Kentucky Summative Assessments

2022–2023 Technical Manual

Pearson

KSA 2022–2023

The *KSA Technical Manual* contains general information on the development, scoring, and analysis of the KSA assessments. The accompanying *Yearbook* contains test performance results in the form of performance statistics and test measurement characteristics to supplement the contents of the technical manual.

Table of Contents

List of Tables

List of Figures

1. Background

Over the last 30 years, Kentucky's assessment program has evolved to such an extent that it is now one of the country's leading assessment programs in preparing students for future success. The assessment program has used resources within Kentucky and external sources to build a system that measures student achievement to both state and national standards. Over the course of its evolution, the Kentucky assessment program has included various forms of assessment components, including brief constructed responses, essays, performance tasks, and portfolios in addition to the conventional multiple-choice items. A major contribution to the maintenance of the assessment program has been through various professional organizations and stakeholder groups within and outside of the Commonwealth of Kentucky. These groups have provided invaluable expertise and feedback on all aspects of the assessment program, from test development to score reporting; they continue to make significant contributions today. This chapter provides a history of the Kentucky assessment program and the contributors who have guided its progression.

1.1. History

1.1.1. Kentucky Instructional Results Information System (1992–1998)

The Kentucky Instructional Results Information System (KIRIS)—used in grades 4, 5, 7, 8, 11, and 12—measured students' knowledge and their application of knowledge through a variety of performance components: essay questions (varying in response length), performance tasks, portfolios, and multiple-choice items. KIRIS covered Reading, Mathematics, Science, Social Studies, and Writing, as well as Arts/Humanities and Practical Living/Vocational Studies. The cornerstone of KIRIS was students demonstrating their understanding of concepts by being required to provide justifications for the responses they provided. The various test item types were administered in three distinct assessment components: a traditional assessment (multiple-choice and open-ended items), a performance event (performance task involving individual and group problem-solving skills), and a portfolio assessment (student-chosen collection of work). Student performance within KIRIS was divided into four achievement categories: *Novice*, *Apprentice, Proficient*, and *Distinguished*.

1.1.2. Commonwealth Accountability Testing System (1998–2010)

Beginning in 1999, the subject areas assessed under KIRIS were carried forward into a new assessment program that blended state- and national-level standards testing. The Commonwealth Accountability Testing System (CATS) consisted of two types of assessments: the Kentucky Core Content Test (KCCT) and the Comprehensive Test of Basic Skills, Fifth Edition (CTBS/5). KCCT, the criterion-referenced portion, was administered to students in grades 4, 5, 7, 8, 10, 11, and 12. For grades 4, 7, and 12, students took part in a writing assessment and created writing portfolios of their best writings produced over time. Student performance on KCCT was divided into the same achievement categories used for KIRIS, but *Novice* and *Apprentice* performance were further divided into low, medium, and high classifications for Reading, Mathematics, Science, and Social Studies. CTBS/5, a nationally norm-referenced assessment, was administered to students in grades 3, 6, and 9 in Reading, Language Arts, and Mathematics.

1.1.3. Unbridled Learning (2010–2016)

In 2009, Kentucky's General Assembly passed Senate Bill 1 that began a reform initiative on the state's accountability system that included new dimensions of student achievement. By 2011, this initiative resulted in the creation of the Unbridled Learning Accountability model that incorporated four strategic priorities for advancing the achievement of Kentucky students: next-generation learners, next-generation professionals, next-generation support systems, and next-generation schools and districts. The aim of this model was college and career readiness for all Kentucky students, which had been defined by the goals put forth by the Partnership for Assessment of Readiness for College and Careers (PARCC) national assessment consortium. In addition to measures of college and career readiness for Kentucky's next generation learners, the new accountability model factors student achievement growth measures and high school graduation rates.

The Unbridled Learning model of accountability covered student achievement on

- Reading, Mathematics, Science, and Social Studies in elementary and middle school grades;
- writing in elementary, middle school, and high school grades; and
- end-of-course tests for high school grades.[1]

The Kentucky Academic Standards (KAS) were adopted to outline the minimum content required for all students before graduating from high school. For Reading, Mathematics and Writing, the content standards were adopted from the Common Core State Standards (CCSS), sponsored by the National Governors Association (NGA) and the Council of Chief State School Officers (CCSSO), while the standards for Science and Social Studies remained from the previous curriculum standards framework.

The Kentucky Performance Rating for Educational Progress (K-PREP) was the collection of tests created and administered to assess the KAS. From 2012 to 2017, K-PREP was a blend of norm-referenced and criterion-referenced test content that provided achievement indices at the state and national levels. The criterion-referenced test portion of K-PREP was built using test content written specifically for Kentucky's assessment, and student performance was divided into the four performance levels used in the previous testing systems: *Novice*, *Apprentice*, *Proficient*, and *Distinguished*. In contrast, the norm-referenced portion consisted of test content from the Stanford Achievement Test Series, Tenth Edition (hereafter Stanford 10) using existing score norms to report Kentucky student achievement on a national scale. Beginning in 2018, Stanford 10 was no longer a component of the K-PREP assessments.

1.1.4. Kentucky's Transition to ESSA (2017–2021)

As Kentuckians engaged in the development of a new accountability system under the Every Student Succeeds Act of 2015 (ESSA) and Senate Bill 1 (2017), the Kentucky Board of Education (KBE) revised its vision and the Kentucky Department of Education (KDE) simultaneously engaged in a comprehensive strategic planning process designed to bring the department's work into alignment with ESSA and new state laws.

[1] Algebra II, English II, Biology, and U.S. History end-of-course exams were implemented in 2011–2012.

The following provided coherence with the state's accountability system: (a) the board's vision that every student is empowered and equipped with the knowledge, skills, and dispositions to pursue a successful future; (b) the department's mission to partner with districts (in the accountability regulation, 703 KAR 5:270), schools, and education stakeholders to indicate the desire for people to invest themselves in students' futures to provide service, support, and leadership to ensure success for every student; and (c) the department's underlying values of equity, achievement, and integrity.

Under ESSA and Senate Bill 1, Kentucky is required to meaningfully differentiate between schools through its accountability system to identify schools each year that need help in improving overall student outcomes or the outcomes of one or more specific group(s) of students. In February 2018, the board approved a new accountability system to be implemented beginning with the 2018–2019 school year, making the 2017–2018 school year a transition year.

In 2020–2021, Kentucky public school students completed the K-PREP Reading and Mathematics assessments annually in grades 3–8 and 10. Other subjects were assessed once per grade level, with Science assessed in grades 4, 7, and 11 and Writing assessed in grades 5, 8, and 11.

1.1.5. Kentucky Summative Assessments (2022–Current)

Starting in spring 2022, Kentucky public school students take the annual summative Kentucky Summative Assessments (KSA) to meet federal and state testing requirements. KSA replaced the previous K-PREP assessment and were developed by Kentucky teachers to align with the KAS in each subject area. KSA are administered in Reading and Mathematics in grades 3–8 and 10; Science in grades 4, 7, and 11; and Social Studies, On-Demand Writing, and Editing and Mechanics in grades 5, 8, and 11.

The KSA assessments are Kentucky's measure of student proficiency and progress on the state content standards that establish goals for what all students should know and be able to do in each grade. KSA are administered online, with only a small percentage of accommodated students taking them on paper. The assessments go beyond multiple-choice items to include extended-response and technology-enhanced items for students to demonstrate critical thinking and problem-solving skills.

1.2. Organizations and Groups Involved

Large-scale assessment programs depend heavily on the input of various professional organizations and stakeholder groups to maintain the confidence of the assessment users in the goals set forth for the assessment program. This next section highlights how various groups have contributed to the KSA program.

1.2.1. Kentucky Department of Education

KDE is headquartered in Frankfort, KY, and leads the design, implementation, and reporting of the accountability model and its components. KDE consists of smaller organizations that provide specific guidance to KSA. The Office of Assessment and Accountability (OAA) works directly on KSA with intra-office support from the Division of Accountability Data and Analysis (data and statistics) and the Division of Assessment and Accountability Support (DAAS). In addition, members of the Office

of Teaching and Learning provide content support on the KSA tests, reviewing and providing feedback on the construction of test forms.

1.2.2. Kentucky Educators

Educators play the next most significant role in the design and maintenance of large-scale assessment programs in the Commonwealth. During the initial development stages of an assessment program, educators are solicited to provide input on assessment design, including the best methods for assessing content. The role of educators in the design and maintenance of an assessment program is based on their unique instructional perspective garnered from their classroom experience and interaction with students. Each year, Kentucky educators are requested to participate in various capacities of test development. For example, as discussed in Chapter 2: Test Development, educators participate in item review meetings to review and discuss item quality, accuracy, and fairness. For these meetings, educators review test items and judge them appropriate for use on future KSA test forms. Here, educators directly affect test content, removing items from consideration or proposing changes to items to make them more appropriate for testing.

Educators participate in other meetings held throughout the lifecycle of an assessment program. During summer 2022, Kentucky educators were assembled virtually to recommend performance standards for the KSA Reading, Mathematics, Social Studies, and Writing tests, using their expertise to provide input on performance level descriptors (PLDs) and cut points for the KSA tests. See Chapter 5: Performance Standards for more details on these standard setting meetings.

1.2.3. School Curriculum, Assessment and Accountability Council

The Governor appoints members to the School Curriculum, Assessment and Accountability Council (SCAAC). The committee's existence was mandated by Executive Order 2021-729 and was created to study, audit, review, and make recommendations concerning Kentucky's system of academic standards, assessing learning, identifying academic competencies and deficiencies of students, holding schools accountable for learning, and assisting schools to improve their performance. SCAAC is comprised of 16 voting members and is authorized to request and receive data from any state or local government agency in the Commonwealth deemed necessary to fulfill the requirements of its mission, including any entity that derives a substantial portion of its funding from public sources.

1.2.4. Kentucky Technical Advisory Committee

Senate Bill 129 (2021) amended KRS 158.6455 by removing specific language around the National Technical Advisory Panel on Assessment and Accountability (NTAPAA) and allowing Kentucky to form its own technical advisory committee, known as the Kentucky Technical Advisory Committee (KTAC). The purpose of the committee is to provide advice and recommendations relating to the development of and modification to the assessment and accountability system, development of administrative regulations governing the assessment and accountability system, setting of standards used in assessment and accountability, and KRS 158.6453, 158.6455, 158.78, or 158.860. When requested, KTAC and KDE convene, along with other organizations (see Section 1.2.5. Contractors), to discuss measurement and/or accountability issues as determined by KDE.

1.2.5. Contractors

1.2.5.1. Human Resources Research Organization

The Human Resources Research Organization (HumRRO), a measurement solutions provider based in Louisville, KY, has a long-standing involvement with the Kentucky assessment program. HumRRO has conducted several alignment and validation studies for presentation to NTAPAA and for state and national conferences. HumRRO also provides quality control verification, replicating measurement analyses performed by prime contractors of state assessment programs, including Kentucky. Chapter 7: Calibration, Equating, and Scoring provides more details regarding HumRRO's involvement in the measurement analyses conducted on KSA by Pearson.

1.2.5.2. Pearson

Pearson's U.S. educational assessment division provides a full range of assessment and measurement services to states and districts throughout the U.S. As the prime contractor for KSA, Pearson works with KDE through its management of project schedules and deliverables, communications, and client meetings to develop valid and reliable assessments that fairly measure the educational progress of Kentucky students. By means of this technical manual and the accompanying documentation, Pearson describes all aspects of the development and delivery of KSA, from item generation to psychometric analysis to score interpretation.

1.2.5.3. Inclusive Large Scale Standards and Assessment

The Inclusive Large Scale Standards and Assessment (ILSSA) group is composed of staff at the University of Kentucky dedicated to the design and implementation of large-scale assessments for students with significant cognitive disabilities. ILSSA has been the contract lead for Kentucky's alternate assessment program since its inception in 1990. ILSSA developed a separate Alternate Kentucky Summative Assessments (AKSA) technical manual for the AKSA assessment program.

1.3. Kentucky Summative Assessment Program

This section provides a brief description of the subject areas and standards assessed through KSA. Chapter 2 outlines the test blueprint for each test.

1.3.1. Reading and Writing

New standards for Reading and Writing were adopted in 2019 based on Senate Bill 175 (2019). Development of the KSA Reading and Writing tests based on these standards represent a comprehensive view of literacy, incorporating reading, composition, and language to ensure that Kentucky students are fully prepared for a successful transition to post-secondary education, work, and the community.

The Reading tests are based on the KAS for Reading. Constructed-response items are explanatory in nature; students are asked to examine text and convey ideas and information to explain their thinking about what they have read. Writing is measured by a combination of the On-Demand Writing test and a brief Editing and Mechanics test that consists of multiple-choice and constructed-response items. The On-Demand Writing test is based on the KAS for Composition. Students respond to one prompt based on a text set. The Editing and Mechanics test is based on the KAS for Language and focuses primarily on Conventions of Standard English, although some items ask students to demonstrate knowledge of language and vocabulary use. More information on the KAS for English language arts (ELA) can be found on the KDE website (ELA).

1.3.2. Mathematics

The KSA Mathematics tests emphasize the balance between the Standards for Mathematical Practices and the Standards for Mathematical Content. The design is created to result in assessments that measure students' abilities to make sense and persevere when solving problems, use quantities appropriately, communicate and critique mathematical thinking, model with mathematics, strategically use tools, attend to precision, and look for and apply structure and patterns to solve problems within grade-level content. The Standards for Mathematical Content are a balanced combination of conceptual understanding, procedural skills/fluency, and application. Additionally, for grades K–8, the percent allocations for content items are based on grade-level domains. For high school, the percentage allocations for content items are based on conceptual categories (as described in the *High School Mathematics Matrix Standards by Course*). More information on the KAS for Mathematics can be found on the KDE website (Math).

1.3.3. Science

In 2015, Kentucky adopted a new set of science academic standards that features assessable performance expectations of what students should know and be able to do with foundations of science and engineering practices, core disciplinary ideas, and crosscutting concepts. In spring 2018, new Science assessments were administered in grades 4 and 7. In spring 2019, a new Science assessment was administered in grade 11. In spring 2022, the new Science assessments in grades 4, 7, and 11 were administered and reported on the KSA scale. The original cut scores were re-evaluated as part of the KSA standard setting conducted in spring 2022 (as described in Chapter 5: Performance Standards). More information on the KAS for Science can be found on the KDE website (Science).

1.3.4. Social Studies

All the KAS standards are eligible to be tested for the KSA Social Studies tests. Each grade-band assessment administered at grades 5, 8, and 11 consists of each discipline strand subdomain (civics, economics, geography, and history) where 50% of the items also reflect the inquiry standards. To achieve the target of the blueprint, test items may be dual-aligned to the KAS for Social Studies. More information on Social Studies can be found on the KDE website (Social Studies).

2. Test Development

Construction of the KSA test forms is a coordinated effort between KDE and Pearson, adhering to guidelines that promote fair and ethical testing practices. The process of constructing test forms begins with the development of content, writing and reviewing items that assess the content appropriately. Developing content for testing is not a simple task and requires detailed specifications, training, and quality control procedures. Using the content developed for testing, specialists work together to assess the appropriateness of the content, including the use of data to determine the statistical quality of the content. This chapter provides a description of the KSA test development process, including item development, content and statistical guidelines considered, and test form design.

2.1. Kentucky Academic Standards Alignment

One emphasis during KSA item and passage development is alignment to the Kentucky Academic Standards (KAS). Pearson began the KSA item development activities by evaluating items developed to assess KAS by a previous Kentucky state assessment contractor. This evaluation was used to create item development plans to bolster the item pool such that the KAS could be more fully represented (as described in the KSA blueprints). This allowed Pearson to create a robust item pool for the KSA assessments that appropriately represents the KAS, using an item bank application that maintains the blueprint requirements to guide the content development process and promote adequate coverage of the KAS for all future administrations of the KSA.

For KSA content development, Pearson designs item writer training materials that include references and discussions to the KAS, with key aspects highlighted for training purposes. Training on the KAS is essential to address interpretations of the standards so that all KSA assessment content is developed to the same guidelines. Item writer training material is reviewed and discussed thoroughly between KDE and Pearson and approved by KDE prior to item writer training. It is crucial that item writer training material is discussed prior to each development cycle for two reasons: (a) content development requirements may change year to year; and (b) interpretations pertaining to assessing KAS may change, dictated by national perspectives.

During item writer training, Pearson presents the KAS and points out key aspects to consider when developing content, including specific decomposition of standards into concrete domain targets (e.g., point of view and the relationship between texts in Reading). The goal of this training is to underscore the breadth of content necessary for assessing Kentucky's students on skills within the KAS framework. Item writers are provided with exemplars to guide their content development.

Pearson conducts internal reviews of content submitted by the contracted item writers. These initial reviews focus on appropriateness and specificity in assessing the KAS. Pearson engages with the item writers to discuss item alignment and suggested content revisions as necessary. Pearson has the authority to, and may, align items to the KAS differently than what was intended by the item writers. Items may be rejected by Pearson due to poor alignment to the KAS. The test content, alignments, and reviews by Pearson are prepared for review by KDE.

KDE reviews the test content and alignments to KAS for appropriateness. Content specialists review each piece of test content and recommend modifications to the KAS alignments as necessary. During this review, KDE and Pearson may discuss differences in interpretations of the KAS and appropriate solutions for assessing Kentucky's students. Once KDE has reviewed and approved the KAS alignment of new test content, Pearson conducts item review workshops with Kentucky educators.

During the item review workshops, participants review each piece of test content for its KAS alignment and content appropriateness. Changes to KAS alignments may be recommended by the committees, but these recommendations must be presented to KDE prior to any changes. KDE and Pearson may discuss recommended changes regarding previous decisions in KAS alignment. Changes in KAS alignment from the committee review must be consistent within the general scope of KAS alignment. Once changes in KAS alignment are applied after committee review and KDE approval, KDE reviews the alignment of new test content for accuracy prior to use by Pearson in building the test forms. KDE has the final authority on KAS alignment of all test content.

2.2. Item Development

Pearson developed item content for the KSA Reading, Mathematics, and Writing assessments. The goal of item development for these subject areas was to build upon item banks for assessing the KAS.

2.2.1. Item Specifications

To develop appropriate content for large-scale testing, individuals tasked with developing test content (i.e., items and passages) must follow specific guidelines that can be general or subject-area specific and give the item writers the parameters for creating content appropriate and suitable for assessing achievement. Appendix A provides passage specifications for Reading and On-Demand Writing as an example.

General guidelines for item writing include the following:

- Items must be clearly and concisely written.
- Items must accurately align to the intended academic standard.
- Items must be unique in approaches to assessing standards.
- Items must be grammatically (and/or mathematically) correct.
- Items should be aligned to *Depth of Knowledge (DOK)* levels to the extent that an adequate range of skill level is represented.

Guidelines of item writing are used to cover the specific aspects of each subject area. For example, Reading items must be answerable using the text and inferences from the text provided and must be specific to the passage provided when items are associated with passages. Multiple-choice answer options for Mathematics items should either be in ascending or descending order when containing numerical values. Item type and format guidelines are also used to promote consistency and appropriateness of items' presentation, task, and, in the case of multiple-choice items, answer options.

The accessibility of items for all intended test takers is also specified through guidelines of *universal design* that include precautions of items' discriminating based on age, gender, ethnicity, disability, socioeconomic status, and English language proficiency.

All guidelines are presented through training workshops and as documentation for use throughout the development of test content. The appendices of this manual contain various materials used within the item development process, including presentations for workshops and item review checklists, as shown below. The materials in these appendices reflect previous years of item development work for KSA. The processes highlighted through these materials are the objects of importance, rather than the actual years.

- Appendix A. Passage Specifications
- Appendix B. Mathematics Item Writer Training
- Appendix C. Social Studies Item Writing Training
- Appendix D. Item Development Review Criteria Checklist
- Appendix E. Item and Passage Writer Source Requirements
- Appendix F. Reading Item Content Review Training
- Appendix G. Mathematics Item Content Review Training
- Appendix H. Social Studies Item Content Review Training
- Appendix I. Item Content Review Checklist
- Appendix J. Mathematics and ELA Item Bias Review Training
- Appendix K. Social Studies Item Bias Review Training
- Appendix L. Item and Passage Bias Review Checklist
- Appendix M. On-Demand Writing Item Content Review Training
- Appendix N. On-Demand Writing Content Review Checklist
- Appendix O. On-Demand Writing Bias Review Checklist
- Appendix P. On-Demand Writing Scoring Rubrics

2.2.2. Item Writing

2.2.2.1. Item Writers/Training

Subject matter experts from the field of education are recruited to develop KSA test content. These individuals enter into an agreement with Pearson that outlines the tasks, proposed compensation, and guidelines for submitting completed work. Pearson then provides extensive training for writers prior to item development. KSA item writer training is provided by subject area, although similar training content is stressed in each training session. During training, the content standards and their measurement specifications are reviewed in detail. Pearson also discusses policies of content security and ownership. Training provides the foundation of best practices for item development.

