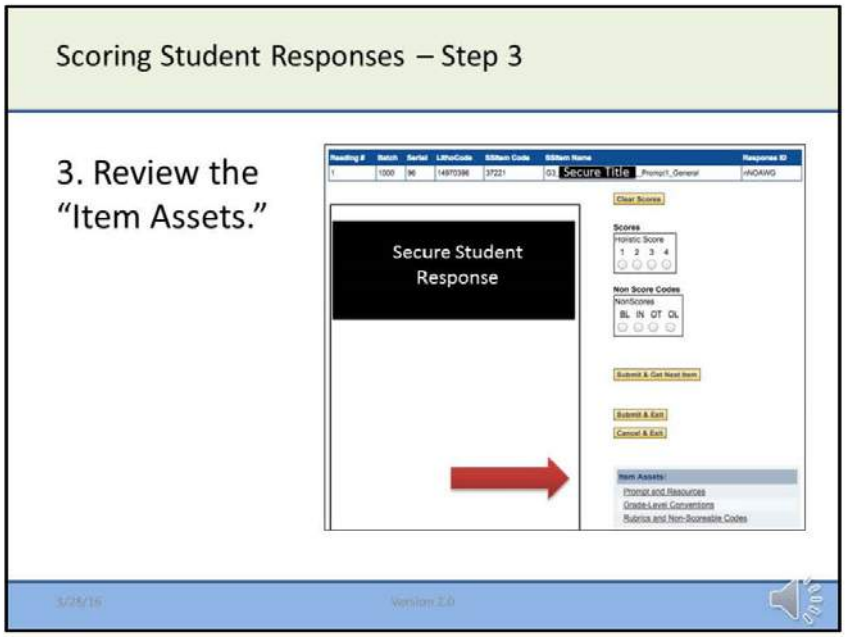
For "Grade/Level," select a grade that you are assigned to score.

Select an item to score.

For grades 3 through 8 and 11, there will be one item listed.

Because grade 10 was a field test this year, you may choose from any of the items listed.

Click "Score Selected Item."



Audio:

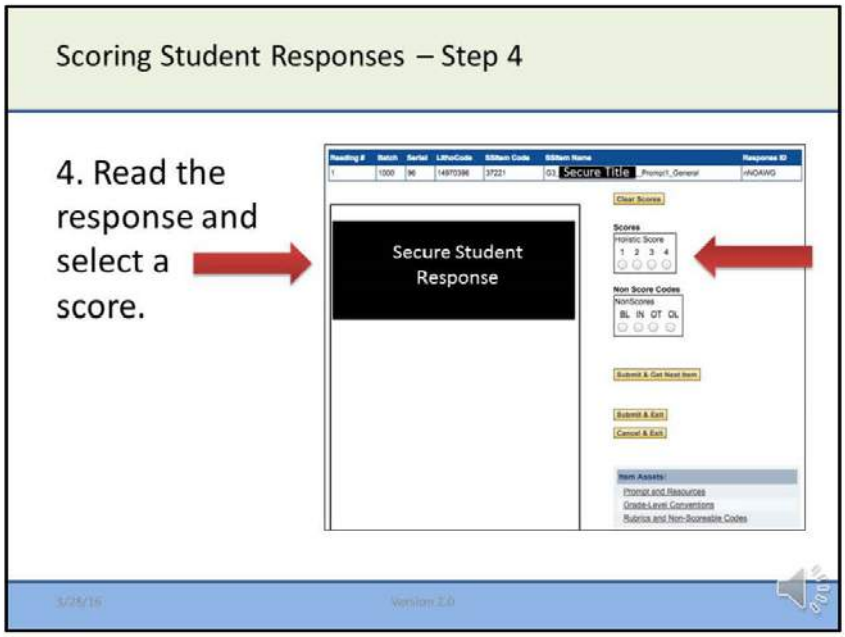
Review the "Item Assets" prior to scoring the first student response.

The assets include a copy of the prompt and resources the student used to write the response, so they provide necessary context.

In addition, there is a copy of the Grade-Level Conventions document as well as the holistic rubric, grade-band rubric, and non-scorable codes.

Revisit the assets as needed as you score responses.

For grade 10, be sure to review the assets any time you select a different item to score because each item has a different prompt.



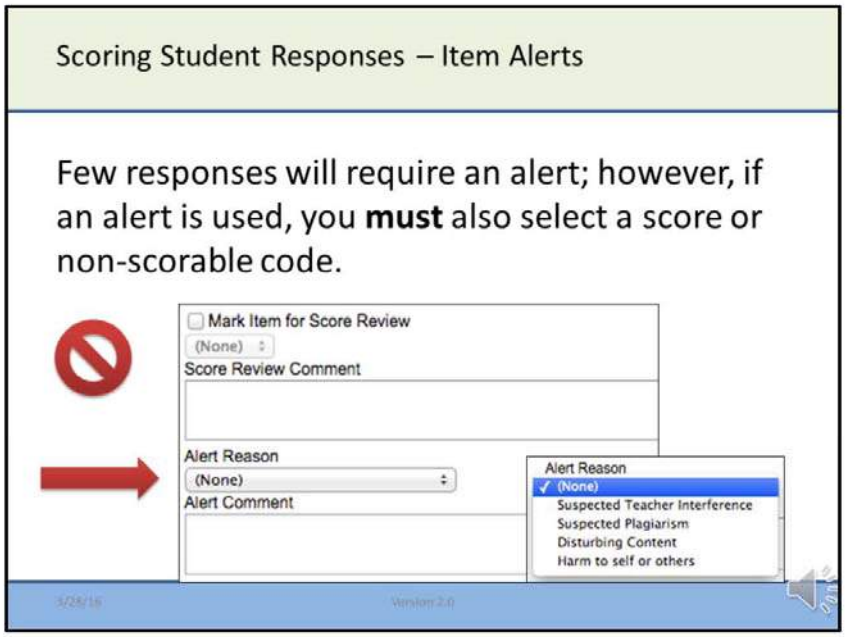
Audio:

Next, read the student response and select a score in the "Holistic Score" box.

Remember, students with responses marked as non-scorable will **not** receive a reported score for the MDPT.

With this in mind, if you think that a response is non-scorable, we recommend that you visit with a peer within the scoring environment to further analyze the response.

Select "Clear Scores," to clear the score and code boxes.



Audio:

At the bottom of the screen, you will see some additional options.

You will **not** use the "Mark Item for Score Review" option.

Collaborate with a peer if you would like someone to review your score before submission.

The alert option allows readers to flag a response if it contains concerning content.

Most responses will **not** require the use of the alerts feature; however, it is important to review these options.

If you suspect teacher interference or plagiarism, collaborate with a peer.

Most likely, these responses should be marked non-scorable and should not require an alert.

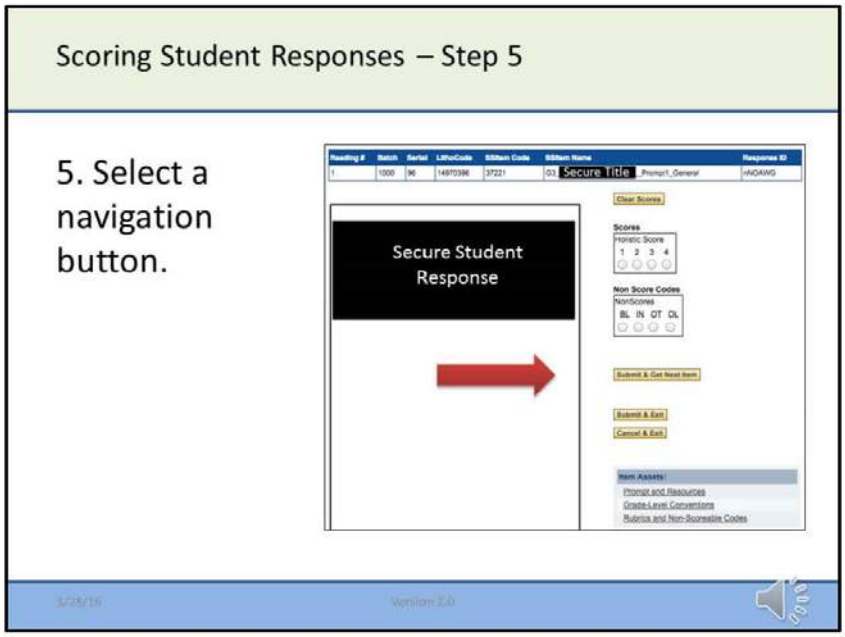
If a response includes disturbing content or indicates possible harm, we recommend that you mark it with the "Harm to self or other" alert.

The "Harm to self or other" alert will automatically contact CETE so that the school and student can be reached immediately.

A "Disturbing Content" alert will be routed to CETE; however, an email will not be generated, so time may pass before the alert is reviewed.

Comments are not required, but they may be helpful.

To submit an alert, you **must** provide a score or non-scorable code and submit the response.



Audio:

Select "Submit & Get Next Item" to assign a score to the response and move to the next response.