2.2.2.2. Item Authoring

Once items are submitted by item writers, Pearson executes a process of review and editing before the items are included into the item banking applications. Pearson uses the Item Content Review Criteria Checklist and Item and Passage Writer Source Requirements before accepting items into the item bank. During this phase of item development, subject matter experts from Pearson review item metadata (e.g., standard/benchmark/objective, answer key, cognitive level) for accuracy, making revisions as needed. Items are also reviewed for appropriate, accurate content, and proper alignment to project specifications. Art specifications and inclusion of item reference objects (e.g., mathematical expressions/equations) are addressed during this review as well.

2.2.2.3. Quality Control

Throughout the item development process, quality control is instituted in a variety of ways. From the initial review of submitted items, multiple staff from Pearson work with and consult over the items. Collaboration on the items includes addressing accuracy in metadata, art, and factual information. Factual information, including art, presented in items is validated through at least two authoritative sources as researched by Pearson. If inaccurate information is found within an item, the correct information is provided.

Items go through many stages during the development process, each with a role of providing quality control measures. For example, *universal design* review provides checks on bias and sensitivity issues on the item, artwork, and stimuli. Scoring rubrics for performance items are also reviewed for what could lead to errors or other issues in handscoring. Furthermore, all revisions to items and other test content are made through the consultation of staff from Pearson for agreement, rather than through a single individual.

2.2.3. Item Review Committees

Kentucky educators and other stakeholders take part in the development of KSA test content through participation in item content and bias and sensitivity review committees. Participants are chosen to be representative of overall demographic characteristics. Beyond this, participants can be classified into three general groups: teacher, non-teacher educator, and general public. Teachers are individuals who are responsible for a classroom. Non-teacher educators have a background in education but are not K–12 classroom teachers. These individuals include curriculum specialists, administrators, and university instructors. Finally, the general public are individuals who are not directly involved with education but who may have been previously involved in education (e.g., retired teachers).

2.2.3.1. Content Advisory Committees

The content advisory committee reviews newly developed items for content, alignment to the standards, and appropriateness at the intended grade level. The participants work in groups, facilitated by Pearson, to recommend that items are accepted for testing, rejected for testing, or conditionally accepted (i.e., acceptance with minor modifications to the items).

2.2.3.2. Bias and Sensitivity Review

In addition to item content reviews, educators/stakeholders review items for fairness in all item material (e.g., passages, art) to prevent the use of material that discriminates or is offensive to any subgroup of students (e.g., gender, ethnicity, disability). From this review, items can be modified to adjust any content that is deemed inappropriate or completely removed from consideration.

2.2.4. Item Editing

After the various reviews are conducted, Pearson and KDE work together to edit items as recommended by the educators and other consultants. Once recommended edits have been made, the items are considered available to be field tested (i.e., administered to students within a standard testing environment for the purposes of collecting item performance data).

2.3. Scoring Guides

For constructed-response items (i.e., short answer and extended-response items), scoring guides are required to describe criteria that differentiate item responses by the achievable score points. Short answer items are worth two points, while the extended-response items are worth four points. A score point of zero can be obtained, but only due to some form of non-response (e.g., blank response or off-topic). Since each constructed-response item presents a different scenario, a unique scoring guide is constructed and used for each item. For On-Demand Writing, however, one scoring rubric is used for all writing prompts across all grades (see Chapter 10: Performance Scoring).

2.4. Test Form Development

Developing test forms is a process by which assessment specialists select and sequence items that assess subject area content as specified by the test design and blueprint documentation. The goal of test form development is to build assessments that allow students to demonstrate achievement to content and performance standards in a fair and appropriate manner. To accomplish this task, specialists work with various forms of specifications that provide parameters for building test forms.

2.4.1. Test Design and Blueprints

The *test design* is the layout of the test in terms of how many items will be administered, what types of items will be administered (e.g., multiple choice, short answer), and the number of sections a test may be divided into. These and other design factors can be considered, allowing assessment specialists to build test forms with the design most suitable for the purpose of the assessment.

Test blueprints, on the other hand, mainly provide specifications on content coverage —the number of items required per domain (i.e., reporting category). This includes how item types are chosen across domains and the number of total points associated. In some cases, though, fulfilling the requirements of a test blueprint is difficult due to item availability and weighing item selection with other considerations, e.g., statistical considerations discussed in the next section. In these cases, test developers provide documentation of the specific reasons that requirements of the test blueprints cannot be fulfilled.

Table 2.1–Table 2.6 present the test blueprints for each KSA subject-area test. For spring 2023, one writing prompt was administered in each grade for the Writing tests: opinion (grade 5) and argumentative (grades 8 and 11). In Mathematics, a matrix design for operational testing was utilized in order to meet sufficient point requirements by domain for reporting in a shortened test form. The distribution of domains present in each operational form (four per grade level) varied across the four forms. However, the blueprint was met across forms at every grade level to provide valid information at the school, district, and state levels.

Table 2.1. KSA Reading Test Blueprint

Grade	Domain	Domain Coverage (%)	Passage Type (% of Items) - Literary	Passage Type (% of Items) - Informative
3	Key Ideas	30–35	50	50
	Craft and Structure	30–35	50	50
	Integration of Knowledge and Ideas	30–35	50	50
4	Key Ideas	30–35	50	50
	Craft and Structure	30–35	50	50
	Integration of Knowledge and Ideas	30–35	50	50
5	Key Ideas	30–35	50	50
	Craft and Structure	30–35	50	50
	Integration of Knowledge and Ideas	30–35	50	50
6	Key Ideas	30–35	45	55
	Craft and Structure	30–35	45	55
	Integration of Knowledge and Ideas	30–35	45	55
7	Key Ideas	30–35	45	55
	Craft and Structure	30–35	45	55
	Integration of Knowledge and Ideas	30–35	45	55
8	Key Ideas	30–35	45	55
	Craft and Structure	30–35	45	55
	Integration of Knowledge and Ideas	30–35	45	55
10	Key Ideas	30–35	40	60
	Craft and Structure	30–35	40	60
	Integration of Knowledge and Ideas	30–35	40	60

Table 2.2. KSA Mathematics Test Blueprint

Domain	Target %						
	Grade 3	*Grade 4*	*Grade 5*	*Grade 6*	*Grade 7*	*Grade 8*	*Grade 10*
Operations and Algebraic Thinking	30–35	15–20	15–20	–	–	–	–
Number and Operations in Base Ten	15–20	25–30	25–30	–	–	–	–
Number and Operations – Fractions	20–25	25–30	25–30	–	–	–	–
Measurement and Data	15–20	10–15	10–15	–	–	–	–
Geometry	10–15	10–15	10–15	–	–	–	–
Ratios and Proportional Relationships	–	–	–	10–15	20–25	–	–
The Number System	–	–	–	30–35	15–20	–	–
Expressions and Equations	–	–	–	25–30	20–25	25–30	–
Geometry	–	–	–	15–20	20–25	25–30	25–30
Statistics and Probability	–	–	–	15–20	20–25	10–15	10–15
The Number System	–	–	–	–	–	10–15	–
Functions	–	–	–	–	–	25–30	22–27
Algebra	–	–	–	–	–	–	22–27
Number and Quantity	–	–	–	–	–	–	10–15
Non-Calculator	60–70	60–70	60–70	30–35	30–35	20–25	20–25

Table 2.3. KSA Science Test Blueprint (2018-Present)

Domain	Target (%)		
	Grade 4	*Grade 7*	*Grade 11*
Physical Science	30–45	35–50	20–35
Life Science	20–35	15–30	30–45
Earth and Space Science	25–40	15–30	20–35
Engineering Design	5–15	5–15	5–15

Table 2.4. Social Studies Test Blueprint

Domain	Target (%)		
	Grade 5	*Grade 8*	*Grade 11*
Civics	25	25	25
Economics	25	25	25
Geography	25	25	25
History	25	25	25

Table 2.5. KSA On-Demand Writing Test Blueprint

Grade	Mode	Domain Coverage (%)
5	Opinion	100
8	Argumentative	100
11	Argumentative	100

Table 2.6. KSA Editing and Mechanics Test Blueprint

Grade	Mode	Domain Coverage (%)
5	Conventions of Standard English	80
	Knowledge of Language and Vocabulary Acquisition and Use	20
8	Conventions of Standard English	80
	Knowledge of Language and Vocabulary Acquisition and Use	20
11	Conventions of Standard English	80
	Knowledge of Language and Vocabulary Acquisition and Use	20

2.4.2. Form Content Alignment

Pearson uses two content specialists for each new KSA test form developed. The first content specialist is responsible for constructing a test form meeting both content and statistical requirements, whereas the second content specialist is responsible for verifying the content alignment of the test form, providing feedback on the match to the test design and blueprint and the accuracy of specified item characteristics (e.g., DOK and answer key). The verification of content alignment may result in feedback suggesting modifications in the items selected for the test form. These suggestions are reviewed and implemented, as necessary, prior to psychometric and KDE review.

During the psychometric review of test forms, the blueprint is reviewed, and feedback is provided with suggestions for improving the match to the test blueprint. KDE also reviews the test forms for blueprint alignment and requests modifications as necessary.

2.4.3. Statistical Guidelines

In addition to content considerations for constructing test forms, statistical considerations must be considered as well. Item statistics are discussed in more detail in Chapter 6: Item Analyses, but a brief mention of the statistics is appropriate here. Statistical guidelines are provided for selecting test items that are fair to all students, including representing a variety of difficulty. Specific guidelines include the following:

- Percent correct is between 30% and 85% for multiple-choice items.
- Item mean score is between 0.60 and 1.70 for short answer items.
- Item mean score is between 1.20 and 3.40 for extended-response items.
- The correlation between item score and total score must be at least 0.20.

Consideration of items outside of these parameters is given when there is little to no choice for meeting test blueprints. In addition, the interaction between percent correct and item-total-score correlation can indicate difficult items that function appropriately within the testing population. For example, an item with a 25% correct response may have an item-total-score correlation slightly above the criterion of 0.20.

Other guidelines must also be considered from a statistical perspective. *Differential item functioning* (DIF) refers to items with a difference in performance across subgroups. For example, an item showing DIF may indicate that males, overall, were more successful on an item than females; or in another case, one ethnicity group outperformed another. Although an important index, it is typically cautioned that statistical results indicating a presence of DIF should be weighed against actual item content. In other words, it is recommended item content is reviewed for bias before an item is judged to be truly exhibiting DIF. Because items are reviewed for bias during the item development phase prior to obtaining statistical data, it is recommended that statistics not become the sole deciding factor in item use given previous scrutiny during item development.

2.4.4. Field Testing

Part of maintaining the integrity of an assessment program over time is to use new items during each assessment cycle. Using new items prevents test content from being compromised due to overexposure, which could lead to questions of test validity. Item development activities occur during each year of the assessment, or as stipulated in work scopes. All newly developed items that pass the item review process are field tested or administered to students to obtain low-stakes performance data.

For the new KSA assessments, items in Reading, Mathematics, Social Studies, On-Demand Writing, and Editing and Mechanics were field tested in 2020 and 2021 with stand-alone field tests (as opposed to embedded within operational forms). Embedding of field test items within operational forms resumed for Spring 2023 administration. For multiple-choice items, the minimum number of responses per field test item can be a few thousand responses. However, for constructed response items (i.e., short answer and extended-response items), only 2,500 responses are selected and scored for item analysis. The selection of responses is random such that all achievable scores are represented for analysis. All item types were field tested as needed for maintaining a suitable pool of items for subsequent test form creation.

After field testing, student performance is analyzed, and decisions are made regarding the future use of the field tested items. In some cases, the statistics of an item will lead to item reviews that may deem the item inappropriate for future use. Performance data from the field item items are also used during test construction for selecting the best available test items.

2.5. Braille and Large Print Test Materials

Federal and state laws require accessibility of test materials for all students. Test materials must be developed to accommodate the various needs of students within a testing population. Visually impaired students participate in the KSA assessment program via Braille or large print versions of the test materials. Test forms for these students are modified reproductions of the test form constructed for the general population. However, it is often the case that some items are not appropriate for translation into Braille. In these situations, items are either replaced with items that can be translated into Braille or they are simply not counted toward students' test scores who use the Braille form.

KSA items that were not appropriate for Braille were removed from inclusion in the Braille students' test scores, thus reducing the maximum number of test points for Braille students. As discussed in Chapter 7: Calibration, Equating, and Scoring, this resulted in separate scoring tables between the general and Braille testing population.

3. Test Administration

To maintain the standardization of administering a large-scale assessment such as KSA, several guidelines must be strictly followed by those involved in the test administration process. These guidelines are developed by internal and external groups and presented in manuals and through training workshops that stress the importance of adhering to these guidelines. For KSA, the *Test Administration Manual (TAM)* is developed in collaboration between KDE and Pearson and outlines administration procedures for before, during, and after the test administration. This chapter highlights some of the topics presented in the TAM regarding overall test administration procedures, including testing dates, student eligibility, and testing accommodations. This chapter also discusses other manuals that are published to guide the KSA administration.

3.1. Test Administration Window

Districts within the Commonwealth of Kentucky begin and end schooling at different times of the year. Therefore, the prescribed test administration window for KSA is based on a district's last day of school, although a general test administration window is specified. Each district is required to administer KSA within the last 14 instructional days of its academic calendar.

In the event of natural disasters or other extenuating circumstances that cannot be controlled by the school or district, the test administration window may be extended. The Department of Education, Office of Assessment and Accountability (OAA) must approve all extensions to the testing window.

3.2. Test Make-Up Procedures

Students may make-up any portion of the KSA assessment during the 14-day administration window or during the four days after the testing window, during which test materials are prepared for return shipping.

3.3. Eligibility Requirements and Exemptions

All students enrolled in grades 3–8, 10, and 11 are required to take KSA, unless they are participating in the Alternate KSA. Participation in the KSA test administration includes the following:

- Students with disabilities
- Students who are retained
- Students who moved during testing
- Students experiencing a minor medical emergency
- English learners (ELs) who are, at least, in their second year of attending a U.S. school.[2]

Students who do not participate in KSA include the following:

- Students participating in the Alternate KSA
- Students expelled and not receiving academic services
- Foreign exchange students
- Students medically unable to take the assessment
- Students who moved out of the Kentucky public school system during the testing window

[2] ELs in their first year must participate in KSA Mathematics where tested at their grade.

Students may be exempt from KSA based on factors not mentioned above. A medical exemption, for example, can be filed for extenuating medical circumstances. Appendix A of the *Yearbook* contains a table of participation rates for each grade-level and subject-area test.

3.4. Accommodations

Testing accommodations are modifications to the testing environment that allow students with special needs to participate in the test administration and demonstrate content achievement. Accommodations used for the test administration are often used during instruction as well, as these accommodations are typically specified in student-specific academic records such as an Individualized Education Program (IEP) or 504 Plan. Accommodations and their acceptable use are clearly defined in the manuals published for KSA test administration. Below is a list of the accommodations used on KSA:

- Use of assistive technology
- Manipulatives
- Reader
- Scribe
- Hand-held calculator (only students that receive specific accommodations can use a hand-held calculator)
- Extended time
- Reinforcement and behavioral modification strategies
- Interpreters for students with deafness or hearing impairment (signing)
- Oral native language support for ELs
- Bilingual/English dictionary

3.5. Test Administration Procedures

Administering a large-scale assessment requires coordination, detailed specifications, and proper training. Along with this, several individuals are involved in the administration process, from those handling the test materials to those administering the tests. Without the proper training and coordination of these individuals, the standardization of the administration could be compromised. KDE works with Pearson to develop and provide the training and documentation necessary for KSA to be administered under standardized conditions throughout all testing environments.

3.5.1. District Assessment Coordinators

Training for KSA test administration is provided to District Assessment Coordinators (DACs) by the Division of Assessment and Accountability Support (DAAS). This training emphasizes the roles and responsibilities of the DACs and Building Assessment Coordinators (BACs) for before, during, and after test administration. The DACs are responsible for all aspects of the KSA test administration, including providing test materials and training to the BACs. The DACs also serve as the point of contact for Pearson in the case of issues with online testing or accommodated test materials (e.g., accommodated test materials ordering).

3.5.2. Grade-Level Scripts

The *grade-level scripts* include explicit directions and scripts to be read aloud to students by test administrators and/or proctors.

3.5.3. Test Administration Manual

The TAM provides test administrators guidelines on preparing online testing environments and the assembly of accommodated test materials for returning to the BACs. Given its content and purpose, the TAM further promotes the standardization of KSA test administration. The assessment coordinators are instructed to read the TAM in preparation for the KSA test administration.

3.6. Test Security

The high-stakes nature of the KSA assessment necessitates the need for test security measures to protect the program's integrity. Policies for KSA test security are outlined in the TAM, and all individuals participating in the KSA test administration must adhere to these policies. Adhering to test security policies includes reporting any suspicions of security breaches immediately to the appropriate authority, as outlined in the TAM. KDE investigates all allegations of test security breaches.

Receipt and shipping of test materials are handled by DACs using tracking sheets provided by Pearson. The TAM provides detailed specifications on inventorying test materials upon arrival and prior to return shipping to Pearson. It is critical that the procedures for shipping are followed to protect the tests from unauthorized exposure.

All administrators/proctors are required to certify their knowledge of and adherence to the policies and guidelines of the KSA test administration. The *Appropriate Assessment Practices Certification Form* certifies that the administrators/proctors have read and understand what is and is not allowed when participating in the KSA test administration.

4. Reports

Multiple reports are used to document student performance on the KSA assessments, presenting different levels of summary information and targeting different audiences. This chapter discusses the various KSA score reports, including specific pieces of information and general cautions on using the reports. Sample score reports are provided in Appendix B of the *Yearbook*.

4.1. Description of Scores

4.1.1. Scaled Score

Scaled scores are derived scores from a statistical transformation of the raw scores, representing a metric that is consistent across test forms and allowing for comparisons across test administrations within a subject and grade. As discussed in more detail in Chapter 7: Calibration, Equating, and Scoring, scaled scores are used to identify the proximity of test performance to established criteria (e.g., passing the test). For KSA, the range of scaled scores is set to 400–600 for each test.

4.1.2. Student Performance Level

Student achievement on KSA is defined by *performance levels* within a classification system of achievement from low proficiency to high proficiency. The KSA has four levels of achievement: *Novice*, *Apprentice*, *Proficient*, and *Distinguished*. These labels are accompanied by performance level descriptors (PLDs) that define the knowledge and skills typical in each level. Performance level summaries are included on the KSA score reports at all levels of reporting (i.e., student, school, district, and state), although the PLD is only included on the ISR as it provides a description of individual student achievement. Chapter 5: Performance Standards discusses the performance level designations and PLDs, and Chapter 7: Calibration, Equating, and Scoring discusses the alignment of scaled scores to the performance levels.

4.2. Description of Reports

4.2.1. Individual Student Report

The Individual Student Report (ISR) provides test score information at the student level for each subject-area test assessed. Scaled scores are reported along with the designated performance level (*Novice*, *Apprentice*, *Proficient*, or *Distinguished*). The performance levels are accompanied with the appropriate PLD that describes the knowledge and skills typically achieved for that performance level. The student's scaled score is also shown against the average scaled score at the school, district, and state level. For Writing, the scaled score is reported with the corresponding performance level and PLD. Like the scaled score for the other subject tests, this score is shown against the mean score at the school, district, and state levels. Additional statements are included as suggestions for continued achievement in each subject area assessed.

4.2.2. School Listing Report

The School Listing report provides a list of all students within a particular school along with their scaled scores and performance levels. This report is created by grade and varies due to the different subject areas assessed within each grade. The school listing report also identifies the students who used test accommodations.

4.2.3. Kentucky Performance Report

The School, District, and State Summary reports provide test score summary information at these three levels of score reporting, providing information for educators and administrators to compare student achievement at various levels.

The School Summary report provides a summary of test performance for all students within a school for a particular subject area and grade, along with summary information at the district and state levels for comparison. This report provides the percentage of students in each performance level along with the percentages at the district and state levels. The school summary report also provides percentages of the school's students that fall above and below the mean scores from the school, district, and state levels.

The District Summary report provides the same information as the School Summary report but aggregated by school. In other words, the summary information is presented for each school within a particular district. The State Summary report provides achievement summary information by district.

4.3. Appropriate Uses for Scores and Reports

The test forms constructed for KSA cover a sampling of curriculum content as specified through the test blueprints; the tests do not assess all possible content on one test form. The content is also assessed through a limited range of item types. Furthermore, the KSA assessments are administered once during the academic year, providing a snapshot of student achievement at a designated point of instruction. Given these limitations of assessment, test scores should only be interpreted and used in the context from which they are obtained. In other words, KSA test scores should be used to describe student achievement on the content assessed (i.e., grade level) and not used to generalize achievement beyond the test. Academic placement decisions and promotions should also not be based solely on KSA test scores but should include other indicators of achievement.