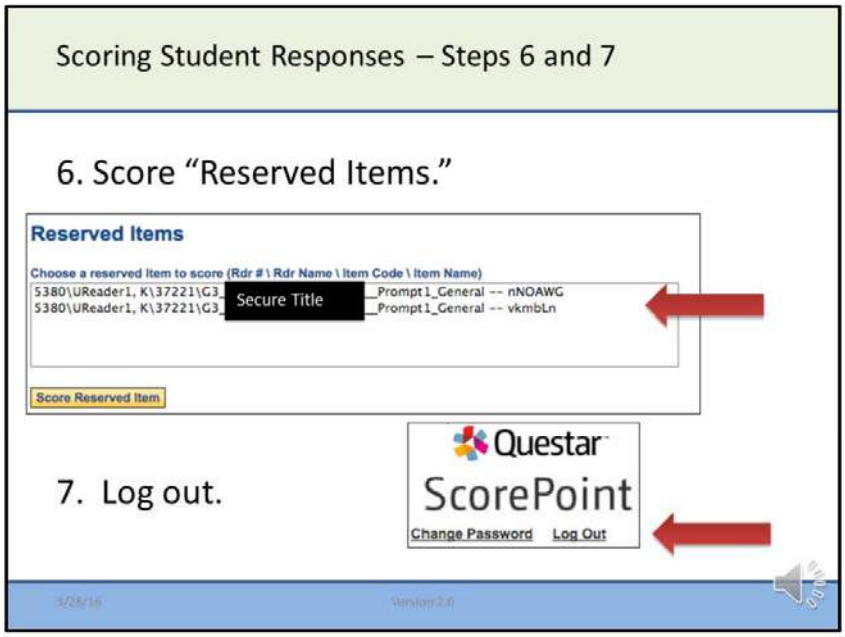
Select "Submit & Exit" to assign a score and move to the item selection page.

Once you select a "Submit" option, you cannot return to the response.

If you need to leave a response without scoring it, select "Cancel & Exit" so that the response will return to the response pool.

You will not be able to return to this response.

When you are ready to log out, use one of the "Exit" options so that you are returned to the item selection page.



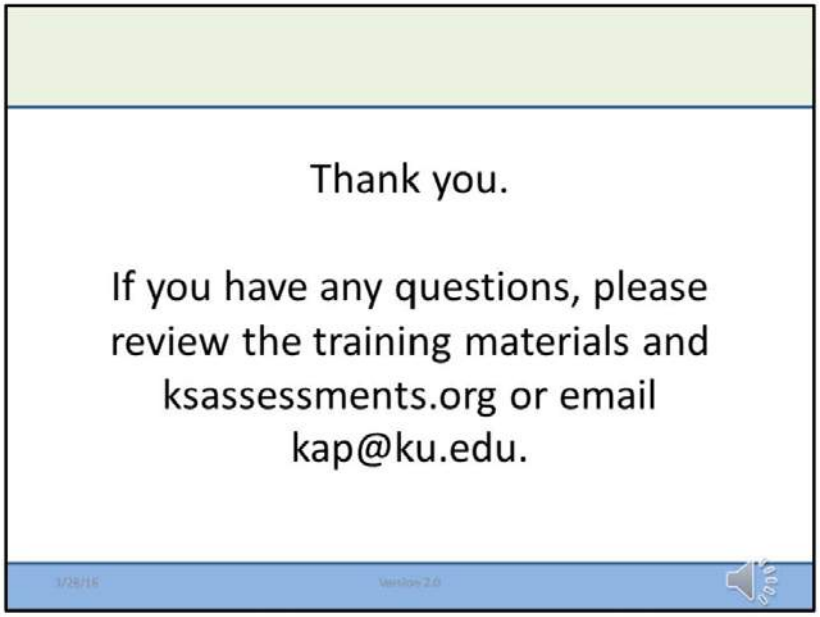
Audio:

Before logging out, look at the "Reserved Items" section.

If you have any responses listed there, please score them.

CETE will periodically clear the reserved responses, so you may not see anything in this section.

To exit ScorePoint, select "Log Out" below the ScorePoint logo.



Audio:

If you have any questions, please review the training materials and the information on CETE's Kansas Assessments webpage.

If you need additional support, please email kap@ku.edu.

In your email, please include a description of your problem; however, do not include screenshots or specific information related to tasks or responses.

Appendix F: HGSS Performance Level Descriptors (PLDs)

Grade 6 Ancient World History		
Level 2	**Level 3**	**Level 4**
Students scoring in the Level 2 range typically can apply social science knowledge to respond to basic historic questions and interact with a variety of familiar primary and secondary source documents.	Students scoring in the Level 3 range typically can apply social science knowledge and strategies to pose or respond to historic questions and interact with a variety of primary and secondary source documents.	Students scoring in the Level 4 range typically can apply social science knowledge, strategies, and analysis to pose or respond to complex historic questions and interact with a variety of primary and secondary source documents, including unfamiliar sources.
Students can identify choices, consequences, rights, responsibilities, beliefs, ideas, diversity, continuity, change, and dynamic relationships of past civilizations and cultures. • Students can identify and trace basic aspects of the rise and fall of different civilizations and cultures. • Students can identify basic elements of world religions and the role religion played in major historical events.	Students can identify and compare choices, consequences, rights, responsibilities, beliefs, ideas, diversity, continuity, change, and dynamic relationships of past civilizations and cultures. • Students can identify and compare aspects of the rise and fall of different civilizations and cultures. • Students can identify and compare elements of world religions and the role religion played in historical events and civilizations.	Students can compare, contrast, and analyze choices, consequences, rights, responsibilities, beliefs, ideas, diversity, continuity, change, and dynamic relationships of past civilizations and cultures. • Students can analyze, compare, and contrast the rise and fall of different civilizations and cultures and posit possible alternative paths history might have taken. • Students can identify, compare, and contrast complex ideas from world religions and their immediate or lasting impact on history, civilization, and culture.
Students can identify the impact of geography on choices, consequences, rights, responsibilities, beliefs, ideas, diversity, continuity, change, and dynamic relationships in the development of civilizations.	Students can identify and analyze the impact of geography on choices, consequences, rights, responsibilities, beliefs, ideas, diversity, continuity, change, and dynamic relationships in the development of civilizations.	Students can evaluate the impact of geography on choices, consequences, rights, responsibilities, beliefs, ideas, diversity, continuity, change, and dynamic relationships in the development of civilizations.
Students can recognize the role of economics in the choices, consequences,	Students can recognize and analyze the role of economics in the choices, consequences,	Students can evaluate the role of economics in the choices, consequences, rights,

Grade 6 Ancient World History		
Level 2	**Level 3**	**Level 4**
rights, responsibilities, beliefs, ideas, diversity, continuity, change, and dynamic relationships in the spread and/or transformation of civilization, cultures.	rights, responsibilities, beliefs, ideas, diversity, continuity, change, and dynamic relationships in the spread and/or transformation of cultures.	responsibilities, beliefs, ideas, diversity, continuity, change, and dynamic relationships in the spread and/or transformation of cultures, and can trace cultural and ideological transformation across societies.
Students can recognize the role of politics and power in the choices, consequences, rights, responsibilities, beliefs, ideas, diversity, continuity, change, and dynamic relationships in the spread and/or transformation of civilizations and cultures. • Students can identify the basic rights and responsibilities of people in various civilizations.	Students can recognize and analyze the role of politics and power in the choices, consequences, rights, responsibilities, beliefs, ideas, diversity, continuity, change, and dynamic in the spread and/or transformation of civilizations and cultures. • Students can analyze and compare and contrast the rights and responsibilities of people in various civilizations.	Students can evaluate the role of politics and power in the choices, consequences, rights, responsibilities, beliefs, ideas, diversity, continuity, change, and dynamic relationships in the spread and/or transformation of civilizations and cultures and can trace cultural and ideological transformation across societies. • Students can evaluate the rights and responsibilities of people in various civilizations and their impact on the development of those civilizations.
Students can identify connections between past civilization's choices, consequences, rights, responsibilities, beliefs, ideas, diversity, continuity, change, and dynamic relationships and the modern world. • Students can identify examples of continuity and change over time.	Students can identify and draw conclusions about the connections between past civilization's choices, consequences, rights, responsibilities, beliefs, ideas, diversity, continuity, change, and dynamic relationships and the modern world. • Students can identify examples of continuity and change over time and analyze their impact on civilizations and cultures.	Students can identify, compare, contrast, analyze, and draw conclusions making connections between the choices, consequences, rights, responsibilities, beliefs, ideas, diversity, continuity, change, and dynamic relationships of past civilizations and the modern world. • Students can evaluate continuity and change over time and their impact on civilizations and cultures.
Students can identify	Students can identify and	Students can identify,

Grade 6 Ancient World History		
Level 2	**Level 3**	**Level 4**
different types of historical documents (e.g., primary vs. secondary sources).	analyze different types of historical documents and the role and importance of each type.	analyze, and evaluate different types of historical documents and the role, importance, and validity of each type.
Students can read and use information from a variety of primary and secondary sources to develop a claim or position using direct evidence and argument, to support their ideas.	Students can read, analyze, and use information from a variety of primary and secondary sources to develop a claim or position using direct evidence and argument, to support their ideas.	Students can read, analyze, and use information from a variety of primary and secondary sources to develop a claim or position using direct evidence and argument, to support their ideas or refute the position of another agent.

Grade 8 US History: Constitutional Age to International Expansion		
Level 2	**Level 3**	**Level 4**
Students scoring in the Level 2 range typically can apply social science knowledge to respond to basic historic questions and interact with a variety of familiar primary and secondary source documents.	Students scoring in the Level 3 range typically can apply social science knowledge and strategies to pose or respond to historic questions and interact with a variety of familiar primary and secondary sources.	Students scoring in the Level 4 range typically can apply social science knowledge, strategies, and analysis to pose or respond to complex historic questions and interact with a variety of primary and secondary source documents, including unfamiliar sources.
Students can identify critical choices, consequences, rights, responsibilities, beliefs, ideas, diversity, continuity, change, and dynamic relationships of the period in the United States from 1789 to the 20th century. • Students can recognize and trace different points of view about such topics as: Westward Expansion, slavery, Reconstruction, and women's and workers' rights.	Students can identify and compare critical choices, consequences, rights, responsibilities, beliefs, ideas, diversity, continuity, change, and dynamic relationships of the period in the United States from 1789 to the 20th century. • Students can analyze different points of view about such topics as: Westward Expansion, slavery, Reconstruction, and women's and workers' rights.	Students can identify, compare, contrast, and analyze critical choices, consequences, rights, responsibilities, beliefs, ideas, diversity, continuity, change, and dynamic relationships of the period in the United States from 1789 to the 20th century. • Students can evaluate different points of view about such topics as: Westward Expansion, slavery, Reconstruction, and women's and workers' rights.
Students can recognize the	Students can recognize and	Students can evaluate the role

Grade 8 US History: Constitutional Age to International Expansion		
Level 2	**Level 3**	**Level 4**
role of economics on the choices, consequences, rights, responsibilities, beliefs, ideas, diversity, continuity, change, and dynamic relationships in the United States prior to the 20th century.	analyze the role of economics on the choices, consequences, rights, responsibilities, beliefs, ideas, diversity, continuity, change, and dynamic relationships of the United States prior to the 20th century.	of economics on the choices, consequences, rights, responsibilities, beliefs, ideas, diversity, continuity, change, and dynamic relationships in the United States prior to the 20th century.
Students can recognize aspects of the role of geography in the choices, consequences, rights, responsibilities, beliefs, ideas, diversity, continuity, change, and dynamic relationships in the United States prior to the 20th century.	Students can recognize and analyze the role of geography on the choices, consequences, rights, responsibilities, beliefs, ideas, diversity, continuity, change, and dynamic relationships of the United States prior to the 20th century.	Students can recognize, evaluate, and analyze the role of geography on the choices, consequences, rights, responsibilities, beliefs, ideas, diversity, continuity, change, and dynamic relationships in the United States prior to the 20th century.
Students can recognize aspects of the role of politics and power in the choices, consequences, rights, responsibilities, beliefs, ideas, diversity, continuity, change, and dynamic relationships in the United States prior to the 20th century. • Students can identify the basic rights and responsibilities of the people and the US government.	Students can recognize and analyze the role of politics and power on the choices, consequences, rights, responsibilities, beliefs, ideas, diversity, continuity, change, and dynamic relationships of the United States prior to the 20th century. • Students can analyze and compare and contrast the basic rights and responsibilities of the people and the US government.	Students can recognize, evaluate, and analyze the role of politics and power on the choices, consequences, rights, responsibilities, beliefs, ideas, diversity, continuity, change, and dynamic relationships in the United States prior to the 20th century. • Students can evaluate the basic rights and responsibilities of the people and the US government.
Students can identify connections between the choices, consequences, rights, responsibilities, beliefs, ideas, diversity, continuity, change, and dynamic relationships of the US prior to the 20th century and the modern world. • Students can identify examples of continuity and change over time.	Students can identify and draw conclusions about the connections between the choices, consequences, rights, responsibilities, beliefs, ideas, diversity, continuity, change, and dynamic relationships of the US prior to the 20th century and the modern world. • Students can identify examples of continuity and change over time	Students can identify, analyze, draw conclusions, and create arguments about the connections between the choices, consequences, rights, responsibilities, beliefs, ideas, diversity, continuity, change, and dynamic relationships of the US prior to the 20th century and the modern world. • Students can evaluate continuity and change over time and their

Grade 8 US History: Constitutional Age to International Expansion		
Level 2	**Level 3**	**Level 4**
	and analyze their impact.	impact.
Students can identify different types of historical documents (e.g., primary vs. secondary sources).	Students can identify and analyze different types of historical documents and the role and importance of each type.	Students can identify, analyze and evaluate different types of historical documents and the role, importance, and validity of each type.
Students can read and use information from a variety of primary and secondary sources to develop a claim or position using direct evidence and argument, to support their ideas.	Students can read, analyze, and use information from a variety of primary and secondary sources to develop a claim or position using direct evidence and argument, to support their ideas.	Students can read, analyze, and use information from a variety of primary and secondary sources to develop a claim or position using direct evidence and argument, to support their ideas or refute the position of another agent.

Grade 11 US History		
Level 2	**Level 3**	**Level 4**
Students scoring in the Level 2 range typically can apply social science knowledge to respond to basic historic questions and interact with a variety of familiar primary and secondary sources.	Students scoring in the Level 3 range typically can apply social science knowledge and strategies to pose or respond to historic questions and interact with a variety of primary and secondary sources.	Students scoring in the Level 4 range typically can apply social science knowledge, strategies, and analysis to pose or respond to complex historic questions and interact with a variety of primary and secondary sources, including unfamiliar sources.
Students can identify critical choices, consequences, rights, responsibilities, beliefs, ideas, diversity, continuity, change, and dynamic relationships of 20th and 21st century U.S. history. • Students can identify and trace the chain of events leading to important events such as: the labor rights movement, the Great Depression, World Wars 1 & 2, the Cold War,	Students can identify and compare critical choices, consequences, rights, responsibilities, beliefs, ideas, diversity, continuity, change, and dynamic relationships of 20th and 21st century U.S. history. • Students can analyze the chain of events leading to important events such as: the labor rights movement, the Great Depression, World Wars 1 & 2, the Cold War,	Students can identify, compare, contrast, and analyze critical choices, consequences, rights, responsibilities, beliefs, ideas, diversity, continuity, change, and dynamic relationships of 20th and 21st century U.S. history. • Students can evaluate the chain of events leading to important events such as: the labor rights movement, the Great Depression, World

Grade 11 US History		
Level 2	**Level 3**	**Level 4**
and the civil rights movement. • Students can identify the causes and effects of the U.S. becoming a world power. • Students can identify the dynamic relationship between the United States' global actions and its people.	and the civil rights movement. • Students can analyze the causes and effects of the U.S. becoming a world power. • Students can analyze the dynamic relationship between the United States' global actions and its people.	Wars 1 & 2, the Cold War, and the civil rights movement. • Students can evaluate the causes and effects of the U.S. becoming a world power, including identifying how other countries interact with the United States. • Students can examine the dynamic relationship between the United States' global actions and its people.
Students can recognize aspects of the role of economics in the choices, consequences, rights, responsibilities, beliefs, ideas, diversity, continuity, change, and dynamic relationships of the 20th and 21st century United States.	Students can recognize and analyze aspects of the role of economics in the choices, consequences, rights, responsibilities, beliefs, ideas, diversity, continuity, change, and dynamic relationships of the 20th and 21st century United States.	Students can recognize, evaluate, and analyze the role of economics on the choices, consequences, rights, responsibilities, beliefs, ideas, diversity, continuity, change, and dynamic relationships of the 20th century and 21st century United States.
Students can identify the role of geography in shaping the choices, consequences, rights, responsibilities, beliefs, ideas, diversity, continuity, change, and dynamic relationships of the 20th and 21st century United States.	Students can identify and analyze the role of geography in shaping the choices, consequences, rights, responsibilities, beliefs, ideas, diversity, continuity, change, and dynamic relationships of the 20th and 21st century United States.	Students can recognize, evaluate, and analyze the role of geography on the choices, consequences, rights, responsibilities, beliefs, ideas, diversity, continuity, change, and dynamic relationships of the 20th and 21st century United States.
Students can identify the role of politics and power in shaping the choices, consequences, rights, responsibilities, beliefs, ideas, diversity, continuity, change, and dynamic relationships of the 20th and 21st century United States. • Students can recognize basic changes made to the rights and responsibilities of the	Students can identify and analyze the role of politics and power in shaping the choices, consequences, rights, responsibilities, beliefs, ideas, diversity, continuity, change, and dynamic relationships of the 20th and 21st century United States. • Students can analyze changes made to the rights and responsibilities of the	Students can recognize, evaluate, and analyze the role of politics and power in the choices, consequences, rights, responsibilities, beliefs, ideas, diversity, continuity, change, and dynamic relationships of the 20th and 21st century United States. • Students evaluate changes made to the rights and responsibilities of the

Grade 11 US History		
Level 2	**Level 3**	**Level 4**
people and the US government.	people and the US government.	people and the US government.
Students can identify connections between the choices, consequences, rights, responsibilities, beliefs, ideas, diversity, continuity, change, and dynamic relationships of the 20th century U.S. and the modern world. • Students can identify examples of continuity and change over time during the 20th and 21st centuries in the United States.	Students can identify and draw conclusions about the connections between the choices, consequences, rights, responsibilities, beliefs, ideas, diversity, continuity, change, and dynamic relationships of the 20th century U.S. and the modern world. • Students can identify examples of continuity and change over time and analyze their impact on the United States.	Students can identify, analyze, draw conclusions, and create arguments about the connections between the choices, consequences, rights, responsibilities, beliefs, ideas, diversity, continuity, change, and dynamic relationships of the 20th century U.S. and the modern world. • Students can analyze examples of continuity and change over time and evaluate their impact on the United States.
Students can identify different types of historical documents (e.g., primary vs. secondary sources).	Students can identify and analyze different types of historical documents and the role and importance of each type.	Students can identify, analyze, and evaluate different types of historical documents and the role, importance, and validity of each type.
Students can read and use information from a variety of primary and secondary sources to develop a claim or position using direct evidence and argument to support their ideas.	Students can read, analyze, and use information from a variety of primary and secondary sources to develop a claim or position using direct evidence and argument to support their ideas.	Students can read, analyze, and use information from a variety of primary and secondary sources to develop a claim or position using direct evidence and argument to support their ideas or refute the position of another agent.