For example, the ISR communicates an individual student's test scores and interpretations of achievement based on those scores. The types of score information presented on an ISR depend on the grade level of the student. The ISR provides a snapshot of achievement and explains the meaning of each piece of information provided, providing valuable information to students and parents. It is important that users of these reports do not extend the score information beyond the interpretations provided.

Test scores are also summarized in the summary reports at the school, district, and state levels, providing valuable achievement information to educators and administrators. These reports are useful for evaluating curriculum and instruction and delineating areas at a group level where progress in achievement may be necessary.

4.4. Cautions for Score Interpretations and Use

KSA test results can be interpreted in many ways and used to make inferences about a student, educational program, school, or district. These results must be used appropriately to prevent inaccurate interpretations.

4.4.1. Understanding Measurement Error

When interpreting test scores, it is important to remember that test scores always contain measurement error. For example, test scores are expected to vary if the same student tested multiple times using equivalent test forms due to fluctuations in a student's mood or energy level or the items and tasks presented on a particular test form. Because measurement error can vary, they can cancel out when scores are aggregated across students. Chapter 8: Reliability provides information on evidence gathered that indicates that measurement error on the KSA assessments is within an acceptable range.

4.4.2. Interpreting Scores at Extreme Ends of the Distribution

Test scores at the extreme ends of the score range should be interpreted with caution. A perfect score does not indicate that a perfect score would be obtained if the test were longer. In addition, because test scores are expected to change with multiple testing attempts, students with high scores on one test may achieve lower scores the next time they test. Similarly, students with low scores on one test may achieve higher scores the next time they test. This is due to the *regression to the mean* phenomenon. Changes in a student's test score over multiple testing events may be due to regression toward the mean rather than differences in achievement. Scores at the extreme ends of the score range must be viewed cautiously and not interpreted beyond the context from which they occur.

4.4.3. Limitations When Comparing Scaled Scores at Reporting Group Levels

Test scores of demographic or program groups can be compared within a subject-area and grade-level test to see which group has the highest (and lowest) average performance. The mean scaled score provides a convenient representation of where the center of a set of scores lies, but it does not provide all information regarding the score distribution. Two groups with similar mean scaled scores can have different score distributions. Therefore, conclusions about the overall distributions cannot be made when viewing group mean test scores.

4.4.4. Inappropriateness of Comparing Scaled Scores Between Content Tests

Test scores between subject-area tests are not on the same scale and should therefore not be compared. As discussed in Chapter 7: Calibration, Equating, and Scoring, test scores within a particular subject-area and grade-level test are placed on the same scale such that scores can be compared across test administrations.[3] The constructs (traits) measured across subject-area tests vary to the extent that the scores cannot be used interchangeably for comparisons.

4.4.5. Program Evaluation

Test scores can be a valuable tool for evaluating programs, but any achievement test can give only one part of the picture. As addressed in Standard 13.9 in the *Standards for Educational and Psychological Testing*, "In evaluation or accountability settings, test results should be used in conjunction with information from other sources when the use of the additional information contributes to the validity of the overall interpretation" (AERA et al., 2014, p. 213). KSA does not measure every factor that contributes to the success or failure of a program. Test scores, therefore, should be considered as only one component of an evaluation system.

[3] For 2023, equating for KSA applies to all subject areas that were tested.

5. Performance Standards

Descriptions of student performance are used to help enhance the reporting of student scores beyond an overall reported score and references to other students or groups of students. Performance levels and descriptions of performance divide the test scores into meaningful categories and align to performance ranging from low to high. For Kentucky, these categories are called *Novice*, A*pprentice*, *Proficient*, and *Distinguished*. PLDs accompany these labels to describe typical performance of students within each group.

This chapter describes the development of the PLDs for the KSA and the standard setting that took place in July 2022 to set the KSA cut scores to distinguish performance among the four performance levels. In addition, the KSA standard validation process that took place in June 2023 to review the current cut scores for the KSA in Mathematics, Reading, Social Studies, and Writing is summarized in this chapter. A separate comprehensive report provides full details of this process, including descriptive information about the panelists involved.

5.1. Performance Level Descriptors

In spring 2022, a draft set of PLDs representing an increasing set of expectations across the Kentucky performance levels were created by KDE content staff and consultants with support from Pearson content specialists. The final approved KSA PLDs are located online at KDE website (PLDs). In July 2022, Kentucky educators were convened to operationalize the PLDs through standard setting, a process of determining test score thresholds, or cut points, to divide the test scores into the four performance level groups.

5.2. Standard Setting Process for KSA

From July 25–29, 2022, after the first operational KSA administration, a standard setting committee meeting was conducted to provide cut score recommendations for the Kentucky summative assessments for Mathematics, Reading, Social Studies, Writing, On-Demand Writing, and Editing and Mechanics. Science standards were set originally in 2018 and 2019. At these same July meetings, a validation of the Science cut scores within the KSA framework was also carried out by Kentucky educators.

A total of 26 committees were convened, one for each subject-area and grade-level assessment. The committees were comprised of teachers and non-teacher educators, with some panelists participating in multiple committees. Panelists were selected to provide content and grade-level expertise and be representative of the state teaching population, including geographic region, gender, ethnicity, educational experience, community size, and community socioeconomic status.

The bookmark standard setting method was used (Lewis et al., 1996; Mitzel et al., 2001) to recommend the performance level cut scores for each assessment (i.e., the *Apprentice*, *Proficient*, and *Distinguished* cut). This is a content- and item-based method that leads panelists through a standardized process through which they consider student expectations, as defined by the PLDs, and the individual items that could be administered to students to recommend cut scores for each performance level.

The key material used by the committee was a set of test items arranged in order of difficulty. Panelists identified and discussed the knowledge, skills, and abilities required to respond to the test items and divided the items into two groups: (a) items that a student who is minimally qualified for a performance level would likely answer correctly; and (b) items too difficult for students at that same performance level. This process was repeated for each performance level cut score in each subject area and grade.

The process started with panelists reviewing the design of the specific assessment and experiencing the different item types. Based on their experience with the test items, a review of the draft PLDs followed where panelists created borderline descriptions. During this process, committees modified the PLDs to create descriptors of the knowledge, skills, and abilities that students with performance at the borderline of the performance level (i.e., students who just barely enter a performance level) would be expected to demonstrate.

Panelists then completed rounds of judgments, reviewing and discussing judgment feedback between rounds. During this process, panelists reviewed items in the ordered item set regarding a performance level and answered the judgment question, "Would a student with performance at the borderline of the performance level likely get the item correct?"

For the purposes of the standard setting, "likely" was defined as two out of three students at the borderline of the performance level. The cut score recommendation for the performance level was determined as the last item that the borderline student would be expected to answer correctly. This process was repeated for each performance level. The standard setting committees for Mathematics, Reading, Social Studies, Editing and Mechanics, and On-Demand Writing completed three judgment rounds. Each recommended cut score from the standard setting committee was the median of the recommendations from the individual panelists in the committee.

For Science, the panelists completed two rounds of judgments. As part of the feedback from Round 1, the panelists were provided the items in the ordered item set that were associated with the current performance level cuts scores along with a reasonable range for each performance level. During Round 2, based on their recommended performance level cut scores and the current performance level, the panelists stated whether they would validate the current cut score or recommend new cut scores. If at least half of the panelists recommended retaining the current performance level cut score, the recommendation was to use the current cut scores, otherwise the recommendation was to use the new cut score recommendations. This judgment was made for each individual performance level.

In addition to separate Editing and Mechanics and On-Demand Writing performance level recommendations, an overall Writing performance level determination was also needed based on the combination of Editing and Mechanics and On-Demand Writing. Panelists recommended the general rules for determining the overall Writing performance levels for all grades. It was noted that On-Demand Writing performance should be weighted more than the Editing and Mechanics performance.

Further, if the Editing and Mechanics performance level is the same or one level different from the On-Demand Writing performance level, the Writing performance level should be the same as the On-Demand Writing performance level. Lastly, if the

Editing and Mechanics performance level is two levels or greater different from the On-Demand Writing performance level, the overall Writing performance level would be one performance level different from the On-Demand Writing performance level in the same direction as the Editing and Mechanics performance level.

After the standard setting committee finished, a vertical articulation committee composed of panelists from the standard setting committees convened to consider the recommended cut scores for each assessment. The articulation committee considered the recommended cut scores, the impact on Kentucky students, and the patterns of the performance standards across grades before adjusting the cut scores as needed to promote articulation and consistency across the assessment program.

To create a common point of reference across the assessments, cut scores and measures of student achievement on all KSA assessments are translated to a scale that ranges from 400 to 600 points. The scaled scores for the performance level cut scores (i.e., the *Apprentice*, *Proficient*, and *Distinguished* cuts) were determined using a common scaling slope for all subject areas except On-Demand Writing, as described in Chapter 7: Calibration, Equating, and Scoring.

KDE reviewed the recommendations from the standard setting panels after articulation for reasonableness within a policy perspective to determine if any additional adjustments were warranted. Final cuts were presented to Commissioner Jason Glass on August 4, 2022, where he reviewed and approved them. Participation rates for the standard setting meeting for some panels was low enough that a recommendation of a standards validation meeting be carried out in spring 2023, which was also supported by the Commissioner.

Table 5.1 presents details of the current cut scores, including references to the underlying theta scales for each respective grade and subject area in addition to the transformed KSA scaled score values (described in Chapter 7). Performance data (i.e., impact data) provided to panelists and KDE at the time of the standard setting are also included for each performance level. Table 5.2 represents the current rules for deriving overall Writing performance indicators.

Table 5.1. Final Cut Scores and Impact Data

Subject	Grade	Theta Cuts			Scaled Score Cuts			Final Impact Data			
		N-A	*A-P*	*P-D*	*N-A*	*A-P*	*P-D*	*N*	*A*	*P*	*D*
Reading	3	-0.5891	0.1892	1.0742	500	513	528	28%	27%	27%	18%
	4	-0.3950	0.3841	1.2615	503	516	531	28%	25%	29%	18%
	5	-0.1847	0.7292	1.6826	507	522	538	27%	27%	28%	18%
	6	-0.3442	0.4746	1.3341	504	518	532	26%	29%	30%	15%
	7	-0.5546	0.0919	0.9348	501	512	526	29%	25%	29%	17%
	8	-0.3741	0.3271	1.0785	504	515	528	28%	28%	28%	16%
	10	-0.5286	0.1987	1.1255	501	513	529	29%	26%	28%	17%

Subject	Grade	Theta Cuts			Scaled Score Cuts			Final Impact Data			
		N-A	*A-P*	*P-D*	*N-A*	*A-P*	*P-D*	*N*	*A*	*P*	*D*
Mathematics	3	-0.2926	0.6838	1.9209	505	521	542	30%	32%	28%	10%
	4	-0.1984	0.6893	1.9941	507	521	543	31%	30%	30%	9%
	5	-0.6449	0.3230	1.6406	499	515	537	30%	32%	28%	10%
	6	-0.9178	-0.1552	0.9485	495	507	526	30%	32%	28%	10%
	7	-0.8682	-0.2962	0.7121	496	505	522	31%	30%	29%	10%
	8	-0.8917	-0.3103	0.8629	495	505	524	35%	26%	29%	10%
	10	-0.9654	-0.3707	0.6498	494	504	521	30%	32%	28%	10%
Science	4	-0.8878	0.2775	1.2302	495	515	531	16%	55%	13%	16%
	7	-1.0819	-0.0138	1.1226	492	510	529	35%	45%	18%	2%
	11	-1.0422	0.1515	1.3864	493	513	533	41%	44%	14%	1%
Social Studies	5	-0.3812	0.3316	1.2264	504	516	530	32%	29%	26%	13%
	8	-0.4339	0.2141	1.1015	503	514	528	36%	27%	25%	12%
	11	-0.5108	0.2409	1.0571	501	514	528	36%	28%	24%	12%
Editing and Mechanics	5	-0.1041	0.7086	1.4513	508	522	534	20%	32%	28%	20%
	8	-0.3897	0.3925	1.3579	504	517	533	21%	30%	33%	16%
	11	-0.3448	0.6818	1.7087	504	521	538	22%	31%	30%	17%
On-Demand Writing	5	-4.8047	0.3950	3.2464	486	512	526	19%	40%	35%	6%
	8	-6.6816	-2.1289	4.4445	477	499	532	19%	39%	36%	6%
	11	-7.9679	-0.9581	5.4224	470	505	537	20%	37%	36%	7%

Note. N = *Novice*, A = *Apprentice*, P = *Proficient*, D = *Distinguished*

Table 5.2. Overall Writing Performance Level Profiles

Subject	Performance Level	On-Demand Writing			
		Novice	*Apprentice*	*Proficient*	*Distinguished*
Editing & Mechanics	*Distinguished*	*Apprentice*	*Proficient*	*Proficient*	*Distinguished*
	Proficient	*Apprentice*	*Apprentice*	*Proficient*	*Distinguished*
	Apprentice	*Novice*	*Apprentice*	*Proficient*	*Proficient*
	Novice	*Novice*	*Apprentice*	*Apprentice*	*Proficient*

5.3. Standards Validation Process for KSA

Because of challenges around the recruitment and retention of subject-matter experts at the 2022 standard setting meeting, KDE determined that a standards validation process would be appropriate in 2023. The purpose of the standards validation process was to allow panelists the opportunity to review the performance level cut scores and either confirm that they were appropriate or determine what adjustments would be appropriate for the current cut scores.

Pearson, in collaboration with KDE and with the assistance of ACS Ventures, LLC, recruited a team of Kentucky educators to review and evaluate the current cut scores and determine if any would need to be updated. Committees of Kentucky educators were identified to complete a review of the current scores using policies and procedures that were consistent with the practices followed in 2022 with slight modifications to reflect the nature of the standards validation aspect of the work.

From June 5–8, 2023, a series of standard validation committee meetings were conducted to review and recommend whether any changes were appropriate to the current cut scores for the Kentucky summative assessments for Mathematics, Reading, Social Studies, On-Demand Writing, and Editing and Mechanics.

There were 23 committees that reviewed the set of current cut scores for each assessment. The committees were comprised of teachers and non-teacher educators; some panelists participated in multiple committees. Panelists were selected for the standards validation committee to provide content and grade-level expertise and be representative of the state teaching population, including geographic region, gender, ethnicity, educational experience, community size, and community socioeconomic status. Extraordinary efforts were introduced to bring as many Kentucky educators into the standards validation process as possible. Many of the committees had ten or more panelists (8 of the 23 committees) engage in the process, while the smallest number committees were comprised of six panelists (4 of the 23 committees).

The validation process closely mirrored the bookmark method that was used for the standard setting meeting in 2022 (Lewis et al., 1996; Mitzel et al., 2001; Schultz & Mitzel, 2009). In 2023, the bookmark procedures were modified slightly in comparison to the procedures followed in 2022. The primary differences in procedures included:

- Two rounds of ratings were completed, in contrast to the three rounds of ratings completed in 2022.
- Panelists were informed of the location of the current cut scores. Items within a reasonable band around each cut score were also identified for the panelists. Items around the current cut score, referenced as a performance level error band, were within ½ of a conditional standard error of measurement (CSEM) and considered to represent item difficulties that were generally consistent with the current cut scores.[4] These item clusters around the current cut scores are collectively referenced as an error band.

All committees met virtually and accessed materials using the Pearson Standard Setting website, which provides secure transmission of the data and information necessary to complete all tasks. The process started with a general orientation session, with the lead facilitator providing a brief overview of the goals and purpose of the meeting, along with the reason behind the need for a standards validation activity. A representative from KDE also reviewed the task being presented to the panelists and summarized the activities completed.

Panelists were then split into breakout rooms, one for each grade/subject area, reviewing the design of the specific assessment and experiencing the different item types. After reviewing the current test, panelists completed a review of the borderline PLDs developed in 2022 with the facilitator leading a discussion of the key aspects of the borderline PLDs and the knowledge and skills defined at each performance level.

Panelists then completed two rounds of judgments, reviewing and discussing judgment feedbacks between rounds. During this process, panelists reviewed items

[4] For the On Demand Writing, pages within 1 CSEM were identified and considered to be consistent with the current cut score recommendations.

in the ordered item set regarding a performance level and answered the judgment question, "Would a student with performance at the borderline of the performance level likely get the item correct?"

For the purposes of the standards validation, "likely" was defined as two out of three students at the borderline of the performance level. The cut score recommendation for the performance level was determined as the last item that the borderline student would be expected to answer correctly. This process was repeated for each performance level.

After the first round of judgments, all panelists' ratings were summarized, with the median value considered to be the cut score from the committee. Panelist were provided a series of feedback data and information to help facilitate their review and discussion before completing their second round of ratings. The feedback provided to panelists included:

- the overall median recommendation, along with the minimum and maximum recommendations received across all panelists;
- information on the range and distribution of individual panelist recommendations to allow each panelist to see how their recommendation compared to other members of the committee; and
- impact data, or the percent of students classified into each of the four performance categories using the committees' cut score recommendations.

The facilitator led a discussion of the cut score recommendations with the panel, after providing all feedback to the committee. The discussion included a review of specific items that were centered around each of the cut score recommendations, the rational of panelist for the placement of their cut scores, and a discussion of the impact data and whether the panelists felt that the impact was consistent with their expectations for student performance.

Once cut scores were identified, the cut score recommendations from the 2023 meeting were compared to the currently implemented cut scores defined in 2022, presented in Table 5.1. The performance level error bands described earlier in this section (error bands were defined as plus or minus ½ of a CSEM from the current cut score) were used to determine if the new cut score recommendations were within a reasonable range of scores around the current cut scores. The results of the comparison are shown within Table 5.3. As can be seen in the tables, in all instances but one, the updated cut score recommendations were consistent with the cut scores established in 2022. For the Mathematics Grade 4 Distinguished cut score, the current cut score resides at page 52 in the ordered item set, with the error bands ranging from pages 50 to 54. The cut score recommendation from the 2023 standards validation was set at page 49, just below the error band.[5]

[5] The detailed results from each panel can be found in the Standards Validation Executive Summary document.

Table 5.3. Consistency of 2023 Cut Score Recommendations with 2022 Cut Scores

Subject	Grade	Within error band		
		N-A	*A-P*	*P-D*
Reading	3	Yes	Yes	Yes
	4	Yes	Yes	Yes
	5	Yes	Yes	Yes
	6	Yes	Yes	Yes
	7	Yes	Yes	Yes
	8	Yes	Yes	Yes
	10	Yes	Yes	Yes
Mathematics	3	Yes	Yes	Yes
	4	Yes	Yes	Lower
	5	Yes	Yes	Yes
	6	Yes	Yes	Yes
	7	Yes	Yes	Yes
	8	Yes	Yes	Yes
	10	Yes	Yes	Yes
Social Studies	5	Yes	Yes	Yes
	8	Yes	Yes	Yes
	11	Yes	Yes	Yes
Editing and Mechanics	5	Yes	Yes	Yes
	8	Yes	Yes	Yes
	11	Yes	Yes	Yes
On-Demand Writing	5	Yes	Yes	Yes
	8	Yes	Yes	Yes
	11	Yes	Yes	Yes

Note. N = *Novice*, A = *Apprentice*, P = *Proficient*, D = *Distinguished*

Due to the overall consistency with the current cut scores, it was determined during the standards validation meeting that a vertical articulation process was no longer appropriate for any of the subject areas. Panelists who had been selected were informed that the vertical articulation was cancelled, and they were not required to attend the workshop for that given day. Because of the very high consistency with the current cut scores identified in 2022, KDE determined that the current cut scores would continue to be used for all KSAs without any adjustments.

6. Item Analyses

Item statistics are crucial for maintaining the integrity of an assessment program, primarily to help test developers construct test forms that provide appropriate information about student achievement. More specifically, item statistics are used to select test items that are appropriate in difficulty, differentiate between students who have and who not mastered the content, and are fair to all students. As mentioned in Section 2.4.3, several statistical indices are used to judge the appropriateness of using items on a test form. This chapter discusses the statistical indices used in judging the quality of items for the KSA assessments.

6.1. Item Mean Scores

Item difficulty denotes how successful students, as a group, are on items. For multiple-choice items, the *p-value* is used to define the proportion of students who answered an item correctly. Although the *p*-value is commonly represented as a proportion, it is often referred to as a "percent." As an example, an item with a *p*-value of 0.55 indicates that 55% of students who responded to that item answered it correctly. This index can also be thought of as the average item score when considering that a correct response is symbolized as '1' and an incorrect response is symbolized as '0'. For constructed-response items, the average item score across a group of students provides the same information of item difficulty. For example, an item with a maximum score of 4 points may have a mean value of 2.13, which is the average item score from all students that attempted that item. In this case, students could obtain scores of 0, 1, 2, 3, or 4 depending on the alignment between the item response and scoring criteria used for these items.

Appendix C of the *Yearbook* presents item difficulties from the KSA assessments. To cover the range of students' skill level, test items should range from easy to difficult with a concentration toward the middle of the continuum. The *Yearbook* includes the single-point item difficulties by *p*-value ranges, including the average *p*-value for all items, for each grade and subject area. The *Yearbook* also contains summaries of item difficulty for the multi-points items.

6.2. Item-Test Score Correlations

Judging items' appropriateness for testing goes beyond the difficulty level of the items; the items must also differentiate between students who have mastered the content and those who have not. Correlations between item score and total test scores are used to evaluate how well items *discriminate* between "high" and "low" proficiency students. In general, the higher the correlation, the better an item is at discriminating among high- and low-proficiency students. Another way of looking at this index is that higher correlations mean that students who should have answered the item correctly, based on their total test score, did answer the item correctly, whereas students who should not have answered this item correctly did not. This is a general expectation, given that some students will answer an item correctly by chance.

Given the nature of correlations, this statistical index has a theoretical range of -1.0 to +1.0, although values do not reach the extreme ends of this range. When the correlation is negative or near zero, the item does not discriminate well, which may lead to further investigations of the item. Appendix D of the *Yearbook* presents summaries of the item-test score correlations for the single-point and multi-points items, including the median correlation across all items, for each grade and subject area.

In addition to the correlation between item score and total test score, each multiple-choice answer option can be compared against the total test scores. Although not provided in the *Yearbook*, the option-test score correlation treats each answer option separately as the "correct" response and is the relationship between the option *p*-value and total test scores. The option-test score correlation for the item's true correct response will be the same as the item-test score correlation.

With this statistic, it is assumed that the option-test score correlation for each incorrect answer option (i.e., distractor) will be lower than that of the correct answer. In fact, the correlation for the distractors should be less than 0 because students who answer an item incorrectly should have lower test scores than those who answered the item correctly. However, a distractor correlation may be positive (slightly above 0), indicating that even students with higher test scores chose that wrong answer. Positive correlations for item distractors may indicate that something is systematically causing students to choose the incorrect answer option. In this case, the item's content and answer option should be reviewed.

6.3. Differential Item Functioning

During item development, items are reviewed for potential bias against any student subgroup (e.g., gender, ethnicity, disability). Items that are identified as displaying potential bias are either revised or removed from consideration for future use. Once items have been field tested, statistics are often computed and used to call to attention items in which subgroups of students performed significantly different from each other. In other words, an item may show that males outperformed females and that the difference may be more than just a chance occurrence.

DIF exists when an item appears to favor one subgroup or present a disadvantage to another group after students across both groups have been matched on proficiency. In DIF procedures, the subgroups of interest are categorized into two groups: focal and reference groups. The focal group is the group of interest; the reference group is the group to which the focal group is compared to. For example, in gender DIF analyses, females are the focal group and males are the reference group; in ethnicity DIF analyses, African Americans are a focal group, and the White subgroup is the reference group. DIF analyses on ethnicity can be extended to other ethnic groups to represent the focal group and comparing them each to the White subgroup. Because students are matched on proficiency across focal and reference groups, statistical differences found between the groups are not confounded by student proficiency.

DIF for the KSA assessments is analyzed by a statistical procedure based on the Mantel-Haenszel chi-square statistic (M-H χ^2) for multiple-choice items (Holland & Thayer, 1988). The chi-square statistic determines whether the odds of a correct response on an item is the same for both focal and reference groups across all levels of proficiency. The Mantel-Haenszel odds ratio (α_{M-H}) is the odds of a correct response of the reference group divided by the odds of a correct response of the focal group. Data for these Mantel-Haenszel procedures are drawn from 2×2×k (score levels) contingency tables for each item. As shown in Table 6.1, the number of focal and reference group members scoring in each possible item response is captured.

Table 6.1. Item 2×2 Contingency Table for the *k*th Score Level

	Item Score		
Group	*Correct (1)*	*Incorrect (0)*	*Total*
Focal (f)	n_{f1k}	n_{f0k}	n_{fk}
Reference (r)	n_{r1k}	n_{r0k}	n_{rk}
Total (t)	n_{t1k}	n_{t0k}	n_{tk}

For classifications of DIF, the Mantel-Haenszel Delta DIF statistic (MHD; Dorans & Holland, 1993) is computed from the Mantel-Haenszel odds ratio and used in conjunction with M-H χ^2 to classify items into three categories distinguishing magnitudes of DIF: negligible DIF (A), moderate DIF (B), and large DIF (C).

Classification is based on the following guidelines:

- M-H χ^2 not significantly different from 0 or |MHD| less than 1 results in a classification of *A*.
- M-H χ^2 significantly different from 0 and |MHD| value at least 1 but less than 1.5 ***or*** M-H χ^2 not significantly different 0 and |MHD| greater than 1 results in a classification of *B*.
- M-H χ^2 significantly different from 0 and |MHD| at least 1.5 results in a classification of *C*.

In addition to these classifications, notation of DIF includes a positive (+) sign, indicating that the item favors the focal group, or a negative (-) sign, indicating that the item favors the reference group. Items designated with B or C DIF classifications are recommended for review before continued use on assessments, although caution must be exercised when analyzing DIF to prevent over-interpretation of the statistics.

The *standardized mean difference* (SMD; Zwick et al., 1993) procedure is used for detecting DIF for constructed-response items. A summary statistic, SMD is used as an effect size estimate comparing the mean item score between the reference and focal groups. Although the numerical result of this statistical procedure is different from the M-H statistics, the classification of the results is the same—the results are classified into three categories indicating the magnitude of DIF with additional notation indicating the favored group.

Appendix E of the *Yearbook* presents the number of items flagged for DIF through three student subgroup comparisons: Male-Female, White-Black, and White-Hispanic. During test construction, classifications of DIF from prior test administrations are available for most items chosen for test forms. When items previously flagged for DIF are chosen for operational test forms, content specialists review these items to determine whether the item content lends itself to DIF. All items, however, are examined for fairness at the time of item development, presented at bias and sensitivity committee reviews prior to field testing (see Chapter 2). Items judged as having bias within the content, regardless of the point when item bias is judged, are not used for testing.

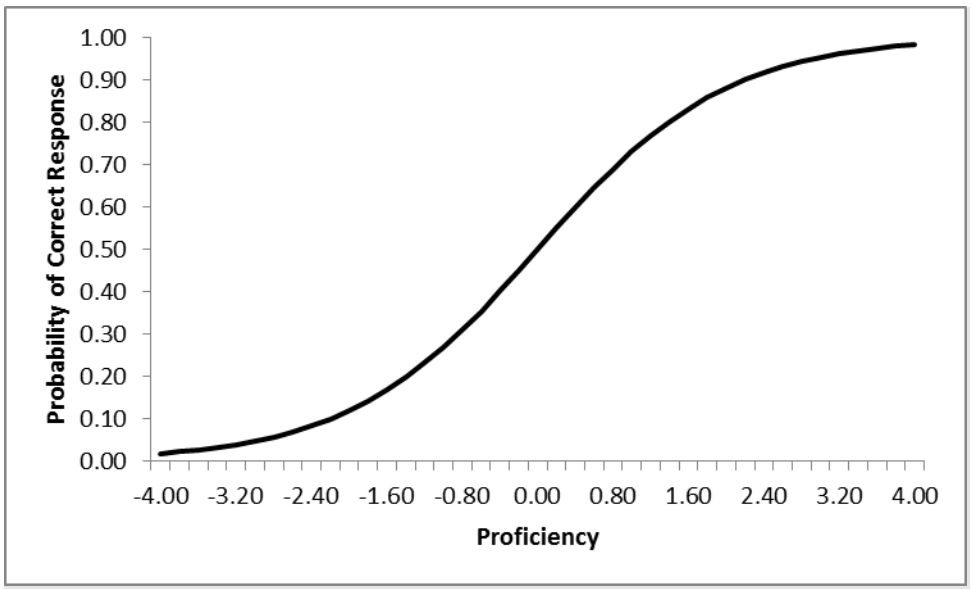
6.4. Item Response Theory

Item response theory (IRT) is a measurement framework that analyzes test item properties and item responses simultaneously. Measurement models under IRT specify the probability of a correct response to an item dependent upon proficiency and item characteristics. The simplest IRT model is the *one-parameter logistic* (1PL) measurement model (Rasch, 1980), represented as:

$$P_i(\theta) = \frac{e^{(\theta - b_i)}}{1 + e^{(\theta - b_i)}},$$

where $P_i(\theta)$ is the probability that a student with proficiency θ answers item i correctly, b_i is the difficulty of item i, and e is the base of natural logarithms with an approximate value of 2.718. This equation above specifies the probability of a correct answer to an item with a particular difficulty for a person with a particular proficiency. Figure 6.1 presents a graphical display of the 1PL model for an item.

Figure 6.1. Graph of 1PL Model



However, this model only applies to multiple-choice items. Given that KSA includes constructed-response items, a separate model is required for estimating proficiency and item difficulty simultaneously for these items. In IRT, the item difficulty is different from the item mean score discussed in Section 6.1. The item difficulty is represented on a *logit scale* with a typical range of -2.0 to +2.0. Item difficulty values near -2.0 indicate very easy items, while values near +2.0 indicate very difficult items.

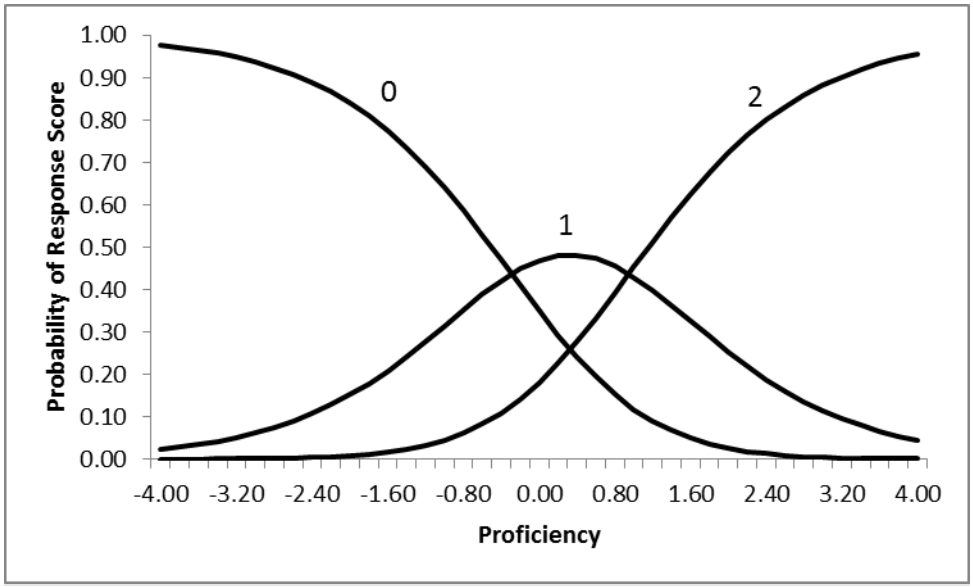
The Partial Credit Model (PCM; Masters, 1982) is an extension of the 1PL model to items that contain multiple steps in the solution process. The PCM can be written as:

$$P_{ix}(\theta) = \frac{exp\left[\sum_{j=0}^{x}(\theta - \delta_{ij})\right]}{\sum_{r=0}^{m_i}\left[exp\sum_{j=0}^{r}(\theta - \delta_{ij})\right]},$$

where $P_{ix}(\theta)$ is the probability that a student with proficiency θ responds in category x on item i with m steps, and δ_{ij} is the *step difficulty* associated with category j of item i (j=1,…, m).

The difference between the 1PL model and PCM is that PCM has multiple difficulties associated with an item as opposed to the single item difficulty in the 1PL model. However, the difficulties in PCM represent the difficulty in transitions from one score category to the next. An item with three score categories (e.g., 0 to 2 points) would have two transitions, or steps: score 0 to score 1 (δ_{i1}) and score 1 to score 2 (δ_{i2}). Figure 6.2 displays score category response curves under PCM for a 3-point item. In this graph, the intersection of response category curves 0 and 1 and the intersection of response category curves 1 and 2 indicate the difficulty of transitions from one score category to the next.

Figure 6.2. Graph of Partial Credit Model for 3-Point Item



In addition to item difficulty, IRT provides other indices for item analyses, such as item fit. Item fit analyses evaluate how well the IRT model(s) used for item analysis explains the responses to items. In the case of KSA, it is how well the 1PL model and PCM explain the response patterns of the items. The underlying investigation compares observed and expected item response patterns after the item parameters have been estimated.

Item fit for KSA is investigated through *mean-square* fit statistics that provide evidence on how well the pattern of observed responses are predicted by the 1PL and PCM measurement models. *Outfit* mean-square statistics are influenced by unexpected response patterns to items far from a student's proficiency measure. *Infit* mean-square statistics are influenced by unexpected response patterns to items near a student's proficiency measure. Linacre (2011a) provides a classification of fit mean-square estimates useful for interpretation, as shown in Table 6.2.

Table 6.2. Criteria for Item Fit Statistics

Mean-Square	Interpretation
> 2.0	Distorts or degrades the measurement system
1.5 – 2.0	Unproductive for construction of measurement, but not degrading
0.5 – 1.5	Productive for measurement
< 0.5	Unproductive for measurement, but not degrading; may produce misleadingly good reliabilities and separations.

Mean-square values near 1.0 indicate little distortion of the measurement system, while values less than 1.0 indicate observed response patterns that are too predictable (model overfit). Values greater than 1.0 indicate unpredictable observed response patterns (model underfit).

Figure 6.3 shows observed (×) and expected (□) performance on an item near average difficulty with infit and outfit indices near 1. The observed item response pattern nearly matches the expected item response patterns given the Rasch measurement model. Figure 6.4, however, shows observed and expected performance on a difficult item with an infit index near 1, but an outfit index near 1.5. In this case, the observed response patterns on the lower end of the scale influenced the outfit index.

Figure 6.3. Observed and Expected Performance on Item of Average Difficulty

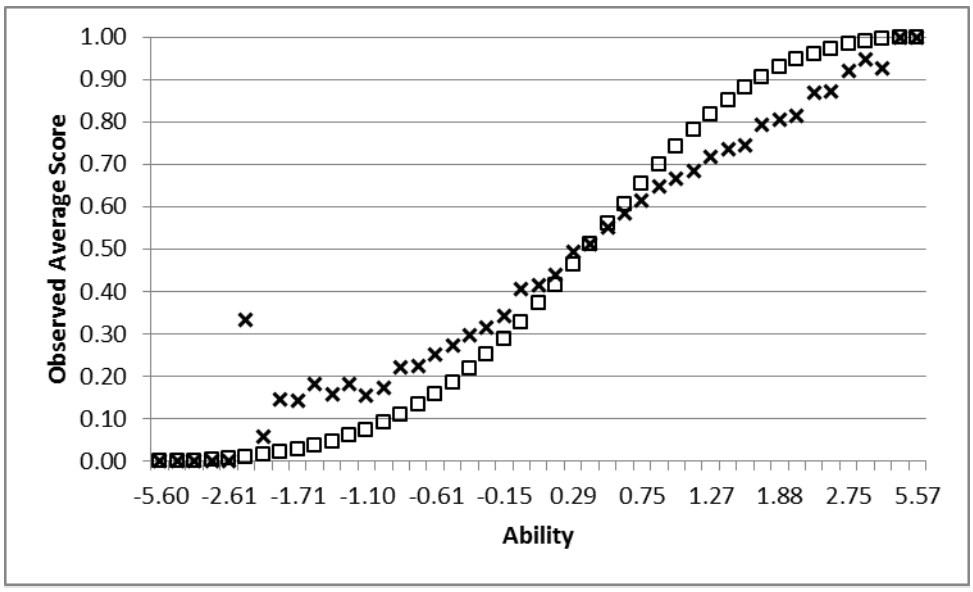
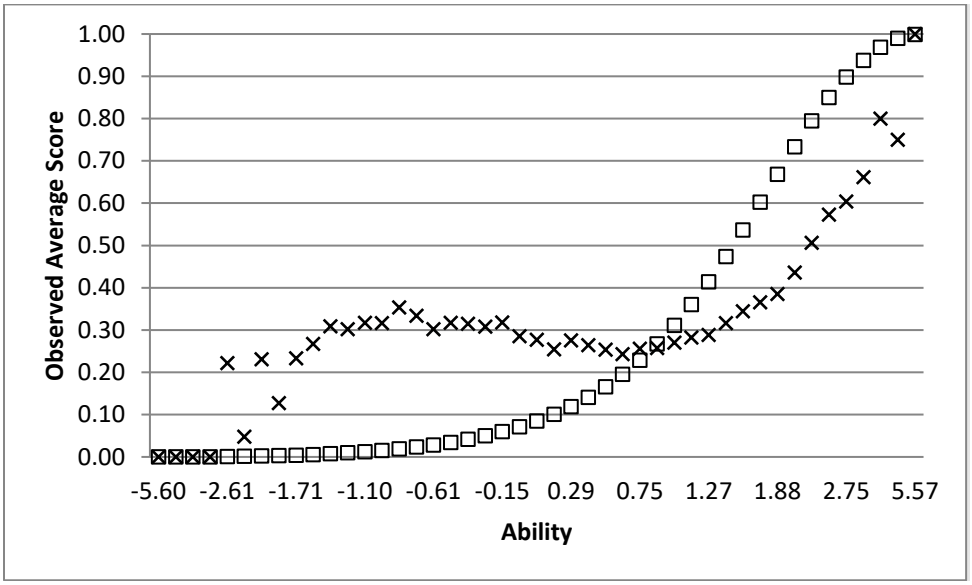


Figure 6.4. Observed and Expected Performance on Difficult Item



Appendix F of the *Yearbook* summarizes the IRT parameter estimates (i.e., item difficulty and item fit).

7. Calibration, Equating, and Scoring

Total test scores for students are often the sum of the correct responses and/or the points achieved on constructed-response items. These *raw scores* provide a simple and meaningful way to summarize a student's performance on a test. Students can also be ranked based on their test performance using the raw scores, and group statistics can be computed (e.g., average, standard deviation) and interpreted. However, raw scores can be of limited value when comparing across test forms.

Large-scale assessment programs typically construct new test forms year-to-year to prevent overexposure of test content and maintain a thorough coverage of curriculum across years. The test forms constructed across years are designed to reflect the same level of difficulty and content, even though the set of items is different across forms. However, no test form has the same level of difficulty as other test forms of similar content, so statistical processes are used to account for the differences. Part of the statistical process is a transformation of raw scores to a metric that allows comparisons of test scores across test forms of similar content. This chapter discusses the item calibration, test equating processes, and score transformations of the KSA assessments.

7.1. Measurement Models

The Rasch 1PL model and PCM were introduced in Section 6.4 to discuss the item parameters estimated under the IRT measurement framework. These models are revisited here in the context of the estimated person proficiency parameters, θ. Under IRT, a proficiency estimate is generated for each student based on their response patterns and the simultaneous estimation of the item parameters. The item and proficiency parameters are on the same logit scale, although the proficiency parameter often results in a wider range of values.

Under Rasch modeling, there is one-to-one correspondence of proficiency parameter to raw score value. In other words, for each possible raw score (total test score) value, there is one person proficiency parameter estimated. For example, if there are 40 raw score points possible on a test, there will be 41 proficiency estimates, one for each raw score (including 0). The proficiency estimates will also increase from the lowest to highest value in relation to the ascending order of the raw scores.

Problems arise in the proficiency estimation for 0 and perfect scores. Proficiency estimates are determined through a maximum likelihood function of the likelihood of proficiency for a student given all item responses. The maximum likelihood cannot be determined in the cases of all-correct or all-incorrect items responses, as the likelihood function continues toward infinity. Therefore, an adjustment (e.g., 0.25) is made to 0 and perfect raw scores so that the maximum likelihood function can result in a proficiency estimate.

7.2. Process

Pearson performed item calibrations to obtain the Rasch item parameters and proficiency estimates for the KSA assessments, and HumRRO performed an independent execution of the analyses as a third-party verifier of the process and results. Pearson created analysis specifications that outlined the process and methodology for scaling the KSA assessments, including timelines, file and document locations, and process checkpoints during which Pearson, HumRRO, and KDE would verify results and discuss any immediate concerns. During the analysis process, a

conference call was held each day to discuss progress and address any concerns before moving forward.

The process used approximately the entire testing population of KSA, although exclusion rules were applied to remove students who did not use the standard test form during assessment. The exclusion rules applied to students who use accommodated test forms (e.g., large print, audio, or Braille) or paper test forms. In the case of Braille students, some test items are considered not appropriate for Braille reproduction and were there removed from administration and scoring for those students. Content specialist reviewed the removal of such items and confirmed it did not affect the blueprint coverage of Braille forms. As a result, separate analyses may be conducted for Braille students due to the difference in maximum test score. The spring 2023 KSA administration had Braille exclusions for the Reading tests.