Appendix G: HGSS Standard-Setting Agenda

May 24, 2015	
Time	Activity
10:00 a.m. – 11:00 a.m.	Orientation session
11:00 a.m. – 11:30 a.m.	Panelists take the HGSS tests
11:30 a.m. – 12:00 p.m.	Exercises to better understand the HGSS PLDs
12:00 p.m. – 1:00 p.m.	Lunch
1:00 p.m. – 3:30 p.m.	Exercises to better understand the HGSS PLDs—Continued
3:30 p.m. – 3:45 p.m.	Break
3:45 p.m. – 5:00 p.m.	Machine-scorable test parts activities
	Practice and readiness form
	Round 1
	Break (for data analysis)
	Round 1 feedback and Round 2
	Break (for data analysis)
	Round 2 feedback
5:00 p.m.	materials collection and dismiss for day
May 25, 2015	
8:00 a.m. – 9:00 a.m.	Review of scoring rubrics and sample MDPT responses
9:00 a.m. – 9:30 a.m.	Training on weighting HGSS test parts
9:30 a.m. – 11: 00 a.m.	Weighting HGSS test parts
11:00 a.m. – 11:15 a.m.	Break
11:15 a.m. – 11:45 a.m.	Final review of cut scores and weighting
11:45 a.m. – 12:00 p.m.	Evaluation form completion and materials collection

Appendix H: Participant Survey

HGSS Standard-Setting Participant Survey

Grade Level: _____

Directions: Please circle or write in your answers. Your responses will be aggregated in technical documents for this event. Your individual responses will not be reported or linked to you.

1. Gender: Female Male

2. Ethnicity:

White	Hispanic or Latino	Black or African America
Asian/Pacific Islander	Native American or American Indian	Other

3. Current Assignment: Classroom Teacher Educator (Non-Teacher) Other

a) If you are a Classroom Teacher, do you teach special-needs students? Yes No

b) If you are a Classroom Teacher, do you teach ELL students? Yes No

c) If "Other" please provide additional information below (occupation, educational focus, etc.)

4. Are you familiar with the Kansas HGSS Standards? Yes No

5. Work Setting: Urban Suburban Rural

6. District Name: _____

7. How many years (total) have you been teaching? _____

8. Please list the grades and the number of years you taught HGSS at each grade.

9. Please describe your professional development activities in HGSS within the past two years:
(Please use the back side if necessary)

Appendix I: Nondisclosure Agreement

Under this agreement, you will have access to secure and confidential materials and data belonging to the Kansas Assessment Program (KAP), a program of the Achievement and Assessment Institute (CETE) at the University of Kansas. These materials, documents, data, and other information are privileged and confidential and may not be used, shared, discussed, or otherwise published with any person who has not signed this confidentiality agreement without the expressed written consent of the Director of CETE.

These data or materials may not be copied, published, announced, or in any other way made public. Any person who knowingly discloses, publishes, or uses KAP tools, resources or materials without prior authorization or approval may be subject to legal action. This Confidentiality Agreement will be enforced by the KU Center for Research acting as fiscal agent on behalf of KAP. The Agreement will be construed under the laws of the state of Kansas and the venue for enforcement will be the Douglas County Kansas District Court.

By signing this agreement, you acknowledge that the KAP materials and data constitute confidential materials of CETE. You further understand that any disclosure, unauthorized use, or reproduction of these materials would damage the confidentiality of KAP, is illegal, and can result in legal action taken against the signatory. You agree to keep any such materials and data, as identified by KAP, secure and confidential.

ACCEPTED AND AGREED TO:

Signature_____ Date _____

Full name (please print) _____

Title _____

Organization Name _____

Address _____

Appendix J: Multidisciplinary Performance Task Rubrics

Grades 6–8

In the response, the student demonstrates:	
4	Consistent and effective command of the skills needed to complete an on-demand.
3	Mostly consistent and adequate command of the skills needed to complete an on-demand writing task.
2	Somewhat consistent and minimal command of the skills needed to complete an on-demand writing task.
1	Inconsistent and ineffective command of the skills needed to complete an on-demand writing task.

Nonscorable Codes

BL	Blank	The response is blank.
IN	Insufficient	The response does not include enough student writing to score.
OT	Off Task	Insufficient OT Off Task The response is unrelated to the resources and/or prompt.
OL	Other Language	The response is in a language other than English.

High School Rubric

In the response, the student demonstrates:	
4	Consistent and effective command of the skills needed to complete an on-demand.
3	Mostly consistent and adequate command of the skills needed to complete an on-demand writing task.
2	Somewhat consistent and minimal command of the skills needed to complete an on-demand writing task.
1	Inconsistent and ineffective command of the skills needed to complete an on-demand writing task.

Nonscorable Codes

BL	Blank	The response is blank.
IN	Insufficient	The response does not include enough student writing to score.
OT	Off Task	Insufficient OT Off Task The response is unrelated to the resources and/or prompt.
OL	Other Language	The response is in a language other than English.

Appendix K: Readiness Form

READINESS FORM

Panelist ID: _____ Grade: _____
Table #: _____

PLEASE COMPLETE THE FOLLOWING QUESTION(S)

1. I have completed the orientation and training, and I understand the purpose of this part of the standard-setting event. I also clearly understand my role in this event and what I am being asked to do next. I am ready to begin this round.

   YES _____ NO _____ Your Initials _____

   If you answered ***NO***, please raise your hand and ask the facilitator for additional help.

NOTE: ANSWER QUESTION 2 ONLY IF YOU ANSWERED "NO" TO QUESTION 1.

2. I have received additional help and training. I now clearly understand my role and the task that I am being asked to do next. I am now ready to begin this round.

   YES _____ NO _____ Your Initials _____

Appendix L: Evaluation Form

Final Evaluation Form

Your opinions will provide us with a basis for evaluating both the materials and the training. **DO NOT** put your name on this form. We want your opinions to remain anonymous.

GRADE: _____

I. Opening Session *Indicate the extent to which you agree or disagree with the following statements.*	Strongly Disagree	Disagree	Somewhat Disagree	Somewhat Agree	Agree	Strongly Agree
a. The opening session provided adequate background about the assessment program.						
b. The opening session provided a clear understanding of the purpose of the meeting.						
c. The opening session provided an appropriate context for my role in the meeting.						
d. The opening session addressed many of my questions and concerns.						
e. The opening session was well organized.						
f. The opening session presenter clearly explained the procedures.						

II. Opening Session Time *Indicate how well the orientation time matched your need for the information presented.*	Too Little	About Right	Too Much
a. The amount of time used for the opening session.			

III. Taking the Test *Indicate the extent to which you agree or disagree with the following statements.*	Strongly Disagree	Disagree	Somewhat Disagree	Somewhat Agree	Agree	Strongly Agree
a. Taking the test helped me understand the assessment.						

IV. Performance Level Descriptors (PLDs) *Indicate the degree to which you understood each PLD level.*	NO UNDERSTANDING	SLIGHT UNDERSTANDING	MODERATE UNDERSTANDING	COMPLETE UNDERSTANDING
Level 2				
Level 3				
Level 4				

V. Low, Middle, High Student Activity *Indicate the degree to which you understood the Low, Middle, and High student activity.*	NO UNDERSTANDING	SLIGHT UNDERSTANDING	MODERATE UNDERSTANDING	COMPLETE UNDERSTANDING
Level 2 Low, Middle, and High				
Level 3 Low, Middle, and High				
Level 4 Low, Middle, and High				

VI. Practice *Indicate the extent to which you agree or disagree with the following statements.*	STRONGLY DISAGREE	DISAGREE	SOMEWHAT DISAGREE	SOMEWHAT AGREE	AGREE	STRONGLY AGREE
a. The practice activity helped me understand my task.						
b. After the practice, I understood my role in the event.						
c. The event staff effectively answered any questions that were asked during practice.						

Use this space for additional comments you wish to share regarding the quality of the opening session, PLD activities, and practice.

VII. Round 1 Experience *Indicate the extent to which you agree or disagree with the following statements.*	Strongly Disagree	Disagree	Somewhat Disagree	Somewhat Agree	Agree	Strongly Agree
a. The rating form was easy to understand.						
b. The expectations for Round 1 were made clear.						
c. I made my ratings independently.						
d. I understood the tasks I was to accomplish for Round 1.						
e. I had the right amount of time to complete the tasks during Round 1.						

VIII. Influential Factors for Round 1 *Indicate how important each of the following elements were as you completed your rating sheets for **Round 1**.*	Not Important	Slightly Important	Moderately Important	Very Important
a. PLDs for Level 2, Level 3, and Level 4				
b. Your perceptions of the Low, Middle, and High students				
c. Your perceptions of the difficulty of the items				
d. Your experience with students at your grade level				
e. Your experience with the Kansas HGSS Standards				

IX. Round 1 Results *Indicate the extent to which you agree or disagree with the following statements.*	Strongly Disagree	Disagree	Somewhat Disagree	Somewhat Agree	Agree	Strongly Agree
a. The **Round 1** results were **clear**.						
b. The **Round 1** results were **useful**.						

X. Round 2 Experience *Indicate the extent to which you agree or disagree with the following statements.*	Strongly Disagree	Disagree	Somewhat Disagree	Somewhat Agree	Agree	Strongly Agree
a. The expectations for Round 2 were made clear.						
b. I made my ratings independently.						
c. I understood the tasks I was to accomplish for Round 2.						
d. I had the right amount of time to complete the tasks during Round 2.						

XI. Influential Factors for Round 2 *Indicate how important each of the following elements were as you completed your rating sheets for **Round 2**.*	Not Important	Slightly Important	Moderately Important	Very Important
a. PLDs for Level 2, Level 3, and Level 4				
b. Your perceptions of the Low, Middle, and High students				
c. Your perceptions of the difficulty of the items				
d. Your experience with students at your grade level				
e. Your experience with the Kansas HGSS Standards				
f. Your Round 1 results				
g. The Round 1 results of the other panelists				
h. Group discussions about the Round 1 Results				

XII. Round 2 Results *Indicate the extent to which you agree or disagree with the following statements.*	Strongly Disagree	Disagree	Somewhat Disagree	Somewhat Agree	Agree	Strongly Agree
a. The **Round 2** results were **clear**.						
b. The **Round 2** results were **useful**.						