Prior to item calibrations, student data are inspected to identify items that potentially may have been scored incorrectly. Items' average scores (*p*-values) and item-total correlations are computed and judged to identify potential mis-keyed items. Items flagged during this analysis are reviewed for their correct answer. If an item is found to be scored incorrectly, the proper adjustment is made, and the scoring process is reinitiated. The scaling analysis depends on accurately scored student data, and all items must be considered to have been properly scored prior to analysis.

Student response data is analyzed through Winsteps Version 3.73 (Linacre, 2011b), a Rasch modeling statistical software. Each KSA assessment is analyzed separately through this software. The output from this process includes item parameters (difficulty) and proficiency estimates, both on a logit scale. The proficiency estimates are used to derive *scaled scores* for performance comparisons across test forms.

Equating is the statistical process by which scores on test forms are adjusted so that scores on the forms can be used interchangeably (Kolen & Brennan, 2004). Once equating has been performed across two or more test forms, the difference in difficulty across forms no longer confounds the comparison of performance across forms (i.e., scores from different forms may be directly compared).

Equating test forms can be accomplished in many ways. One method used in large-scale assessments is the common-item nonequivalent groups design (Kolen & Brennan, 2004). This method is used to equate alternate test forms across two different testing occasions with two different testing populations. This is accomplished using a set of common items included on both forms. The testing populations are considered nonequivalent as they do not consist of the same students taking both forms. The equating result is a scale transformation that accounts for differences in difficulty across two (or more) test forms. The result is that scores from both test forms exist on a single scale. Except for On Demand Writing, this method is used for all subjects.

For On Demand Writing tests, there is no overlap across the writing test forms. Students only took one test form which could be either anchor form or non-anchor form. Anchor forms are intact forms from the previous administration tested again in 2023 to maintain the scale. The testing populations for each form are considered equivalent since the test forms are randomly assigned to students through a spiral process. The equating result is a scale transformation that accounts for differences in

ability scale across the anchor forms and non-anchor forms. The rest of this section describes the equating process for the KSA assessments, as conducted by Pearson.

7.2.1. Linking Items

Part of the design of the equating process is the selection of common items from the test form to which equating will be performed. For equating analyses, items are chosen from previous test forms. Choosing common items requires attention to various item characteristics, both contextually and statistically. Although not presented here, guidelines for choosing common items are presented to test form developers so that these linking *sets* represent a robust subset (i.e., mini version) of the overall test. Linking items are chosen to best represent the range of item difficulty while adhering to the content distribution of the blueprint.

For the KSA tests (except for On Demand Writing tests), the anchor items set was expanded to include all 1- and 2-point items previously administered in the 2022 spring administration as linking items. The benefit of expanding the anchor items set is to minimize item parameters drift that might happen from 2022 to 2023 administration and stabilize the operational scale. For the On Demand Writing tests, two anchor forms were selected on each grade level. On each anchor form, all traits are used as linking items. Table 7.1 presents the distribution of the linking items by item type.

Table 7.1. Number of Linking Items by Item Type in the 2023 KSA tests

Subject	Grade	Multiple-Choice/ Technology Enhanced	Multi-Select/ Short Answer	Extended-Response
Reading	3	40	12	–
	4	39	2	–
	5	36	7	–
	6	34	6	–
	7	39	6	–
	8	42	6	–
	10	36	5	–
Mathematics	3	54	6	–
	4	41	4	–
	5	40	7	–
	6	45	4	–
	7	45	4	–
	8	44	7	–
	10	43	7	–
Science	4	32	10	–
	7	29	6	–
	11	37	5	–
Social Studies	5	27	6	–
	8	26	1	–
	11	41	3	–
Editing and Mechanics	5	8	5	–
	8	16	10	–
	11	8	5	–
Writing	5	–	–	10
	8	–	–	12
	11	–	–	12

7.2.2. Analysis

Post-equating analysis is performed by Pearson and an independent contractor of KDE using analysis specifications created and maintained by Pearson. Four process checkpoints were implemented for verification across the independent replications: (a) initial calibration item parameters; (b) Robust Z statistics for linking item analysis; (c) Equating constant for linking non-anchor forms to anchor forms of On Demand Writing tests; (d) final (equated) item parameters; and (e) raw-score-to-scale-score (RSSS) conversion tables.

These checkpoints represent the five main steps in the analysis process:

1. Calibrate the items through Winsteps software (Linacre, 2011b) using student item response data.
2. For all tests except for On Demand Writing, perform item stability analysis of linking items using Robust Z statistical methodology (Huynh, 2000; Huynh & Rawls, 2009; Huynh & Meyer, 2010) and drop linking items deemed unstable through this statistical index.
3. Use stable linking items as the anchor scale to produce equated item parameters for non-linking operational items.
4. For On Demand Writing, perform an iterative process to center the theta scale of the non-anchor form on the theta scale of the anchor forms by applying the equating constant to the non-anchor form, calculated as the average anchor theta ability minus the average theta ability of the non-anchor form. The initial equating constant will be added to the freely calibrated item parameters of the non-anchor forms. Then anchored item calibration is conducted to the non-anchor form with the adjusted item parameters. A non-anchor form is equated when the difference of the average ability between the non-anchor form and anchor forms is less than 0.001.
5. Produce score conversion tables, including scaled score transformations.

The Robust Z statistical procedure is used to determine if student performance remains stable on items administered across test administrations. If student performance on specific items changes substantially across test administrations when compared to the overall set of linking items, those items are not appropriate for equating one test form onto the other. The criterion for removing linking items is that the robust-Z value is greater than 1.645 (flagged for drift). One anchor item with the largest absolute robust-Z was removed during each iteration. Note that not all anchor items flagged for drift will be removed from post-equating if more than 20% of the anchor items are flagged for drift. When more than 20% of the anchor items are flagged for drift, a set of criteria including ratio of standard deviation (in the range of 0.9-1.1) and correlation (>0.95) of banked item parameters and current calibrated item parameter estimates of anchor items are examined to force anchor items with less drift back into the final linking set until the proportion of removed anchor items is no more than 20%. Each linking set is tested through this procedure. Although items may be considered unstable for equating, they remain as scored items for students' test score.

Table 7.2 presents the total number of unstable linking items dropped and the evaluation summary of the remaining linking items for the 2023 KSA tests; this table excludes On Demand Writing. For 2023, the majority of linking items were considered to be stable and kept in the final equating analyses. These linking items were used to produce equated parameter estimates of non-linking items. These item parameter estimates are produced through item calibration with Winsteps, like the initial step of the analysis, but with the linking items used as an anchor scale.

Table 7.2. Unstable Linking Items Dropped During the Robust Z Procedure

Subject	Grade	No. of Linking Items Dropped	Item Type		Reduced Linking Set Statistics		
			Multiple-Choice/ Technology Enhanced	Multi-Select/ Short Answer	SD Ratio	Correlation	% of Linking Items Remaining
Reading	3	7	4	3	0.9623	0.9681	87.0
	4	5	4	1	0.9656	0.9943	88.0
	5	6	5	1	0.9778	0.9890	86.0
	6	8	6	2	1.0369	0.9944	80.0
	7	9	8	1	0.9941	0.9964	80.0
	8	9	7	2	0.9675	0.9884	81.0
	10	7	6	1	1.0144	0.9966	83.0
Mathematics	3	6	5	1	0.9876	0.9946	90.0
	4	7	6	1	0.9943	0.9951	84.0
	5	9	7	2	1.0016	0.9981	81.0
	6	3	2	1	1.0036	0.9921	94.0
	7	9	7	2	0.9999	0.9956	82.0
	8	6	3	3	0.9855	0.9911	88.0
	10	3	2	1	0.9905	0.9912	94.0
Science	4	5	4	1	0.9795	0.9923	88.0
	7	6	5	1	1.0392	0.9931	83.0
	11	8	7	1	1.0756	0.9859	81.0
Social Studies	5	6	5	1	0.9827	0.9936	82.0
	8	4	4	0	0.9606	0.9806	85.0
	11	6	6	0	0.9737	0.9817	86.0
Editing and Mechanics	5	2	0	2	1.0069	0.9976	85.0
	8	5	1	4	1.0252	0.9926	81.0
	11	2	0	2	1.0512	0.9921	85.0

7.2.3. Quality Control

HumRRO executed the calibration and scaling analyses as a third-party verifier using the analysis specifications created by Pearson. Prior to the analysis, Pearson coordinated a *dry run* execution of the analysis process with HumRRO so that both groups can prepare and execute program codes using mock data. The dry run allowed Pearson and HumRRO to discuss processes ahead of the live analysis, including verification of software versions.

Pearson provided all the necessary item and student data files to HumRRO at the time the files were available. As the third-party verifier, HumRRO compared analysis results with those obtained by Pearson and provided feedback on the comparison. Pearson, HumRRO, and KDE also participated in a conference call each day during the analysis to share general impressions and discuss any concerns with the current results. To use the daily conference call effectively, Pearson proposed a schedule of analysis such that Pearson and HumRRO would perform the same analyses concurrently to be able to address any issues and concerns immediately (during the conference calls).

As part of the feedback on the replications, HumRRO provided outputs detailing the comparisons of results. These outputs are stored internally by both Pearson and HumRRO as documentation of the verification process.

7.3. Scaled Scores

Scaled scores can be derived through either linear or nonlinear transformations of the raw scores. For KSA, the scaled scores are derived through linear transformations of the respective IRT theta metric for a given subject area and grade using the following general form:

$$SS = m\theta + b,$$

where *m* is the slope, *θ* is the IRT person proficiency estimate obtained through the calibration (Winsteps), and *b* is the intercept. Using this equation, a scaled scored can be computed for each raw score possible, given the correspondence of raw score to proficiency estimate (*θ*) from Rasch modeling of student response data. The scaled score metric for the KSA assessments was chosen to range from 400 to 600 where the slope (*m*) was set to 16.67, the intercept (*b*) was set to 510, and θ is the person proficiency estimate, with the exception of On Demand Writing where the slope (*m*) was set to 5 and the intercept (*b*) set to 510.

Scaled scores for each domain (i.e., reporting category) of each subject area were also computed to help illustrate students' specific strengths and weaknesses. These were transformed on the same metric as individual student scores and used for aggregate summary information at the school, district, and state levels. More specifically, student scores were aggregated across these levels to provide indices of how each aggregate level compared with the others on each domain.

The scaled score system was created to indicate student performance in line with the state performance standards and as articulated by the PLDs. Performance levels are the best indicators to use for comparing performance across grades or subjects. Using scaled scores in this way provides a meaningful context for assessing achievement. Table 7 presents the scaled score ranges for each KSA performance level—*Novice*, *Apprentice*, *Proficient*, and *Distinguished*.

Table 7.3. Scores by Performance Level

Subject	Grade	*Novice*	*Apprentice*	*Proficient*	*Distinguished*
Reading	3	400–499	500–512	513–527	528–600
	4	400–502	503–515	516–530	531–600
	5	400–506	507–521	522–537	538–600
	6	400–503	504–517	518–531	532–600
	7	400–500	501–511	512–525	526–600
	8	400–503	504–514	515–527	528–600
	10	400–500	501–512	513–528	529–600
Mathematics	3	400–504	505–520	521–541	542–600
	4	400–506	507–520	521–542	543–600
	5	400–498	499–514	515–536	537–600
	6	400–494	495–506	507–525	526–600
	7	400–495	496–504	505–521	522–600
	8	400–494	495–504	505–523	524–600
	10	400–493	494–503	504–520	521–600
Science	4	400–494	495–514	515–530	531–600
	7	400–491	492–509	510–528	529–600

Subject	Grade	*Novice*	*Apprentice*	*Proficient*	*Distinguished*
Science	11	400–492	493–512	513–532	533–600
Social Studies	5	400–503	504–515	516–529	530–600
	8	400–502	503–513	514–527	528–600
	11	400–500	501–513	514–527	528–600
Editing and Mechanics	5	400–507	508–521	522–533	534–600
	8	400–503	504–516	517–532	533–600
	11	400–503	504–520	521–537	538–600
Writing	5	400–485	486–511	512–525	526–600
	8	400–476	477–498	499–531	532–600
	11	400–469	470–504	505–536	537–600

7.3.1. Results

Appendix G of the *Yearbook* contains the derived scaled scores for each KSA assessment in tables. Each table contains the scaled scores and conditional standard error of measurement (CSEM) that represents the standard deviation of observed scores of students with the same true score, as discussed in Chapter 8: Reliability. Appendix H of the *Yearbook* provides score frequency distributions for each KSA assessment*;* Appendix I of the *Yearbook* provides descriptive statistics (mean, standard deviation, minimum, maximum) for the scaled scores for each KSA assessment for the overall testing population and by subgroups (gender, ethnicity, migrant status, economic disadvantaged or not, and accommodations). Appendix J of the *Yearbook* provides performance level distributions for each KSA assessment.

7.3.2. Considerations and Limitations

There are limitations on using scaled scores for interpreting student performance. First, the scaled scores are not on a *vertical scale*, which limits interpretations on performance differences on a subject-area test across grades. Second, scaled scores should not be used for interpreting performance differences between assessments within the same grade. Differences in scaled scores do not reflect actual differences in raw scores or proficiency estimates from which they are derived. For example, a scaled score difference of five points can be the result of a small difference in proficiency estimate. Also, differences in scaled scores within a test vary along scale.

8. Reliability

Reliability is the consistency of the results obtained from a measurement. When a score is reported for a student, the expectation is that if the student had instead taken a different but equivalent version of the test, a similar score would have been achieved. A test that does not meet this expectation (i.e., a test that does not measure student proficiency and knowledge consistently) has little or no value. Furthermore, the proficiency to measure consistently is a prerequisite to making appropriate interpretations of scores on the measure (i.e., showing evidence of valid use of the results).

8.1. Estimating Reliability

Internal consistency methods use a single administration to estimate test score reliability. For state assessments where student testing time is at a premium, internal consistency procedures have a practical advantage over reliability estimation procedures that require multiple tests. One method for computing reliability estimates is through the person ability estimates obtained when test items are calibrated to the IRT framework.

Reliability is estimated as the ratio of true score variance to observed score variance where true score variance is the observed score variance *minus* error variance. Appendix K of the *Yearbook* provides reliability estimates, using person ability estimates, for the overall testing population and by gender, ethnicity, and other student subgroups.

8.2. Standard Error of Measurement

A reliability coefficient expresses test score consistency in terms of variance ratios. In contrast, the standard error of measurement (SEM) expresses score inconsistency (i.e., unreliability). The SEM is an estimate of how much error there is likely to be in a student's observed score or, alternately, how much score variation would be expected if the student were tested multiple times with equivalent forms of the test. The SEM is calculated using the following formula:

$$SEM = s_x\sqrt{1-\rho_{XX'}},$$

where s_x is the standard deviation of the total test scores, and $\rho_{xx'}$ is a reliability estimate for the set of test scores.

8.2.1. Use of the Standard Error of Measurement

The SEM can be helpful for quantifying the extent of error in student scores due to factors unrelated to the test itself. An SEM band placed around the student's observed score would result in a range of values most likely to contain the student's true score. The true score may be expected to fall within one SEM of the observed score 68% of the time, assuming that measurement errors are normally distributed.

For example, if a student has an observed score of 45 on a test with a reliability of 0.88 and a standard deviation of 9.48, the SEM would be

$$SEM = 9.48\sqrt{1-0.88} = 3.28$$

Placing a one-SEM band around this student's observed score would result in a score range of 41.72 to 48.28 (i.e., 45 $\pm$ 3.28). Furthermore, if it is assumed the errors are normally distributed and if this procedure were replicated across repeated testing occasions, this student's true score would be expected to fall within the $\pm$1 SEM band 68% of the time (assuming no learning or memory effects). Thus, the chances are better than two out of three that a student with an observed score of 45 would have a true score within the interval 41.72 − 48.28. This interval is called a confidence interval or band. Increasing the range of the confidence interval improves the likelihood that the confidence interval includes the true score. For example, an interval of $\pm$1.96 SEMs around the observed score covers the true score with 95% probability and is referred to as a 95% confidence interval.

Appendix K of the *Yearbook* provides the SEM for the KSA assessments along with the reliability estimates. The SEM is reported for total scores for the testing population, gender, ethnicity, and other student subgroups.

8.2.2. Conditional Standard Error of Measurement

Although the overall SEM is a useful summary indicator of a test's precision, the measurement error on most assessments varies across the score range. This means the measurement accuracy of a test is likely to differ for students depending on their score. The SEM is defined as the standard deviation of the observed scores of students with a particular *true* score, or a score without any measurement error. This standard deviation is called the conditional standard error of measurement (CSEM). The reasoning behind the CSEM is as follows: If a group of students all have the same true score, a measure without error would assign these students the same score (the true score). Any differences in the scores of these students must be due to measurement error. The conditional standard deviation defines the amount of error.

True scores are not observable, so the CSEM cannot be calculated simply by grouping students by their true score and computing the conditional standard deviation. However, IRT allows the CSEM to be estimated for any test where the IRT model holds. Under the Rasch IRT model, the mathematical statement of CSEM for each person is as follows:

$$\sigma_{\hat{\theta}} = \frac{1}{\sqrt{\sum_{i=1}^{L} p_{vi}(1 - p_{vi})}},$$

where v represents a person, i represents an item, L represents the number of items on the test, $\hat{\theta}$ represents proficiency, and p_{vi} represents the probability that a person will answer an item correctly. p_{vi} is defined as follows:

$$p_{vi} = \frac{e^{\theta_v - b_i}}{1 + e^{\theta_v - b_i}},$$

where θ_v represents person $v's$ proficiency, and $b_i$represents item $i's$ difficulty.

Appendix G of the *Yearbook* provides the conditional standard errors of scaled scores are provided in the score conversion tables. The conditional standard error values can be used in the same way to form confidence bands as described for the test-level SEM values.

8.3. Scoring Reliability for Open-Ended Items

8.3.1. Reader Agreement

Pearson uses several procedures to monitor scoring reliability. One measure of scoring reliability is the between-reader agreement observed in the required second reading of (a) all On-Demand Writing test responses; and (b) a percentage of students' short answer and extended-response item responses for Reading, Mathematics, Editing and Mechanics, Social Studies, and Science. These data are monitored daily during the scoring process. Reader agreement data show the percent perfect agreement of each reader against all other readers, but they do not provide a mechanism for monitoring drift from established criteria by all readers at a particular grade level. Thus, an additional set of data, resulting from a procedure known as validity scoring, are collected daily to check for reader drift and reader consistency in scoring to the established criteria.

When scoring supervisors at Pearson identify ideal student responses (i.e., ones that appear to be exemplars of a particular score value), they route these to the scoring directors for review. Scoring directors examine the responses and choose appropriate papers for validity scoring. Validity responses are usually solid score point responses. The scoring directors confirm the score and enter the student response into the validity scoring pool. Readers score a validity response periodically throughout the scoring process. Validity scoring is blind; because image-based scoring is seamless, readers do not know when they are scoring a validity response. Results of validity scoring are analyzed regularly by Pearson's scoring directors, and appropriate actions are initiated as needed, including the retraining or termination of readers.

Appendix L in the *Yearbook* provides scoring metrics (reliability, validity, and score distributions) for constructed-response items across subject areas. Checks of the consistency of readers of the same composition is one form of inter-rater reliability. Rater agreement is categorized as perfect agreement (no difference between readers), adjacent agreement (one score point difference), or non-adjacent agreement (greater than one score point difference). More detailed information regarding the scoring process of constructed response items is provided in Chapter 10: Performance Scoring.

8.3.2. Score Resolutions

A district may appeal the score assigned to any student's composition about which a question has been raised. In these instances, Pearson provides an individual analysis of the composition in question.

8.4. Reliability of Performance Level Categorization

Every test administration results in some error in classifying students. The concept of the SEM provides a mechanism for explaining how measurement error can lead to classification errors when cut scores are used to classify students into different performance levels. For example, some students may have a true performance level greater than a cut score. However, due to random variations (measurement error), their observed test score may be below the cut score. As a result, the students may be classified as having a lower performance level. As discussed in Section 8.2, a student's true score is most likely to fall into a standard error band around their observed score. Thus, the classification of students into different performance levels can be imperfect, especially for the borderline students whose true scores lie close to the performance level cut scores.

8.4.1. Accuracy and Consistency

Accuracy refers to the extent to which achievement decisions based on test scores match those that would be made if the scores did not contain any measurement error (i.e., true scores). Since true scores are not available, an estimate of the true score distribution must be determined for classification accuracy to be estimated. Consistency, on the other hand, refers to the extent to which achievement classification decisions based on test scores match the decisions based on a second, parallel form of the same test. This index assumes that two parallel forms of the same test are administered to the same group of students. However, this is impractical in Kentucky. Livingston and Lewis (1995) developed techniques to estimate both accuracy and consistency that overcome the constraints of true scores and multiple test forms on the same students. These procedures are used to generate accuracy and consistency indices on the KSA assessments.

8.4.2. Calculating Accuracy

To calculate accuracy, a 4×4 contingency table is created for each subject area and grade. The $[x, y]$ entry of an accuracy table represents the estimated proportion of students whose true scores fall into performance level x and whose observed scores fall into performance level y. Table 8.1 is an example accuracy table where the columns represent test-based student achievement, and the rows represent true performance level decisions. In this example, the total accuracy is approximately 75%, the sum of the diagonal (shaded) cells.

Table 8.1. Example Accuracy Classification Table

	Observed Score				
True Score	*Novice*	*Apprentice*	*Proficient*	*Distinguished*	**Total**
Novice	0.117	0.034	0.000	0.001	0.152
Apprentice	0.019	0.161	0.061	0.002	0.243
Proficient	0.000	0.034	0.294	0.061	0.389
Distinguished	0.000	0.000	0.036	0.179	0.215
Total	0.136	0.229	0.391	0.243	1.000

It is useful to consider decision accuracy based on a dichotomous classification of *Novice* or *Apprentice* vs. *Proficient* or *Distinguished* because Kentucky uses *Proficient* and above as proficiency for Adequate Yearly Progress (AYP) decision purposes and as an index for tracking students' readiness to college and careers. To compute decision accuracy in this case, the table is dichotomized by combining cells associated with *Novice* with *Apprentice* and combining *Proficient* with *Distinguished*. The sum of the shaded cells in Table 8.2 indicates classification accuracy around the *Proficient* cut point of approximately 90%. The percentage of students incorrectly classified as *Apprentice* or lower, when their true score indicates *Proficient* or above, is approximately 3%.

Table 8.2. Example Accuracy Classification Table for *Proficient* Cut Point

	Observed Score				
True Score	*Novice*	*Apprentice*	*Proficient*	*Distinguished*	**Total**
Novice	0.117	0.034	0.000	0.001	0.152
Apprentice	0.019	0.161	0.061	0.002	0.243
Proficient	0.000	0.034	0.294	0.061	0.389
Distinguished	0.000	0.000	0.036	0.179	0.215
Total	0.136	0.229	0.391	0.243	1.000

8.4.3. Calculating Consistency

Consistency can be calculated in the same manner, via a 4×4 contingency table, albeit with data indicating an estimate of the joint distribution of classifications on (hypothetically) two independent, parallel test forms. Table 8.3 presents sample statistics of consistency classification. Based on this sample data, the overall consistency is approximately 67%, and the consistency at *Proficient* is 87%. The agreement rates are lower than those for accuracy because both classifications contain measurement error, whereas true score classification is assumed to be without error in the accuracy table.

Table 8.3. Example Consistency Classification Table

	Second Form				
True Score	*Novice*	*Apprentice*	*Proficient*	*Distinguished*	**Total**
Novice	0.111	0.043	0.009	0.001	0.164
Apprentice	0.019	0.147	0.073	0.004	0.243
Proficient	0.006	0.038	0.252	0.075	0.371
Distinguished	0.000	0.002	0.056	0.163	0.221
Total	0.136	0.230	0.390	0.243	1.000

8.4.4. Calculating Kappa

Another way to express overall consistency is to use Cohen's kappa (κ) coefficient (Cohen, 1960) that assesses the proportion of consistent classifications beyond chance. The coefficient is computed as follows:

$$\kappa = \frac{P - P_c}{1 - P_c},$$

where P is the proportion of consistent classifications, and P_c is the proportion of consistent classification by chance. Using Table 8.3, P is the sum of the shaded cells whereas P_c is

$$\sum_x C_{x.} C_{.x},$$

where $C_{x.}$ is the proportion of students whose observed performance level would be x on the first form, and $C_{.x}$ is the proportion of students whose observed performance level would be x on the second form. Therefore, the kappa coefficient using the data from Table 8.3 is 0.548.

Appendix N of the *Yearbook* contains a summary table of the classification accuracy and consistency indices, including kappa coefficients, for overall performance level classification and at the *Proficient* cut point for each subject area and grade.

9. Validity

Validation is the process of collecting evidence to support inferences from test results. A prime consideration in validating a test is determining if it measures what it purports to measure (i.e., if the test measures the construct of interest). During this process, several threats to validity must be considered. For example, the test may be biased against a particular group, test scores may be unreliable, students may not be properly motivated to perform on the test, or the test content may not span the entire range of the construct to be measured. Any of these threats to validity could compromise the interpretation of test scores.

Beyond verifying that the test is measuring what it is supposed to measure, it is equally important that the interpretations made by users of the test's results are limited to those that can be legitimately supported by the test. The topic of appropriate score use is discussed in Section 4.4: Cautions for Score Interpretations and Use and Section 7.3.2: Considerations and Limitations.

Demonstrating that a test measures what it is intended to measure and that interpretations of the test's results are appropriate requires an accumulation of evidence from several sources. These sources generally include expert opinion, logical reasoning, and empirical justification. What constitutes a sufficient collection of evidence in the demonstration of test validity has been the subject of considerable research, thought, and debate in the measurement community over the years. Several different conceptions of validity and approaches to test validation have been proposed, and, as a result, the field has evolved. However, more recent thinking has led to a new framework of providing validity evidence (Kane, 2006).

9.1. Argument-Based Approach to Validity

The *Standards for Educational and Psychological Testing* (AERA et al., 2014) recommends establishing the validity of a test using a *validity argument*. This term is defined in the *Standards* as "An explicit justification of the degree to which accumulated evidence and theory support the proposed interpretation(s) of test scores for their intended purposes" (p. 225).

Kane (2006), following the work of Cronbach (1988), presents an argument-based approach to validity that seeks to address the shortcomings of previous approaches to test validation. The argument-based approach creates a coherent framework (or theory) that clearly lays out theoretical relationships to be examined during test validation.

The argument-based approach given by Kane (2006) delineates two kinds of arguments: (a) the interpretative argument and (b) the validity argument. An *interpretative argument* specifies the inferences and assumptions made in the process of assigning scores to students and the interpretations made of those scores. The interpretative argument provides a step-by-step description of the reasoning (if-then statements), allowing one to interpret test scores for a particular purpose. Justification of that reasoning is the purpose of the *validity argument* that is a presentation of all the evidence supporting the interpretative argument.

The interpretative argument is usually laid out logically in a sequence of stages. For achievement tests like the KSA assessments, the stages can be broken out as *scoring*, *generalization*, *extrapolation,* and *implication*.

9.1.1. Scoring

The scoring part of the interpretative argument deals with the processes and assumptions involved in translating the observed responses of students into observed student scores. Critical to these processes are the quality of the scoring rubrics; the selection, training and quality control of scorers; and the appropriateness of the statistical models used to equate and scale test scores. Empirical evidence that can support validity arguments for scoring includes inter-rater reliability of constructed-response items and item-fit measures of the statistical models used for equating and scaling. The KSA assessment uses IRT models, so it is also important to verify the assumptions underlying these models.

9.1.2. Generalization

The second stage of the interpretative argument involves the inferences about the *universe score* made from the observed score. Any test contains a sample of the items that could potentially appear on the test. The universe score is the hypothetical score a student would be expected to receive if the entire universe of test items could be administered. Two major requirements for validity at the generalization stage are that (a) the sample of items administered on the test is representative of the universe of possible items; and (b) the number of items on the test is large enough to control for random measurement error. The first requirement entails a major commitment during the test development process to ensure that content validity is upheld and test specifications are met. For the second requirement, estimates of test reliability and the SEM are key components to demonstrating that random measurement error is controlled.

9.1.3. Extrapolation

The third stage of the interpretative argument involves inferences from the universe score to the *target score*. Although the universe of possible test items is likely to be quite large, inferences from test scores are typically made to an even larger domain. For example, not every standard and benchmark of the KSA assessments is assessed by the test. Some standards and benchmarks are assessed only at the classroom level because they are impractical or impossible to measure with a standardized assessment. It is through the classroom teacher that these standards and benchmarks are assessed. However, the KSA tests are used for assessment of proficiency with respect to all standards. This is appropriate only if interpretations of the scores on the test can be validly extrapolated to apply to the larger domain of student achievement. This domain of interest is called the target domain, and the hypothetical student score on the target domain is called the target score. Validity evidence in this stage must justify extrapolating the universe score to the target score. Systematic measurement error could compromise extrapolation to the target score.

The validity argument for extrapolation can use either analytic evidence or empirical evidence. Analytic evidence largely stems from expert judgment. A credible extrapolation argument is easier to make to the degree the universe of test questions largely spans the target domain. Empirical evidence of extrapolation validity can be provided by criterion validity when a suitable criterion exists.

9.1.4. Implication

The implication stage of the interpretative argument involves inferences from the target score to the decision implications of the testing program. For example, a college admissions test may be an excellent measure of student achievement and a predictor of college GPA. However, an administrator's decision of how to use a particular test for admissions has implications that go beyond the selection of students who are likely to achieve a high GPA. No test is perfect in its predictions, and basing admissions decisions solely on test results may exclude students who would excel if given the opportunity.

9.2. Validity Argument Evidence

The following sections present a summary of the validity argument evidence for each of the four parts of the interpretive argument: scoring, generalization, extrapolation, and implication. Much of this evidence is presented in greater detail in other chapters in this manual. In fact, most of this manual can be considered validity evidence for the KSA assessment (e.g., item development, performance standards, scaling, equating, reliability, performance item scoring, and quality control). Relevant chapters are cited as part of the validity evidence given below.

9.2.1. Scoring

Scoring validity evidence can be divided into two sections: (a) evidence for the scoring of performance items; and (b) evidence for the fit of items to the measurement model.

9.2.1.1. Scoring of Performance Items

The scoring of constructed-response items and written compositions on the KSA assessments is a complex process that requires its own chapter to describe fully, as provided in Chapter 10: Performance Scoring. The chapter's documentation of the processes of rangefinding, rubric review, recruiting and training of scorers, and quality control provides some of the evidence for the validity argument that the scoring rules are appropriate. Further evidence comes from Appendix L and M of the *Yearbook* reporting inter-rater agreement and inter-rater reliabilities. The results in those tables show both measures are generally high for the KSA assessments.

9.2.1.2. Model Fit

IRT models provide a basis for the KSA assessments and can be used for the selection of items to go on the test and the equating and scaling procedures. A failure of model fit would make the validity of these procedures suspect. Item fit is often examined during test construction. Any item displaying misfit is scrutinized before a decision is made to put it on the test. Further evidence of the fit for the IRT models comes from dimensionality analyses. IRT models for the KSA assessments assume the domain being measured by the test is relatively unidimensional. To test this assumption, a principal components analysis is performed. Appendix O of the *Yearbook* provides eigenvalues representing unexplained variance in the data. These values are obtained from the Winsteps software during the item calibration process. Any eigenvalue greater than 2 may signify a secondary dimension within the assessment.

To go along with the principal component analyses, confirmatory factor analyses were conducted to test the model of one factor construct within the KSA assessments. Indices of model fit are used to determine how well this model fits the

data. McDonald and Ho (2002) define absolute fit indices as determining how well an *a priori* model fits the sample data. The chi-square statistic assesses the magnitude of discrepancy between the sample and fitted covariance matrices (Hu & Bentler, 1999). However, this statistic is sensitive to sample size and often rejects the model when large samples are used (Bentler & Bonnet, 1980). Alternatives to the chi-square, the goodness-of-fit statistic (GFI; Jöresky & Sörbom, 1993) and adjusted goodness-of-fit (AGFI; Tabachnick & Fidell, 2007), are also sensitive to sample size, which has led to researchers reporting them along with other fit indices (Hooper et al., 2008).

The root mean square error of approximation (RMSEA), a comparative fit index, indicates how well the model would fit the population covariance matrix (Byrne, 1998). This fit index favors parsimony as it is sensitive to the number of estimated parameters in the model. Of the few suggestions of index threshold cutoffs of good fit, the most stringent criterion is 0.06 as suggested in Hu and Bentler (1999). A confidence interval can also be constructed for RMSEA, with a lower limit close to 0.0 signifying a well-fitting model, as well as an upper limit less than 0.08.

The root mean square residual (RMR) and standardized root mean square residual (SRMR) are the square root of the difference between the residuals of the sample covariance matrix and the hypothesized covariance model. The SRMR has a range of 0 to 1, with 0 indicating perfect fit. Byrne (1998) suggests well-fitting models having an SRMR less than 0.05. Hooper et al. (2008) caution that SRMR will tend to be low with a high number of parameters and models with large sample sizes. Hu and Bentler (1999) suggested a two-index presentation when reporting model fit evaluation. One proposed combination is the RMSEA, with confidence interval, and the SRMR. Appendix P of the *Yearbook* provides the estimates of these indices. These estimates provide support of the one-factor construct for the KSA assessments.

Another check for unidimensionality can be made at the item level. The content measured by each item on the test should have a strong relationship with the content measured by the other items. An item-total correlation (also called point-biserial correlation for multiple-choice items) is the correlation between an item and the total test score. Conceptually, a high item-total correlation (i.e., 0.30 or above) for an item indicates that students who performed well on the test got the item right and students who performed poorly on the test got the item wrong. In other words, the item discriminated well between high- and low-proficiency students. Assuming the total test score represents the extent to which a student possesses the construct being measured by the test, high item-total correlations indicate the items on the test require possession of this construct to be answered correctly. Appendix D of the *Yearbook* presents the item-total correlations.

9.2.2. Generalization

Two major requirements for validity allow generalization from observed scaled scores to universe scores. First, the items administered on the test must be representative of the universe of possible items. Evidence regarding this requirement comes from content validity that is documented through evidence that the test measures the state standards and benchmarks to the extent possible. Second, random measurement error on the test is controlled. Evidence that measurement error is controlled comes largely from reliability and other psychometric measures. Evidence is also presented concerning the use of the KSA assessments for different student populations.

9.2.2.1. Evidence of Content Validity

The KSA assessments are based on content standards and benchmarks along with extensive content limits that help define what is to be assessed. Committees of educators collaborate with item development experts, assessment experts, and KDE staff to review newly developed and field tested items so that tests adequately sample the relevant domain of material the test purports to cover. These review committees participate in this process to further advance test content validity for each test.

A sequential review process for committees is used by KDE as outlined in Chapter 2: Test Development. In addition to providing information on the difficulty, appropriateness, and fairness of items and performance tasks, committee members provide a check on the alignment between the items and the benchmarks measured. When items are judged to be relevant (i.e., representative of the content defined by the standards), this provides evidence to support the validity of inferences made regarding knowledge of this content from the results. When items are judged to be inappropriate for any reason, the committee can either suggest revisions (e.g., reclassification, rewording) or elect to eliminate the item from the item pool. In essence, these committees review and verify the alignment of the test items with the objectives and measurement specifications so that the items measure the expected content. The nature and specificity of these review procedures provide strong evidence for the content validity of the test.

As discussed in Chapter 2, Pearson works with trained item writers to write items specifically to measure the objectives and specifications of the content standards for the tests. Many different people with different backgrounds write the items, preventing bias that might occur if items were written by a single author. The input and review by these assessment professionals provide further support of the item being an accurate measure of the intended objective.

9.2.2.2. Evidence of Control of Measurement Error

Reliability and the SEM are discussed in Chapter 8: Reliability. Appendix G of the *Yearbook* has tables reporting the conditional SEM for each scaled score point, and Appendix K of the *Yearbook* provides the reliability estimates. Further evidence is supplied to demonstrate that the IRT model fits the data well. Item-fit statistics and tests of unidimensionality also apply here, as they did in the section describing evidence argument for scoring. Appendices O and P of the *Yearbook* provide the results of these analyses.

9.2.2.3. Validity Evidence for Different Student Populations

It can be argued from a content perspective that the KSA assessments are not more or less valid for use with one subpopulation of students relative to another. The assessments measure the statewide content standards that are required to be taught to all students. In other words, the tests have the same content validity for all students because what is measured is taught to all students, and all tests are given under standardized conditions to all students. Every effort is made to eliminate items that may have ethnic or cultural biases. As described in Chapter 2, item writers are trained on how to avoid economic, regional, cultural, and ethnic biases when writing items. After items are written and passage selections are made, committees of Kentucky educators are convened by KDE to examine items for potential subgroup bias. Items are further reviewed for potential bias by Pearson and KDE after field test data are collected.

9.2.3. Extrapolation

Validity for extrapolation requires evidence that the universe score is applicable to the larger domain of interest. Although it is usually impractical or impossible to design an assessment measuring every concept or skill in the domain, it is desirable for the test to be robust enough to allow some degree of extrapolation from the measured construct. The validity argument for extrapolation can use either analytical evidence or empirical evidence. The argument for KSA uses analytical evidence.

The standards create a common foundation to be learned by all students and define the domain of interest. As documented in this manual, the KSA assessments are designed to measure as much of the domain defined by the standards as possible. Although a few benchmarks from the standards can only be assessed by the classroom teacher, most benchmarks are assessed by the test. Thus, it can be inferred that only a small degree of extrapolation is necessary to use test results to make inferences about the domain defined by the standards.

The use of different item types also increases the validity of the KSA assessments. The combination of multiple-choice, short answer, and extended-response items results in assessments measuring the domain of interest more fully than if only one type of response format was used.

9.2.4. Implication

Inferences are made at different levels based on the KSA assessments. Individual student scores are reported, as well as aggregate scores for schools and districts. Inferences at some levels may be more valid than those at others. For example, the KSA assessments report individual student scores, but some students may feel that few ramifications of the test directly affect them; such students may fail to put forth their full effort. Although this manual documents evidence showing that the KSA assessments are valid measures of student achievement on the standards, individual and school-level scores are not valid if students do not take the test seriously.

One index of student effort is the percentage of blank or off-topic responses to constructed-response items and written compositions. Because constructed-response items require more time and cognitive energy, low levels of non-response on these items provide evidence of students giving their full effort. Appendices L and M of the *Yearbook* includes non-response rates for the short answer and extended-response items.

One of the most important inferences to be made concerns the student's proficiency level, especially for accountability tests like the KSA assessments. Even if the total correct score can be validated as an appropriate measure of the standards, it is still necessary that the scaling and performance level designation procedures be validated. Because scaling and standard setting are both critical processes for the success of the Kentucky assessments, separate chapters are devoted to them in this manual. Chapter 5 discusses the details of setting and validating performance standards, and Chapter 7 discusses scaling. These chapters serve as documentation of the validity argument for these processes.

At the aggregate level (school, district, or state), the implication validity of school accountability assessments like the KSA assessments can be judged by the impact the testing program has on the overall proficiency of students. Validity evidence for this level of inference will result from examining changes over time in the percentage

of students classified as proficient. There exists a potential for negative impacts on schools as well, such as increased dropout rates and narrowing of the curriculum. Future validity studies need to investigate possible unintended negative effects as well.

9.3. Summary of Validity Evidence

Validity evidence is described in this chapter as well as other chapters of this technical manual. In general, validity arguments based on rationale and logic are strongly supported for the KSA assessments. The empirical validity evidence for the scoring and the generalizability validity arguments for KSA is also quite strong. Reliability indices, model fit, and dimensionality studies provide consistent results, indicating that the KSA assessment is properly scored, and scores can be generalized to the universe score.

10. Performance Scoring

Some items on the KSA assessments require students to construct their own response. For example, students may be required to provide a short, written response to demonstrate the application of a mathematical formula or a scientific concept. The KSA tests include short answer and extended-response items, in addition to multiple-choice items, to tap higher-order thinking skills. Short answer items are designed such that students can respond in a few words to a small number of sentences, whereas extended-response items are designed so that students may respond completely in no more than one page. For On-Demand Writing, students are required to write an essay based on a given prompt. Except for accommodations, all constructed-response items are delivered online and scored against a rubric by human scorers who are trained with material specific to the items tested. For example, an extended-response item on photosynthesis will have score requirements detailing the required knowledge of photosynthesis to achieve each possible score point.

Pearson's Performance Scoring Center (PSC) hires and trains scorers for the constructed-response items. Scorers review student responses and provide scores based on the requirements of the rubrics applied. The process of scoring constructed-response items is a coordinated effort that involves PSC, KDE, and hired external staff. PSC and KDE work together before, during, and after scoring the constructed-response items to fulfill standards of quality in scoring. This chapter provides a discussion of the process, including preparation of training materials.

10.1. Rubric Creation

The On-Demand Writing tasks were scored analytically with trait scoring. Grade 5 used Clarity and Coherence, Support, Sourcing, Organization, and Language/Conventions. Grades 8 and 11 used Clarity and Coherence, Counterclaims, Support, Sourcing, Organization, and Language/Conventions. The scoring rubric was created with input from multiple groups within Pearson and KDE. The rubric was used for the first time to score the field test responses from the stand-alone field test administered in fall 2020 (see Chapter 6: Item Analyses).