XIII. Group Discussion *Indicate the extent to which you agree or disagree with the following statements.*	Strongly Disagree	Disagree	Somewhat Disagree	Somewhat Agree	Agree	Strongly Agree
a. The group discussions aided my understanding of the issues.						
b. The time provided for discussions was adequate.						
c. Everyone had equal opportunity to contribute ideas and opinions.						
d. The discussions about the Performance Level Descriptors (PLDs) were helpful to me.						
e. The discussions about the Low, Middle, and High student were helpful to me.						

XIV. MDPT *Indicate the extent to which you agree or disagree with the following statements.*	Strongly Disagree	Disagree	Somewhat Disagree	Somewhat Agree	Agree	Strongly Agree
a. The scoring rubric helped me understand the role of the MDPT in the HGSS assessment.						
b. The sample responses helped me understand the role of the MDPT in the HGSS assessment.						

XV. Test Part Weights *Indicate the extent to which you agree or disagree with the following statements.*	Strongly Disagree	Disagree	Somewhat Disagree	Somewhat Agree	Agree	Strongly Agree
a. The final weight for the MDPT part is reasonable.						
b. The final weight for the first machine scorable Part 1 is reasonable.						
c. The final weight for the second machine scorable Part 2 is reasonable.						

XVI. Results for Level 2 Cut Score *Indicate the extent to which you agree or disagree with the following statements.*	Strongly Disagree	Disagree	Somewhat Disagree	Somewhat Agree	Agree	Strongly Agree
a. The cut score for this achievement level is **appropriate** based on the PLDs and the Low, Middle, and High student activities.						
b. The cut score for this achievement level is **defensible** due to panelists' adherence to procedures.						

XVII. Results for Level 3 Cut Score *Indicate the extent to which you agree or disagree with the following statements.*	Strongly Disagree	Disagree	Somewhat Disagree	Somewhat Agree	Agree	Strongly Agree
a. The cut score for this achievement level is **appropriate** based on the PLDs and the Low, Middle, and High student activities.						
b. The cut score for this achievement level is **defensible** due to panelists' adherence to procedures.						

XVIII. Results for Level 4 Cut Score *Indicate the extent to which you agree or disagree with the following statements.*	Strongly Disagree	Disagree	Somewhat Disagree	Somewhat Agree	Agree	Strongly Agree
a. The cut score for this achievement level is **appropriate** based on the PLDs and the Low, Middle, and High student activities.						
b. The cut score for this achievement level is **defensible** due to panelists' adherence to procedures.						

If you did not **agree** or did not **strongly agree** with the **appropriateness** or **defensibility** of any of the cut scores, please explain what adjustments you would make to improve the appropriateness or defensibility of the cut scores.

XIX. Successfulness of Agenda Items *Indicate how successful you believe each task or event was in the standard-setting process.*	Not Successful	Slightly Successful	Moderately Successful	Very Successful
a. Opening session				
b. Taking the operational test				
c. Review of the Performance Level Descriptors				
d. Discussions about the Low, Middle, and High student groups				
e. Practice activities				
f. Discussions after Round 1				

XX. CETE Staff *Indicate how helpful you felt each staff member was during the standard-setting process.*	Not Helpful	Slightly Helpful	Moderately Helpful	Very Helpful
a. Orientation Leader				
b. Facilitator				
c. Table Leaders				
d. Content Specialists				

XX. CETE STAFF *Indicate how helpful you felt each staff member was during the standard-setting process.*	NOT HELPFUL	SLIGHTLY HELPFUL	MODERATELY HELPFUL	VERY HELPFUL
e. Other staff (please specify here): _____				
_____				

Thank you for your participation in this event. Please use this space for any additional comments about this event that you wish to share.

Appendix M: Score Reports

Student Report

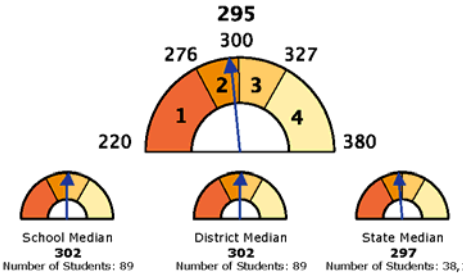
STUDENT REPORT: Last, First
GRADE: 3 English Language Arts / STATE ID: 0000000000
SCHOOL: School Name
DISTRICT: Name / #D0000

2015—2016
KANSAS ASSESSMENT PROGRAM

The KAP assessments measure students' understanding of the Kansas College and Career Ready Standards at each grade. The English language arts questions ask students to select the right answer, organize information, and respond to a writing prompt. The overall score combines a score from the reading, writing, and listening section and a score from the on-demand writing task section.

Overall English Language Arts Score: Level 3

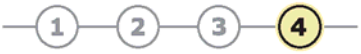
Reading, Writing, and Listening Score



In the reading, writing, and listening section, your student is performing at Level 2. Questions in this section ask students to select the right answer and to sort, match, label, and order items.

Students who score at this level can typically
- read and understand readily accessible grade-level texts
- identify main ideas and clear details
- determine meanings of common words and straightforward figurative language
- identify text features and structures that organize a text
- identify relationships between parts of a text
- revise or edit a text to use appropriate general language, grammar, punctuation, and spelling
- use writing strategies appropriate for different types of text
- structure a text to support a purpose or opinion

On-Demand Writing Task Score



Your student's on-demand writing task score is a 4. On-demand writing tasks ask students to engage with texts and other resources and then compose a related writing sample.

Students who perform at this level can typically
- maintain a setting and characters in narrative writing or maintain a clear main idea in informational and opinion writing
- use significant descriptions and supporting evidence and details
- clarify relationships and use appropriate transitions
- follow complex grammar, punctuation, and spelling rules

Performance Level Descriptions *(applies to all scores)*

Level 1: A student at Level 1 shows a limited ability to understand and use the reading, writing, and listening skills and knowledge needed for college and career readiness.

Level 2: A student at Level 2 shows a basic ability to understand and use the reading, writing, and listening skills and knowledge needed for college and career readiness.

Level 3: A student at Level 3 shows an effective ability to understand and use the reading, writing, and listening skills and knowledge needed for college and career readiness.

Level 4: A student at Level 4 shows an excellent ability to understand and use the reading, writing, and listening skills and knowledge needed for college and career readiness.

For more details about how your student performed on specific types of test questions, see the back of this report. →

STUDENT REPORT

STUDENT: Last, First
STATE ID: 0000000000

GRADE: 3 English Language Arts

Your Student's Performance

Exceeds | Meets | Below | Insufficient Data

OVERALL READING

In this area, your student performed below students who received the minimum Level 3 score. The reading portion requires students to read and analyze literary and informational texts and answer questions related to main ideas, text structure, language use, word meanings, and making and supporting conclusions.

READING: Literary Texts

In this area, your student performed as well as students who received the minimum Level 3 score. This portion requires students to answer questions based on literary texts (such as stories and poems).

READING: Informational Texts

In this area, your student performed below students who received the minimum Level 3 score. This portion requires students to answer questions based on informational texts (such as science articles and historical speeches).

READING: Making and Supporting Conclusions

In this area, your student performed below students who received the minimum Level 3 score. These questions require students to read literary and informational texts and then make conclusions and use details and evidence to support ideas.

READING: Main Idea

In this area, your student performed below students who received the minimum Level 3 score. These questions require students to read literary and informational texts and then determine central ideas, key events, and topics and identify supporting details.

OVERALL WRITING

In this area, your student performed below students who received the minimum Level 3 score. The writing portion requires students to read short writing samples and answer questions related to revising, editing, vocabulary, and language use. (This portion does not include the on-demand writing task.)

WRITING: Revising

In this area, your student performed below students who received the minimum Level 3 score. These questions require students to revise provided text by applying writing skills, including using specific story-telling strategies, revising text into a logical order, adding context and detail, and identifying words or phrases to strengthen the text.

WRITING: Editing

In this area, your student performed better than students who received the minimum Level 3 score. These questions require students to clarify messages in a variety of texts by following grade-appropriate grammar, capitalization, punctuation, and spelling rules.

WRITING: Vocabulary and Language Use

In this area, your student performed below students who received the minimum Level 3 score. These questions require students to revise texts by using accurate language and vocabulary that is appropriate to a text's purpose and audience.

OVERALL LISTENING

In this area, your student performed below students who received the minimum Level 3 score. The listening portion requires students to listen to a recording and show understanding by interpreting the speaker's point of view, identifying central ideas and supporting evidence, and making conclusions.

Standard error of measurement for this report:

Student — 8.1 | School — 3.6 | District — 3.6 | State — 0.2

The standard error indicates how much a student's score might vary if the student took many equivalent versions of the test (tests with different items but covering the same knowledge and skills).