10.2. Rangefinding

Rangefinding is a process by which samples of students' responses from a previous test administration are selected to be used as scorer training material. In practice, the student responses are selected from the field test (i.e., the first time items are administered to students in a testing environment). Pearson scoring directors construct the training sets by selecting student responses to each constructed item that represent the range of student performance.

During this process, the scoring directors use the scoring rubric and any other item ancillary material as guides to determine the level of performance exhibited in each response. Proposed anchor and practice sets are reviewed by educators and responses approved by the rangefinding committee are used in scorer training. After rangefinding, additional practice and qualifying sets are built using the same scoring rationale agreed upon during the rangefinding meeting. The anchor set consists of multiple responses per possible score point and is arranged from low to high. The practice and qualification sets consist of a set number of randomly arranged responses.

10.3. Scoring Process

10.3.1. Recruitment

Recruiting scorers is the responsibility of Pearson, who keeps a database of individuals with scoring experience. The recruiting of scorers is done by the Pearson People Department, distributed scoring division. The number of scorers recruited for any project is based on the amount of time allocated for the scoring activity and the volume of scores to be assigned. Pearson recruits slightly more scorers than the projected need to accommodate for some attrition.

10.3.2. Training

Highly qualified scorers are essential to scoring students' responses to constructed-response items and writing prompts. Thus, the careful selection of professional scorers is critical in scoring the KSA assessments. Pearson actively seeks candidate scorers from all ethnic backgrounds to maximize the diversity of the scorer pool. Included in this pool is a core group of veteran scorers whose insight, flexibility, and dedication have been demonstrated while working on a range of assessments over time. Scoring supervisors are chosen from the pool of scorers based on demonstrated expertise in all facets of the scoring process, including strong organizational abilities and training skills. Supervisors are adept at helping scorers understand the scoring requirements of KDE.

Upon being hired, scorers sign a confidentiality agreement in which they pledge to keep all information and student responses confidential. Scorers and scoring supervisors are trained to thoroughly learn the rubric and score responses according to the scoring guides developed for KSA. At the beginning of the Kentucky scoring project, all scoring supervisors and scorers assigned to the project complete training specific to the KSA assessment. Thorough training is vital to the successful completion of any scoring assignment. Subject-specific leaders follow a series of prescribed steps so that training is consistent and of the highest quality. PSC staff develops its training materials to facilitate learning through visual, auditory, and kinesthetic channels.

Scoring supervisor training occurs first as supervisors assist in the training of scorers. A primary goal of this session is that scoring supervisors clearly understand the scoring protocols and the training materials so that all responses are scored in a manner consistent with the scores assigned to the anchor papers and according to the intentions of KDE. Scoring supervisors read and discuss the assessment items along with the rubrics used to score them. They are asked to carefully read and annotate all training materials so they can readily assist in scorer training and respond to scorers' questions during training and scoring.

Online training of scorers takes place after supervisors have been trained. The online training agenda includes an introduction to the Kentucky assessment program. It is important for scorers to understand the history and goals of the assessments and the context within which students' responses are evaluated. This gives them a better understanding of what types of responses can be expected. The scorers receive a description of the scoring criteria applied to the responses. Next, the trainers present the first item to be scored and the scoring rubric itself.

The primary goal of training is to convey to the scorers the decisions made during training, to show what type(s) of responses correspond to each score point, and to help scorers internalize the scoring protocol so they may effectively apply those

decisions. Scorers are better able to comprehend the scoring guidelines in context, so the rubric is presented in conjunction with the anchor papers. Anchor papers are the primary points of reference for scorers as they internalize the scoring rubric. There are three to four anchor papers for each score point value per item. The online training system directs scorers' attention to the score point description from the scoring guide, as well as the illustrative anchor papers, thereby enabling scorers to immediately connect the language of the scoring rubric with actual student performance.

After presentation of the anchor papers and annotations, each scorer is shown practice sets. Practice papers represent each score point and are used during training to help scorers become familiar with applying the scoring rubric. Some papers clearly represent the score point, while others are selected because they represent borderline responses. Use of these practice sets provides guidance to scorers in defining the line between score points. The final task of the training process is to review the qualification sets. Scorers must score the responses in the qualification set to successfully demonstrate their readiness for live scoring, or they are dismissed from the project.

10.3.3. Quality Control

As part of quality control, items are double-scored for score consistency analyses. All On-Demand Writing responses are double-scored, whereas 20% of responses to the constructed-response items (i.e., short answer and extended-response items) are double-scored for the other subject areas.

Validity scoring is also conducted throughout scoring. Validity responses are usually solid score point responses considered as exemplar responses. They are routed throughout the scoring queue of student responses such that they are scored by scorers in random fashion. Scorer agreement with validity responses is closely monitored via real-time reports, and disagreement with a predetermined number of validity responses can result in dismissal from the project.

A variety of reports are produced throughout the scoring process to allow scoring supervisory staff to monitor the progress of the project, the reliability of scores assigned, and individual scorers' work:

- *Daily and Cumulative Inter-Rater Reliability Reports by Item and Scorer*. These reports provide information about how many times scorers are in exact agreement, assign adjacent scores or require resolutions. The reliability is computed and is monitored daily and cumulatively for the project.
- *Daily and Cumulative Frequency Distributions*. These reports show how many times each score point is assigned to each item. The frequency distributions are produced both daily and cumulatively for the entire scoring project. This report allows scoring supervisors and subject leaders to see whether scorers tend to score consistently high or low.

With the help of the individual scorer reliability and validity reports, the scoring lead staff can closely monitor each scorer's performance. To document retraining efforts for scorers with low reliabilities, PSC maintains a *Scorer Intervention Log*. Entries on this form describe the feedback given to a scorer regarding their problematic scoring and enumerate the interventions taken. Scorers are dismissed if they have been counseled, retrained, and given every reasonable opportunity to improve and are still performing below the acceptable standard.

Appendix L of the *Yearbook* contains summaries of the inter-rater agreement rates and score point distribution for the constructed-response items (short answer, extended-response, and writing prompts). Appendix M of the *Yearbook* contains a summary of *total* scores and inter-rater agreement rates for Writing by grade.

10.4. Security

Scorers assigned to the Kentucky assessment program must sign a nondisclosure agreement before they can see any KSA test materials. All materials provided to scorers are also secured via security guidelines and infrastructure by Pearson. Finally, all operational scoring is conducted by using Pearson's image-based scoring system, a computer-based application that operates over a secure network. Each scorer must log in with a unique ID and password. Only scorers for the KSA project have access to the project materials. The image for scoring presented to scorers does not contain any identifying information about the student or the student's school or district.

11. Quality Control Procedures

Large-scale assessment programs involve constant activity from test development to score reporting. Several individuals and procedures are involved to maintain the workflow from one output to the next. It is crucial that each process consists of a quality control system that allows for system outputs to be checked and verified for accuracy before the next phase of the assessment cycle is implemented. Given the number of systems and processes put in place for an assessment cycle, the quality control systems must be constantly monitored and adjusted when the need occurs. Systems of quality control help safeguard KSA from situations that could affect the reputations of both Pearson and KDE. This chapter highlights how quality control measures are implemented throughout the assessment program.

11.1. Test Construction

Guidelines of test development are outlined in Chapter 2: Test Development, from item development to form construction. These guidelines help test developers, including content support and psychometrics, to build test forms that are defensible in terms of content representation and statistical measurement. The selection and placement of items are vetted through several reviews within Pearson and KDE. The development of forms is an iterative process of item selections as test developers strive to assemble the best selection of content (items) to judge student achievement and maintain statistical quality appropriate for the assessment.

11.2. Performance Scoring

Quality control measures are implemented throughout all phases of the performance scoring process, starting with the scorer recruiting and screening process designed to locate and employ the most highly qualified individuals available. At the beginning of each scoring project, scorers receive thorough training on the specific items and rubrics they will score, regardless of their previous scoring experience. Training is provided by individuals who, after fulfilling rigorous internal guidelines for knowledge and presentation skills, are considered qualified trainers. During scoring, scorers are constantly monitored for scoring accuracy and consistency. More details on the performance scoring process and quality control are presented in Chapter 10: Performance Scoring.

11.3. Equating

Test form equating is the process by which test forms are made equitable for within-year or across-year comparisons. Quality control for the psychometric analyses begins with the receipt of student data and continues through the review of the results:

- Student data are inspected for completeness and accuracy according to data layout specifications. Omissions and other data issues are investigated before subsequent analyses.
- Item scoring is inspected through statistical key checks that capture and compare the distribution of student responses, within each item, to predetermined criteria (e.g., minimum acceptable *p*-value and item-total correlation). Any item with statistical values below the minimum acceptable value is reviewed to verify that it was scored correctly. If an item is found to have been scored incorrectly, the item is rescored and a new student data file is produced.

- IRT analyses, including item calibrations and scaling, are performed by Pearson staff and one external third-party consultant. The results from these replications are compared for consistency, and any unexpected differences are resolved. Conference calls are also held daily during the psychometric analyses.
- A summary of the psychometric analyses is provided to KDE for review.

11.4. Scoring and Reporting

Before reporting, script and conversion programs with mock data are run to check that accurate reports are being produced. A random sample of reports are also selected during processing and checked against raw data to verify the accuracy of the actual reports. Test files are used to produce reports for the software quality assurance team to review. These mockups are sent to KDE for approval of the format and layout of the report. Once these mockups are approved, the data are checked again using production data. Data files are provided to KDE prior to the release of the score reports, which are used by KDE to confirm that the reported data are correct and prepare performance reports for release within the state.

For shipping, score reports are assembled by Pearson's pre-mailing staff. Strict quality control is observed during pre-mailing so that all score report shipments are complete. Once all score reports are assembled and quality checked, they are distributed using quality shipping procedures agreed to by KDE.

12. Glossary of Terms

Classical test theory: a measurement theory that prescribes a relationship between true score and score error in defining an observed score.

Classification accuracy: the extent to which achievement classifications from test scores match classifications if test scores contained no error of measurement.

Classification consistency: the extent to which achievement classifications from test scores match classifications from test scores of a parallel form of the same test.

Constructed-response item: a test item that requires a form of written response by the examinee.

Criterion-referenced test: a test that measures achievement according to defined criteria of mastery.

Cut point: a numerical value differentiating two categories of performance classification.

Differential item functioning (DIF): the difference in performance on an item between subgroups of students, after controlling for differences in group achievement or score level.

Equating: the statistical process of adjusted test scores across test forms so that scores on equivalent test forms can be used interchangeably.

Field test items: items used on a test for gathering performance data while not contributing to examinees' test scores.

Item response theory (IRT): the measurement theory that prescribes relationships of item difficulty and examinee proficiency for indices of test performance.

Item-test correlation: the correlation between item score and total test score.

Multiple-choice item: a test item that requires selection of response from a group of options.

Performance level: a categorization of achievement from test performance.

Performance level descriptor (PLD): a description of the performance level, outlining the knowledge and skills typical for a performance level.

***p*-value:** the proportion of correct responses to an item (for multiple-choice items).

Quartile: a group of observations representing a fourth of the total group.

Rangefinding: the process by which constructed responses from a previous test administration are selected to be used as scorer training material.

Rasch model: a measurement model that factors proficiency and item difficulty in determining probability of item success.

Raw score: the sum of points for a test or subdomain.

Regression to the mean: the statistical phenomenon describing the tendency of repeated data points to move closer to the average value.

Reliability: the consistency of results obtained from a measurement.

Scaled score: a score derived from a transformation of a raw score.

Scaling: transforming scores into meaningful and comparable units.

Standard error of measurement (SEM): a statistic in classical test theory that expresses the interval of a student's true score.

Standard setting: the process of setting cut points that delineate levels of achievement.

Subdomain: a set of knowledge and skills within a larger content space.

Test blueprint: a detailed prescription of content coverage by test form, provides the number of test items by content and subdomain levels.

Test design: a general summary of test form layout.

True score: a student's expected score resulting from multiple replications of measurement.

Universal design: the idea of making assessment content accessible to the widest possible group of examinees.

Validity: a framework for assessing the appropriateness and plausibility of intended test score use and interpretations.

Vertical scale: a metric of scores across grades from which achievement growth can be inferred.

13. References

American Educational Research Association (AERA), American Psychological Association (APA), & the National Council on Measurement in Education (NCME). (2014). *Standards for educational and psychological testing.* AERA.

Bentler, P. M., & Bonnet, D. C. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin, 88*(3), 588–606.

Byrne, B. M. (1998). Structural equation modeling with LISREL, PRELIS and SIMPLIS: Basic concepts, applications and programming. Lawrence Erlbaum Associates.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*, 37–47.

Cronbach, L. J. (1988). Five perspectives on validity argument. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 3–17). Lawrence Erlbaum.

Dorans, N. J., & Holland, P. W. (1993). DIF detection and description: Mantel-Haenszel and standardization. In P.W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp.35–66). Lawrence Erlbaum Associates.

Holland, P. W., & Thayer, D. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer and H. I. Braun (Eds.), *Test validity* (pp. 129–145). Lawrence Erlbaum Associates.

Hooper, D., Coughlan, J., & Mullen, M. R. (2008). Structural equation modelling: Guidelines for determining model fit. *The Electronic Journal of Business Research Methods*, *6*, 53–60.

Hu, L. T., & Bentler, P. N. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*(1), 1–55.

Huynh, H. (2000, June). *Guidelines for Rasch linking for PACT.* Memorandum to Paul Sandifer on June 18, 2000. Available from Author.

Huynh, H., & Meyer, P. (2010). Use of robust z in detecting unstable items in item response theory models. *Practical Assessment, Research & Evaluation*, *15*(2). http://pareonline.net/getvn.asp?v=15&n=2

Huynh, H., & Rawls, A. (2009). A comparison between robust z and 0.3-logit difference procedures in assessing stability of linking items for the Rasch model. In Everett V. Smith Jr. & Greg E. Stone (Eds.) *Applications of Rasch measurement in criterion-referenced testing: Practice analysis to score reporting.* JAM Press.

Jöresky, K., & Sörbom, D. (1993). *LISREL 8: Structural equation modeling wfith the SIMPLIS command language*. Scientific Software International Inc.

Kane, M. T. (2006). Validation. In R.L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). American Council on Education/Praeger.

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

Lewis, D. M., Mitzel, H. C., & Green, D. R. (1996). Standard setting: A bookmark approach. In D. R. Green (Chair), *IRT-based standard-setting procedures utilizing behavioral anchoring*. Symposium conducted at the Council of Chief State School Officers National Conference on Large Scale Assessment, Phoenix, AZ.

Linacre, J. M. (2011a). A user's guide to Winsteps® Rasch-model computer programs. Winsteps.com.

Linacre, J. M. (2011b). *Winsteps*® Rasch measurement computer program. Winsteps.com.

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

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*(2), 149–174.

McDonald, R. P., & Ho, M.-H.R. (2002). Principles and practice in reporting statistical equation analyses. *Psychological Methods, 7*(1), 64–82.

Mitzel, H. C., Lewis, D. M., Patz, R. J., & Green, D. R. (2001). The bookmark procedure: Psychological perspectives. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods and perspectives* (pp. 249–281). Lawrence Erlbaum Associates.

Rasch, G. (1980). Probabilistic models for some intelligence and attainment tests. University of Chicago Press.

Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics* (5th ed.). Allyn and Bacon.

Zwick, R., Donoghue, J. R., & Grima, A. (1993). Assessment of differential item functioning for performance tasks. *Journal of Educational Measurement, 30,* 233–251.

Appendix A. Passage Specifications

Reading

Pearson Reading Passage Specifications for 2019 Development – Recommendations

Overview:

In order to provide high-quality, authentic passages for the KY Summative Assessments, Pearson recommends the following specifications.

1- **Passage Source**

The majority of passages for the new KY assessment will be permissioned passages licensed through the Copyright Clearance Center. These will be supplemented by public domain passages that are relevant and accessible to students at each grade level.

Passages will include written texts as well as multimedia texts including video, audio and art.

Brian: I would think most of your SS stuff will be public domain (Library of Congress, National Archives, Gutenberg.org, etc.). If there's stuff you need from permissioned sources, make sure you have your budget figured out. It's always created as a guide—more important stuff that will allow you to assess difficult skills is worth more, even if it's a little above what's budgeted per source. Also, we're still negotiating about videos and such—They have the capability for them now, but we haven't been given the green light to find them for reading yet. But maybe worth discussing for social studies if you want them. There's a lot of good historical public domain stuff out there!

2- **Passage Readabilities**

Pearson is partnering with Metametrics to analyze all passages in order to ensure appropriate grade-level placement. Every passage will receive a Lexile score from the Lexile Publisher Assistant program. Pearson recommends that the following scale, developed by Metametrics and utilized by many state assessment and curriculum programs, be used as one measure for grade-level decisions.

Typical Lexile Reader Measures by Grade for English Text

Grade	Reader Measures, Mid-Year *25th percentile to 75th percentile (IQR)**
1	BR120L to 295L**
2	170L to 545L
3	415L to 760L
4	635L to 950L
5	770L to 1080L
6	855L to 1165L
7	925L to 1235L
8	985L to 1295L
9	1040L to 1350L
10	1085L to 1400L
11 & 12	1130L to 1440L

* The Lexile range shown is the middle 50 percent of reader measures for each grade. This means that 25 percent of students had Lexile measures below the lower number and 25 percent had Lexile measures above the higher number.
** Beginning Reader (BR) is a code given to readers and texts that are below 0L on the Lexile scale. The lower the number following the BR code, the more advanced the reader or text is. The higher the number, the less complex the text is or less skilled the reader is.

Source: Metametrics

Prior to publishing passages on field test forms, Metametrics will provide a certified Lexile score for each passage.

***Brian**: Website:* https://accounts.lexile.com/login/

You can create a free account and it will let you check passages up to 1000 words and provide a lexile range (800-900, for example). That's close enough for now, and you can see how it fits on the scale above. We're getting our contract in place with Metametrics now, and once that's done you can use it to find a precise score.

3- Text Complexity

While Lexile score is an important quantitative measure of a passage's overall readability, there are two additional aspects that are equally important. Each passage can be analyzed on a more subjective level for the quality of the text and the anticipated match to expected readers. Assigning a value for text complexity (Readily Accessible; Moderately Complex; or Very Complex) and ensuring a range of passages across complexity levels can help ensure that students receive equitable experiences regardless of which test form they receive. The pyramid below provides a visual representation of the CCSS model, and Pearson recommends that we follow a similar model for KY Summatve Assessments.

***Brian**: This idea will probably be important for Social Studies too since some of the passages are pretty complex in their writing and sentence structure and ideas and such, so will receive a pretty high Lexile*

score. You can cJfset that a bit by showing the importance cf a text, or by showing that even though the language is pretty challenging, the message is pretty straightforward—that sort cf thing.

Qualitative evaluation of the text

Levels of meaning, structure, language conventionality and clarity, and knowledge demands

Quantitative evaluation of the text

Readability measures and other scores of text complexity

Matching reader to text and task

Reader variables (such as motivation, knowledge, and experiences) and task variables (such as purpose and the complexity generated by the task assigned and the questions posed)

Source: http://www.corestandards.org/ELA-Literacy/standard-10-range-quality-complexity/measuring-text-complexity-three-factors/

4- Passage Length

The CCSS recommends minimum and maximum word counts across grade bands. When passages are chosen for test forms, the total overall word count will be considered across all passages which helps ensure equity across forms. Passages can be categorized by passage length to help in the selection process. Pearson recommends that passages be classified as short or long according to the following scales:

G3-5:

Short: 200 – 399 words

Long: 400 – 800 words

Pair: up to 1000 words total

G6-8:

Short: 400 – 699 words

Long: 700 – 1000 words

Pair: up to 1250 words total

G10:

Short: 500 – 999 words

Long: 1000 – 1500 words

Pair: up to 1600 words

***Brian**: I'd suggest you consider passage length (short or long) in conjunction with the number of items you want to develop for a cluster. If it's more than maybe 8 or so, you may want long passages. If it's less, short will probably work fine. Our rule of thumb is to figure 5 minutes average for reading time on a passage. Short will be a little less, long will be a little more.*

The same number of items will be developed for passages regardless of passage length. To provide the most flexibility during test construction, Pearson recommends that blueprints not designate form position by passage length. This should only be metadata attached to each passage which will provide a reference during test construction.

5- Passage Type

Passages will be divided into three categories: Literary, Informational, and Paired Passages.

6- Genres

Genres are more specific sub-types of passage type. By specifying genre, a more diverse range of passages and therefore item standards can be assessed. There are many potential genres. Pearson recommends the following:

A- Literary
   a. Fiction story
   b. Poetry
   c. Drama

B- Informational
   a. Article
   b. Expository
   c. Narrative non-fiction
   d. Functional/Technical?