Additional Resources

For sample test questions, go to **ksassessments.org/practice-tests**
For information on the Kansas College and Career Ready Standards, visit **ksde.org**
To learn about the Kansas Assessment Program, go to **ksassessments.org**
To discover more about this score report, see the 2016 Parent Guide at **kap.cete.us/pg**

Kansas State Department of Education

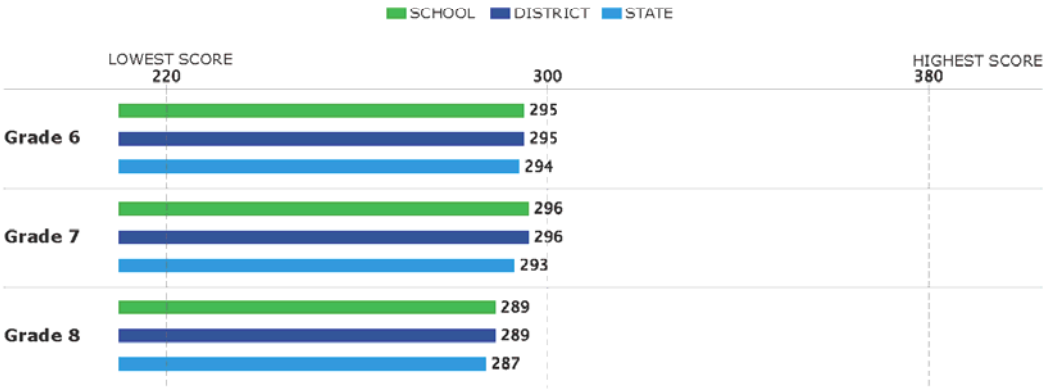
School Report

SCHOOL REPORT: School Name / #0000
SUBJECT: English Language Arts
DISTRICT: District Name / #D0000

2015—2016
KANSAS ASSESSMENT PROGRAM

The KAP assessments measure students' understanding of the Kansas College and Career Ready Standards at each grade. The English language arts questions ask students to select the right answer, organize information, and respond to a writing prompt. In grades 3–8, a student's overall score combines a score from the reading, writing, and listening section and a score from the on-demand writing task section. In grade 10, students took an on-demand writing task field test and did not receive a writing task score.

Reading, Writing, and Listening Scores: Median School, District, and State Performance

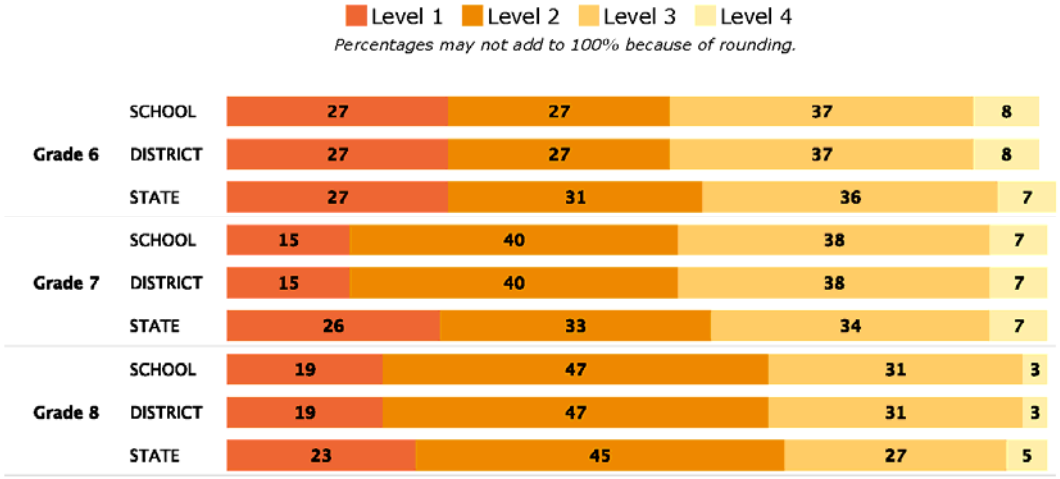


Standard error of measurement for this report:

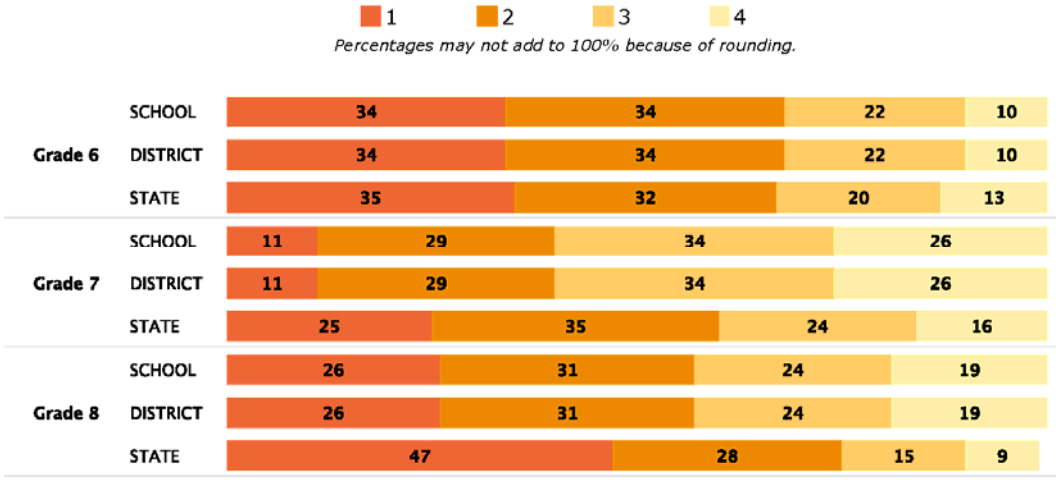
Grade 6: School—3.0 | District—3.0 | State—0.2
Grade 7: School—2.7 | District—2.7 | State—0.2
Grade 8: School—3.0 | District—3.0 | State—0.2

The standard error indicates how much students' scores might vary if the students took many equivalent versions of the test (tests with different items but covering the same knowledge and skills).

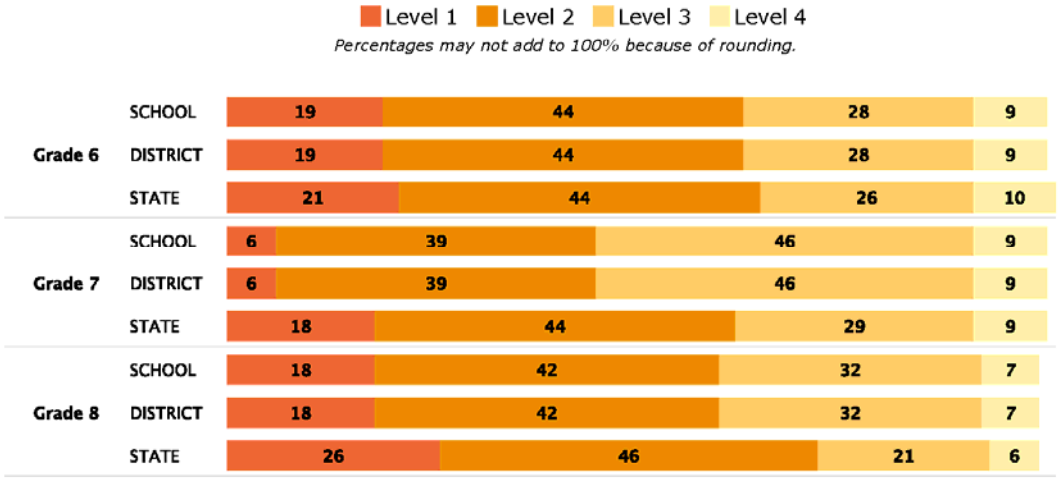
Reading, Writing, and Listening Scores: Percentage of Students in Each Performance Level



On-Demand Writing Task Scores: Percentage by Grade



Overall Scores: Percentage of Students in Each Performance Level, by Grade



Your School's Performance

Exceeds Meets Below Insufficient Data

Grade	6	7	8
OVERALL READING	Below	Below	Below
Literary Texts	Below	Below	Below
Informational Texts	Below	Below	Below
Making and Supporting Conclusions	Below	Below	Below
Main Idea	Below	Below	Below
OVERALL WRITING	Below	Below	Below
Revising	Below	Below	Below
Editing	Meets	Meets	Below
Vocabulary and Language Use	Meets	Below	Below
OVERALL LISTENING	Meets	Meets	Below

OVERALL READING
The reading portion requires students to read and analyze literary and informational texts and answer questions related to main ideas, text structure, language use, word meanings, and making and supporting conclusions.

Literary Texts
This portion requires students to answer questions based on literary texts (such as stories and poems).

Informational Texts
This portion requires students to answer questions based on informational texts (such as science articles and historical speeches).

Making and Supporting Conclusions
These questions require students to read literary and informational texts and then make conclusions and use details and evidence to support ideas.

Main Idea
These questions require students to read literary and informational texts and then determine central ideas, key events, and topics and identify supporting details.

OVERALL WRITING
The writing portion requires students to read short writing samples and answer questions related to revising, editing, vocabulary, and language use. (This portion does not include the on-demand writing task.)

Revising
These questions require students to revise provided text by applying writing skills, including using specific story-telling strategies, revising text into a logical order, adding context and detail, and identifying words or phrases to strengthen the text.

Editing
These questions require students to clarify messages in a variety of texts by following grade-appropriate grammar, capitalization, punctuation, and spelling rules.

Vocabulary and Language Use
These questions require students to revise texts by using accurate language and vocabulary that is appropriate to a text's purpose and audience.

SCHOOL REPORT SCHOOL: SCHOOL NAME

OVERALL LISTENING
The listening portion requires students to listen to a recording and show understanding by interpreting the speaker's point of view, identifying central ideas and supporting evidence, and making conclusions.

Your School's Performance

Exceeds
In this area, your students typically performed better than students who received the minimum Level 3 score.

Meets
In this area, your students typically performed as well as students who received the minimum Level 3 score.

Below
In this area, your students typically performed below students who received the minimum Level 3 score.

Insufficient Data
In this area, your students did not answer enough questions for accurate reporting.

Additional Resources
For sample test questions, go to **ksassessments.org/practice-tests**
For information on the Kansas College and Career Ready Standards, visit **ksde.org**
To learn about the Kansas Assessment Program, go to **ksassessments.org**
To discover more about this score report, see the 2016 Educator Guide at **kap.cete.us/eg**

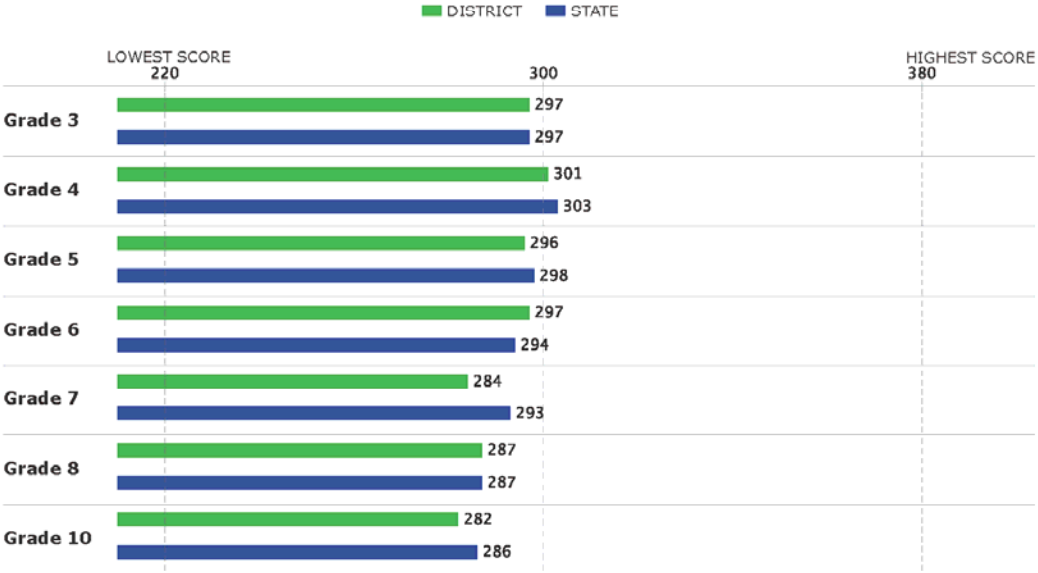
District Report

DISTRICT REPORT: Name / #D0000
SUBJECT: English Language Arts

2015—2016
KANSAS ASSESSMENT PROGRAM

The KAP assessments measure students' understanding of the Kansas College and Career Ready Standards at each grade. The English language arts questions ask students to select the right answer, organize information, and respond to a writing prompt. In grades 3–8, a student's overall score combines a score from the reading, writing, and listening section and a score from the on-demand writing task section. In grade 10, students took an on-demand writing task field test and did not receive a writing task score.

Reading, Writing, and Listening Scores: Median District and State Performance

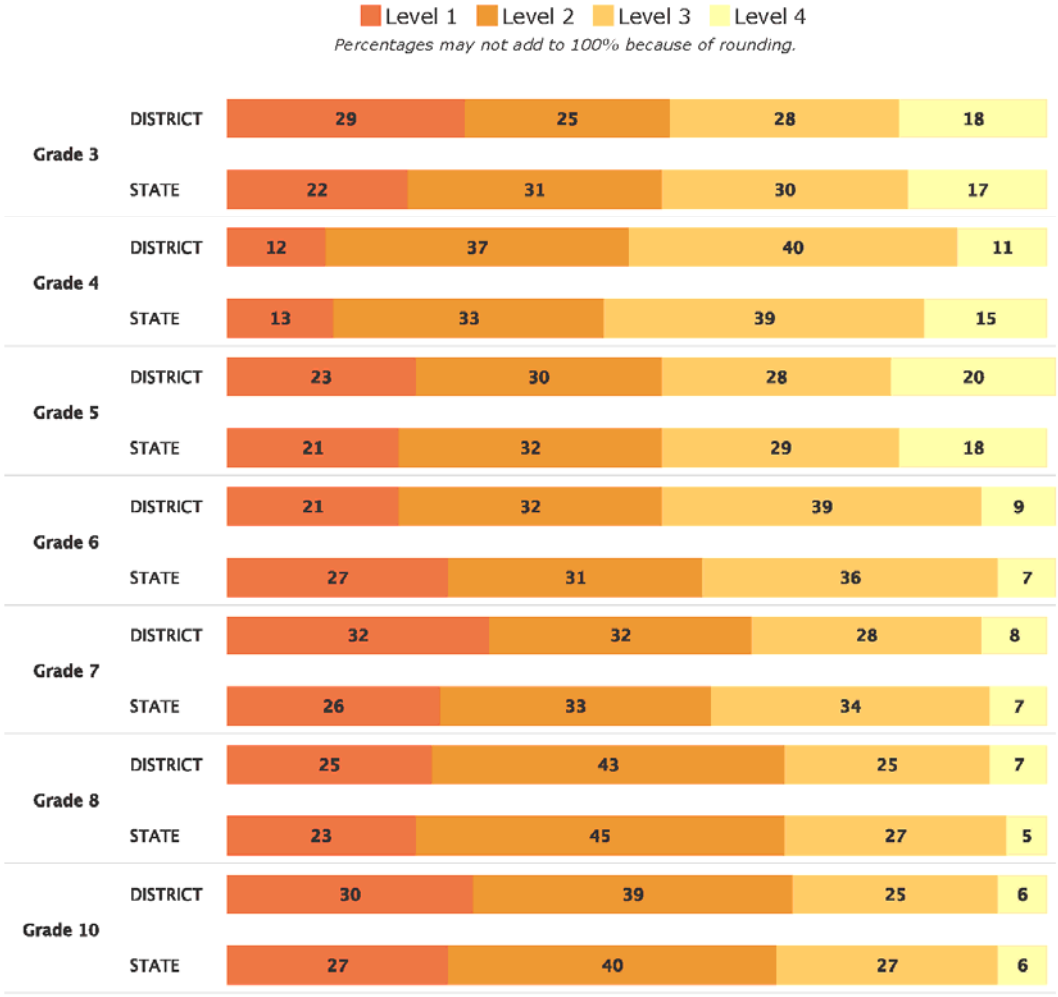


Standard error of measurement for this report:

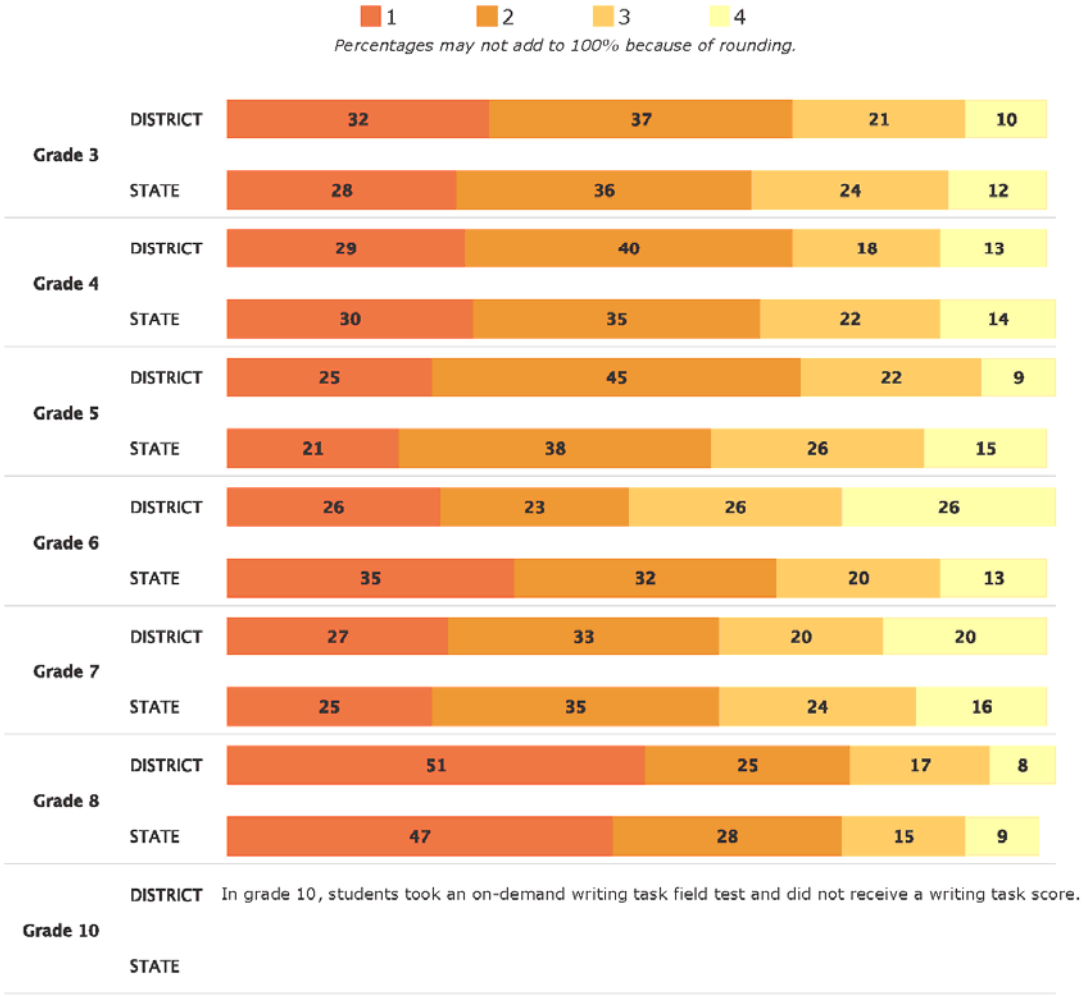
Grade 3: District—2.4 | State—0.2
Grade 4: District—2.4 | State—0.2
Grade 5: District—2.6 | State—0.2
Grade 6: District—2.4 | State—0.2
Grade 7: District—2.5 | State—0.2
Grade 8: District—2.6 | State—0.2
Grade 10: District—2.7 | State—0.2

The standard error indicates how much students' scores might vary if the students took many equivalent versions of the test (tests with different items but covering the same knowledge and skills).

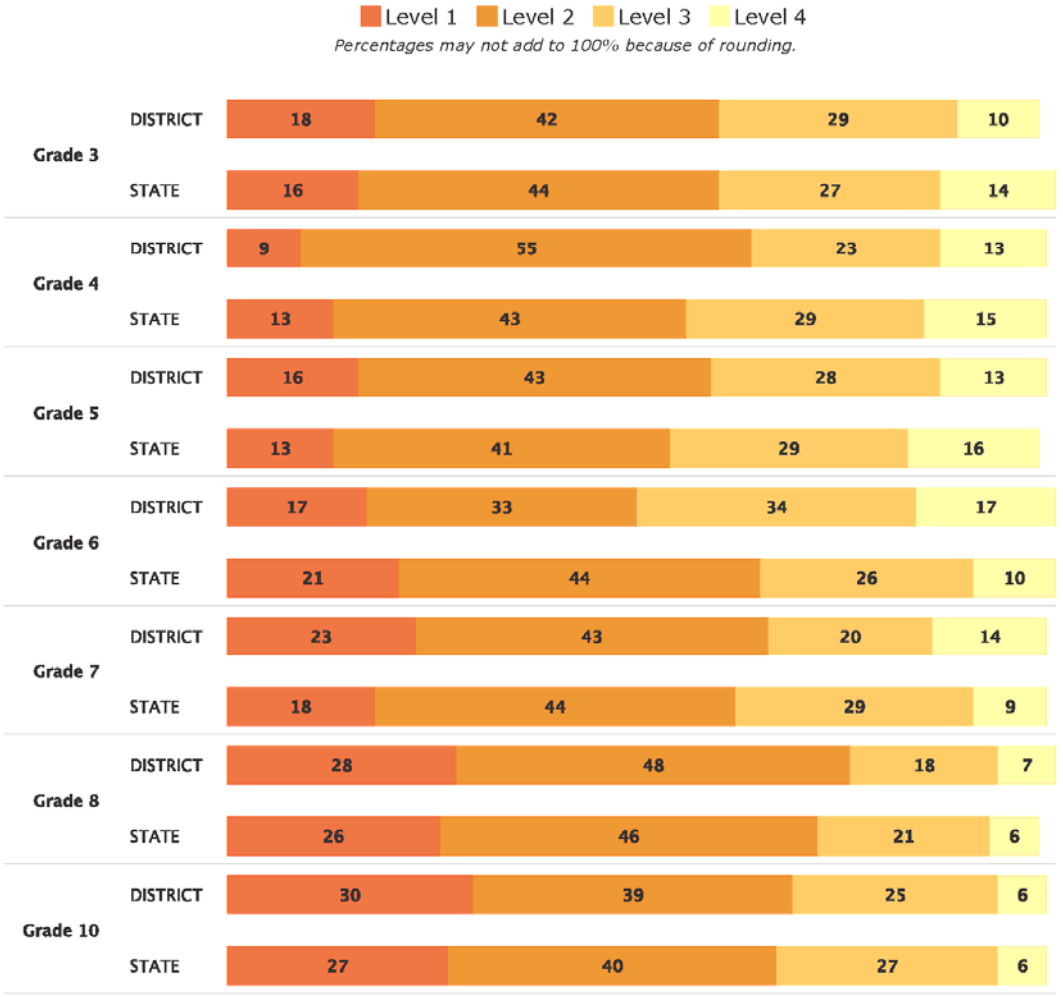
Reading, Writing, and Listening Scores: Percentage of Students in Each Performance Level



On-Demand Writing Task Scores: Percentage by Grade



Overall Scores: Percentage of Students in Each Performance Level, by Grade



Your District's Performance

Exceeds | Meets | Below | Insufficient Data

Grade	3	4	5	6	7	8	10
OVERALL READING	Below	Meets	Below	Below	Below	Below	Below
Literary Texts	Meets	Meets	Below	Meets	Below	Below	Below
Informational Texts	Below	Meets	Below	Below	Below	Below	Below
Making and Supporting Conclusions	Below	Meets	Meets	Below	Below	Below	Below
Main Idea	Below	Below	Below	Meets	Below	Below	Below
OVERALL WRITING	Below	Below	Meets	Meets	Below	Below	Below
Revising	Below	Meets	Below	Meets	Below	Below	Below
Editing	Below	Below	Meets	Meets	Below	Below	Below
Vocabulary and Language Use	Below	Below	Meets	Meets	Below	Below	Below
OVERALL LISTENING	Meets	Meets	Meets	Meets	Meets	Below	Meets

OVERALL READING
The reading portion requires students to read and analyze literary and informational texts and answer questions related to main ideas, text structure, language use, word meanings, and making and supporting conclusions.

Literary Texts
This portion requires students to answer questions based on literary texts (such as stories and poems).

Informational Texts
This portion requires students to answer questions based on informational texts (such as science articles and historical speeches).

Making and Supporting Conclusions
These questions require students to read literary and informational texts and then make conclusions and use details and evidence to support ideas.

Main Idea
These questions require students to read literary and informational texts and then determine central ideas, key events, and topics and identify supporting details.

OVERALL WRITING
The writing portion requires students to read short writing samples and answer questions related to revising, editing, vocabulary, and language use. (This portion does not include the on-demand writing task.)

Revising
These questions require students to revise provided text by applying writing skills, including using specific story-telling strategies, revising text into a logical order, adding context and detail, and identifying words or phrases to strengthen the text.

Editing
These questions require students to clarify messages in a variety of texts by following grade-appropriate grammar, capitalization, punctuation, and spelling rules.

Vocabulary and Language Use
These questions require students to revise texts by using accurate language and vocabulary that is appropriate to a text's purpose and audience.

OVERALL LISTENING
The listening portion requires students to listen to a recording and show understanding by interpreting the speaker's point of view, identifying central ideas and supporting evidence, and making conclusions.

Your District's Performance

Exceeds
In this area, your students typically performed better than students who received the minimum Level 3 score.

Meets
In this area, your students typically performed as well as students who received the minimum Level 3 score.

Below
In this area, your students typically performed below students who received the minimum Level 3 score.

Insufficient Data
In this area, your students did not answer enough questions for accurate reporting.

Additional Resources

For sample test questions, go to **ksassessments.org/practice-tests**
For information on the Kansas College and Career Ready Standards, visit **ksde.org**
To learn about the Kansas Assessment Program, go to **ksassessments.org**
To discover more about this score report, see the 2016 Educator Guide at **kap.cete.us/eg**

Appendix N: Letters from the Commissioner of Education

KANSAS ASSESSMENT PROGRAM

2016 Educator Guide

Understanding the Kansas Assessment Program Score Report

Dear Educators:

Thank you for your participation in the 2016 Kansas Assessment Program.

While assessments are an important tool that can help gauge a student's progress, we recognize they are just one of several measures to consider. Your use of classroom interaction, homework, assessments and other strategies throughout the year are equally important to the process of identifying learning and achievement levels.

The Kansas State Board of Education's new vision for education — Kansas leads the world in the success of each student — reduces what many have considered an overemphasis on state assessments and increases the focus on the needs of the whole child. As we work toward this vision, you will see an increased focus on areas such as kindergarten readiness, Individual Plans of Study focused on career interest, high school graduation rates, postsecondary completion, and social/emotional growth.

Kansas state department of Education

Assessments will continue to serve a role in helping to determine your students' academic readiness, but the State Board and the Kansas State Department of Education think it is time to shrink the assessment footprint on Kansas. We want the goals of each student — from the 5-year-old kindergarten student all the way to the high school graduate considering a career, college or the military — to be important.

Kansas' teachers, students and parents are among the best in the nation, and we all share in the responsibility of making every child successful by achieving their desired future.

Thank you for all of your hard work and commitment to ensuring each student in Kansas is prepared for future success.

Sincerely,

Dr. Randy Watson
Kansas Commissioner of Education

KANSAS ASSESSMENT PROGRAM

2016 Parent Guide

Understanding the Kansas Assessment Program Score Report

Dear Parents:

Thank you for supporting your child's participation in the 2016 Kansas Assessment Program.

While assessments are an important tool to help teachers, parents, and students gauge a student's progress, it is essential to remember they are just one of several measures teachers consider. Your student's teachers use classroom interaction, homework, assessments and many other strategies throughout the year to identify learning and achievement levels.

The Kansas State Board of Education's new vision for education—Kansas leads the world in the success of each student—reduces what many have considered an overemphasis on state assessments and increases focus on the needs of the whole child. As we work toward this vision, you will see schools focus on areas such as kindergarten readiness, Individual Plans of Study focused on career interest, high school graduation rates, postsecondary completion, and social/emotional growth.

Assessments will continue to serve a role in helping to inform about your child's academic readiness, but the State Board and the Kansas State Department of Education believe it is time to minimize the assessment footprint on Kansas. We want the goals of each student—from the 5-year-old kindergarten student all the way to the high school graduate focusing on a career, college, or the military—to be important.

As you review your child's report, please take the opportunity to contact your child's school, teacher or principal to have them explain these results to you in detail.

Kansas students, teachers and parents are among the best in the nation, and we all share in the responsibility of making every child successful by achieving their desired future.

Thank you for being a positive part of your child's education, and thank you for your continued support of Kansas schools.

Sincerely,

Dr. Randy Watson
Kansas Commissioner of Education