C- Mixed
   a. Fiction/Non-Fiction

On-Demand Writing

Editing Task Commissioned Passages

Length:

G5: ~250 words

G8: ~300 words

G11: ~350 words

Genres:

Fiction: Letter to friend or relative; story from various genres; etc. Not poetry or drama.

Nonfiction: Biographical; literary nonfiction (essay about a mountain or nature); description of a historical event; description of a scientific process or machine; etc.

Grade 5:	6 fiction	6 nonfiction
Grade 8:	5-6 fiction	6-7 nonfiction
Grade 11:	5 fiction	7 nonfiction

General info:

- Items will be based on the ACT model for assessing English (Passage with underlining and items correct or improve the underlined portions unless they are already correct, etc.).
- Each commissioned passage should support a variety of items from the Language standards. Each set will include:

4-5 MC items

1-2 MR items (total of 6 MC and MR items)

2 SA items

Writing on Demand Passages

Number of texts

G5: 2 plus a chart or graph

G8: 3 plus one or two charts or graphs

G11: 4 plus two charts or graphs

Word counts

G5: ~ 600 across texts plus chart or graph

G8: ~800 across texts plus charts or graphs

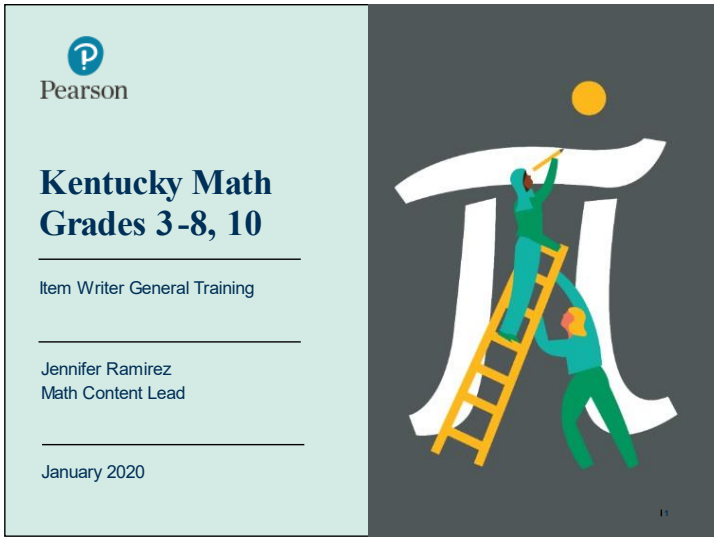
G11: ~1000 across texts plus charts or graphs

Genres

All passages and stimuli from authentic, permissioned or public domain sources. Each set contains related information allowing a student to see information that will lead to an essay taking one side or another on a topic. Science, social studies, current events, etc. Ideally, topics will be interesting and relevant to students, but not controversial.

Recommend as many PD sources as possible, especially for the shorter excerpts and the charts/graphs.

Appendix B. Mathematics Item Writer Training



Item Writer Responsibilities

1. **Confidentiality**
   - Item writers must not copy, discuss, or disclose in any manner the information or materials used during this training, while writing items, or after the assignment has been completed.
2. **Nondisclosure**
   - Item writers must maintain the security of the test items, documents, and materials being created. Item writers will not retain paper or electronic copies of materials after the assignment has been completed.
3. **Ownership**
   - All materials developed for the assessment program must be original and may not appear in any other source. They are the property of Kentucky Assessment and may not be used for any other purpose.

Pearson

Math Item Writer Training 3

ABBI Kentucky Banks

- ❑End-of-Span
  - ❑Bank that houses the G10 Mathematics Items
- ❑K-PREP
  - ❑Bank that houses the G3-8 Mathematics Items
- ❑End-of-Course
  - ❑We will not use this bank for Item development.

Pearson

Math Item Writer Training 4

Asset/Item Types

Multiple Choice (MC)	Multiple Select (MS)	Short Answer (SA)	Extended Response (ER)
• Machine Scored • Max point value: 1 • Only 1 correct answer out of 4 choices	• Machine Scored • Max point value: 2 • Only 2 correct answers out of 5 choices • Partial scoring for one correct answer • Must have only 5 choices	• TE, AI scored, and/or Human Scored • Max point value: 2 • Single or Multiple parts	• TE, AI scored, an/or Human Scored • Max point value: 4 • Single or Multiple parts

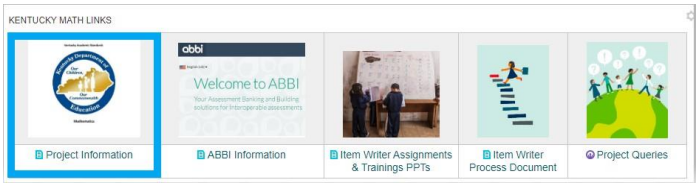
See Item Types trainingfor more information about each item type.

Pearson

Math Item Writer Training 5

Neo - Project Information

The Kentucky Math Project Information link contains all documents that should be used when developing the items.



Neo - Project Information continued

Kentucky Documents:

- Math Standards (KAS)
- ABBI Elements Guide (KY Math)
- Rigor and Cognitive Complexity
- Standards for Mathematical Practice
- Illustrative Mathematics
- Item and Passage Writer Source Requirements
- Universal Design Information

KY Academic Standards (KAS)

- **A combined PDF for all of the grades is provided.**
  - Recommend reading through the Introduction PDF first.
- **For each KAS, Clarifications and Attending to the Standards for Mathematical Practice provided.**
  - Coherence guide is provided in the clarification sections. Use this to ensure that the item is not assessing content that is below or above the grade level.
- **Calculator Designations**
  - No – No calculator will be provided
  - Yes – The grade level appropriate calculator will be provided.
    - G3-5 Desmos Four Function
    - G6-8 Desmos Scientific
    - G10 Desmos Graphing
  - Z – Item dependent. The review committee may decide the calculator usage. If the committee decides it doesn't matter if a calculator is or is not used, then leave as Z and the calculator will be updated once it is placed on a form.
- **Mathematical Practice**
  - There are suggested MP(s) provided for each KAS. These are not set in stone. Based on the item, the item writer will select in the metadata which MP(s) are assessed in the item.

ABBI Elements Guide (KY Math)

- Contains a list of all functionalities that can be used in the Sp2021 Item development.
- Table of Contents are hyperlinked to the section in the document.
- Provides IW with the common uses of the functionality, the scoring of the element, and sample direction lines.

Rigor

- Each standard has been aligned to a specific Rigor. Write the item with the intended rigor in mind.
  - Procedural
  - Conceptual
  - Application

Cognitive Complexity

- Use the chart on page 4 when writing the item. Write to Medium or High.
  - Low
  - Medium
  - High

Pearson

Math Item Writer Training | 9

Standards for Mathematical Practice

- A guide for IWs to use when developing items so that there is a good representation of ALL MPs in the items.
  - Each item must align to one Mathematical Practice Standard.
  - Use this document along with the SMP section in the Kentucky Academic Standards PDF.

Illustrative Mathematics

- Resource for Item Writers to get ideas for item development.
- Caution: These items were developed for classroom use and assess the Common Core State Standards and should only be used when appropriate.

Item Writer Source Requirements

- When using units in real-world contexts, attach a word document of the source in ABBI.
- Title document "UIN_source" and upload under Item Development category.
- This can be uploaded either by the Item Writer or the Research Librarian.

Pearson

Math Item Writer Training | 10

Universal Design Information

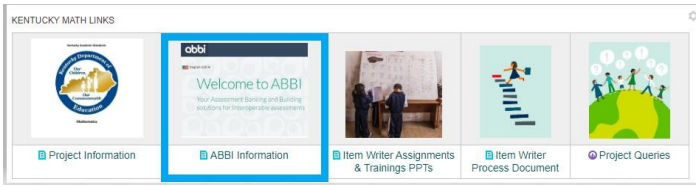
- PowerPoint presentation to aid Item Writers in recognizing potential bias and/or sensitivity issues.
  - Writing Bias Free Items
  - Plain Language Strategies
  - Words and Topics to Avoid

Pearson

Math Item Writer Training | 11

Neo - ABBI Information

The ABBI Information link contains XML source codes for the templates and information guides related to working in ABBI.



Pearson

Math Item Writer Training | 12

Neo - ABBI Information continued

Templates:

- Rubric Templates (KY Math)
- Table Templates (KY Math)

Information Guides:

- ABBI Elements Guide (KY Math)
- Math Equation Creator (KY Math)
- Cluster Set Information Guide
- Entering tables into ABBI
- Inserting a Lilac Feedback Box
- Attaching Sources to Items
- Column Formatting in ABBI
- ABBI Training Documents and Videos
- Metadata (Pearson Use only)

Pearson

Math Item Writer Training | 13

Rubric Templates (KY Math)

- Source code to use when creating rubrics in ABBI.
  - One-point rubric
  - Two-point rubrics
  - Three-point rubric
  - Four-point rubric

Table Templates (KY Math)

- Source code to use when creating tables in ABBI.
  - Horizontal table
  - Vertical table
  - X/Y table
  - Two-way table

ABBI Elements Guide (KY Math)

- Contains a list of all functionalities that can be used in the Sp2021 Item development.
- Table of Contents are hyperlinked to the section in the document.
- Provides IW with the common uses of the functionality, the scoring of the element, and sample direction lines.

Pearson

Math Item Writer Training | 14

Math Equation Creator

- Guidelines for determining when to use Equation Creator in ABBI

Cluster Set Information Guide

- Information about what a Cluster Set includes as well as the process for how to create a cluster set.
- Cluster sets include two to four independent items that assess different standards but share a common stimulus.

Entering Tables into ABBI

- Instructions for how to enter the table element and source code.

Pearson

Math Item Writer Training | 15

Inserting a Lilac Feedback Box

- Use a Lilac Feedback Box (LFB) to communicate questions, comments, concerns internally throughout item development.
- Provides you with instructions on how to insert a LFB into ABBI.

Attaching Sources into ABBI

- Provides you with instructions on how to upload a source document into ABBI.

Column Formatting in ABBI

- Provides you with instructions on how to adjust an item's with or how to change it to a two-column format.

Pearson

Math Item Writer Training | 16

ABBI Training documents and Videos

The ABBI Training Documents and Videos contains links to trainings that are already created in Neo about different ABBI topics:

Topic	Link
Getting Started	Getting Started with ABBI
Asset List	ABBI Guide for Asset List Filtering and Sorting - Video Using the Asset List Customize Search Results
Create, Edit, Review	Creating and Editing Items and Passages ABBI Guide for Creating Items - Video ABBI Guide for Using Edit Mode - Video ABBI Guide for Using Review Mode - Video
Metadata	ABBI Guide for Metadata - Video
How-to guides for ABBI Item Interactions	HOW TO: Item Interactions

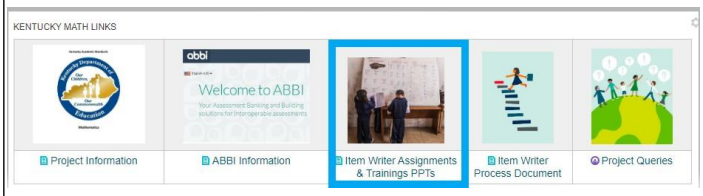
HOW TO: Item Interactions contains information, examples, and how -to guides for each element in ABBI .

Pearson

Math Item Writer Training | 17

Neo – IW Assignments & Training PPTs

- Contains the assignment(s) for each Item Writer. Pearson will notify the IW when an assignment has been posted.
- Links to all Training PowerPoints



Pearson

Math Item Writer Training | 18

Item Writing 101

- For first time Kentucky Item Writers
  - Which documents to have open
  - How to use the documents to develop the item
  - How to enter your item in ABBI

Math Item Types Training

- Description of the various item types used
  - MC/TE/FIB – Multiple choice/Technology Enhanced/Fill -in-the-Blank
  - MS – Multiple Select
  - SA – Short Answer
  - ER – Extended Response

Pearson

Math Item Writer Training | 19

Rubric and Rationale Training

- All Item Types (except MC) require a rubric.
  - The source codes for the rubrics are provided on NEO.
  - Rubrics should mirror what was asked of the student in the item.
- Multiple Choice, Multiple Select, and Inline Choice requires rationales for the correct and incorrect choices.

Item Writer General Training

- The link to this presentation can be found on the NEO Project Information page.

Asset Writer Checklist

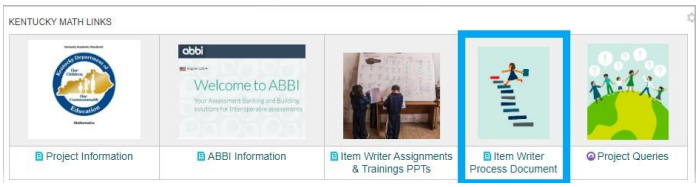
- Helpful checklist available for the IW's use.

Pearson

Math Item Writer Training | 20

Neo – Process Document

- Steps that the Item Writer and Pearson AS will go through when writing, submitting, and reviewing items.
- Contains links, email information (subject lines, who to include), and item status in ABBI to use.

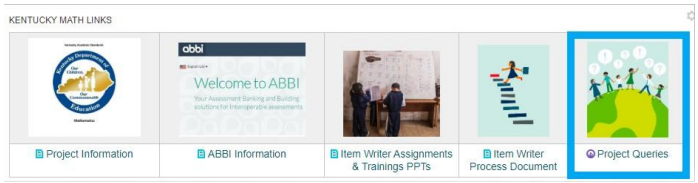


Pearson

Math Item Writer Training 21

Neo – Project Queries

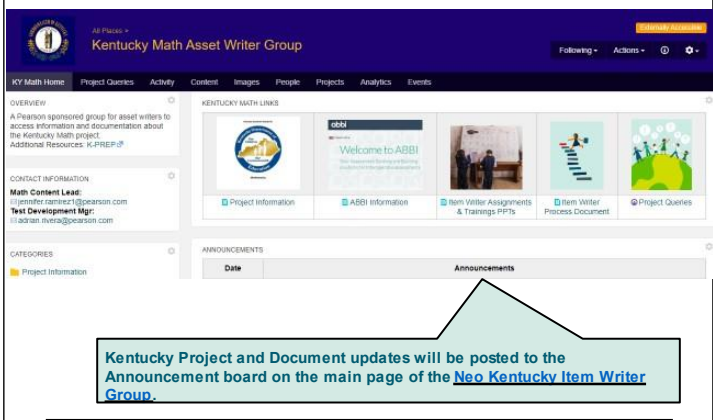
- A place to ask general process questions.
- Do not ask item specific questions here. Use a LFB within ABBI.
- Useful to place to look to see if others had the same question or concern.



Pearson

Math Item Writer Training 22

Neo - Announcements



Kentucky Project and Document updates will be posted to the Announcement board on the main page of the Neo Kentucky Item Writer Group.

Pearson

Math Item Writer Training 23

Contact Information

- All item specific questions should be posted on the Item in ABBI using a LFB.
- Send an email to jennifer.ramirez1@pearson.com when there is a LFB question that requires Pearson feedback.

Math Item Writer Training 24

Conclusion

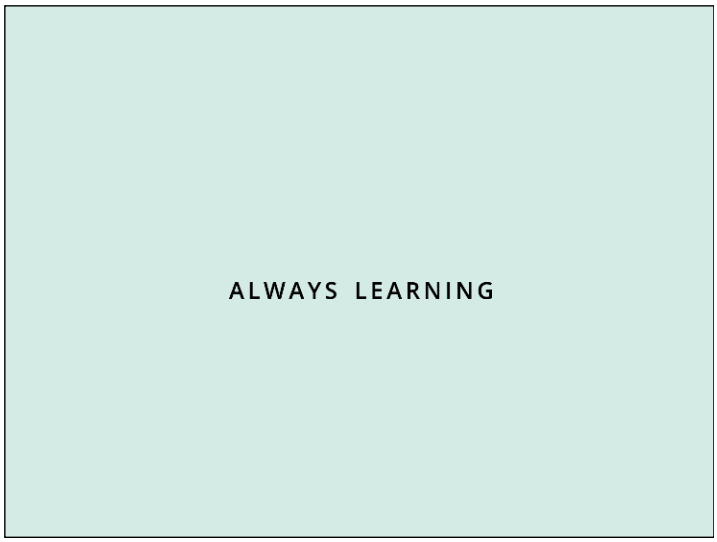
- Review the item after it is written to make sure it aligns to the Kentucky Academic Standard (KAS) and intent of the item type.
- Keep in mind the content limitations of the grade level and previous grade level.
- Refer to the "Asset Writer Checklist" before submitting an item.
- Check ABBI Elements Guide to get the correct direction line verbiage.
- Use the power point trainings as needed.

Questions?

Math Item Writer Training 25

There's so much more to learn

Find out more at
Kentucky Math Item Writer Group on Neo


ALWAYS LEARNING

Appendix C. Social Studies Item Writing Training

Pearson

Kentucky Summative Assessments—Social Studies

Item Writer Training

October 2020

Photograph in the Carol M. Highsmith Archive, Library of Congress, Prints and Photographs Division

Pearson

2

Agenda

I Welcome

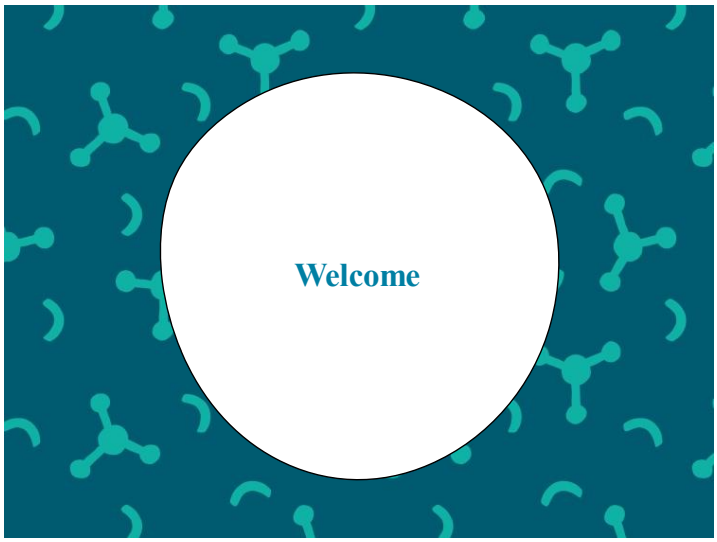
II Kentucky Academic Standards

III Project Overview

IV Lessons Learned: Dos and Don't

V Resources and Wrap -Up

VI Questions



Participants

1. **Pearson**
   - Ariel Juarez, Content Specialist —Social Studies
   - Lydia Mantis, Content Specialist —Social Studies
   - Michael Bardgett, Content Specialist —Social Studies
   - Sharon Staples, Principal Assessment Specialist —Social Studies
2. **Item Writers**
   Many thanks for participating in this training session!

Kentucky Social Studies Item Writing Team

Pearson

14

KAS: Disciplinary Standards

1. **Strands**
   - Civics (blue)
   - Economics (yellow)
   - Geography (green)
   - History (purple)
2. **Concepts and Practices**
   - Important note: Kentucky is embedded across all strands and most grades.

Civics (C)	Economics (E)	Geography (G)	History (H)
Civic and Political Institutions (CP)	Microeconomics (MI)	Migration and Movement (MM)	Change and Continuity (CH)
Roles and Responsibilities of a Citizen (RR)	Macroeconomics (MA)	Human Interactions and Interconnections (HI)	Cause and Effect (CE)
Civic Virtues and Democratic Principles (CV)	Specialization, Trade and Interdependence (ST)	Human Environment Interaction (HE)	Conflict and Compromise (CO)
Processes, Rules and Laws (PR)	Incentives, Choices and Decision-making (IC)	Geographic Reasoning (GR)	Kentucky History (KH)
Kentucky Government (KGO)	Kentucky Economics (KE)	Kentucky Geography (KGE)	

Source: Kentucky Department of Education

Pearson

16

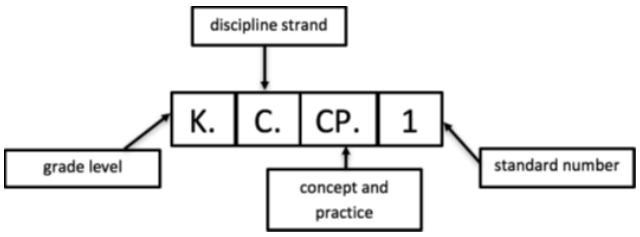
KAS: Organization

1. **By grade level within grade bands (K–5, 6–8, and high school)**
   - Overview
   - Standards
   - Clarifications
2. **By inquiry practices and content progressions**
3. **Tips:**
   - Use the table of contents to access information in multiple ways.
   - Use the keyboard shortcut Ctrl F to easily search by Disciplinary Standard or Inquiry Practice

Pearson

17

KAS: Coding Example



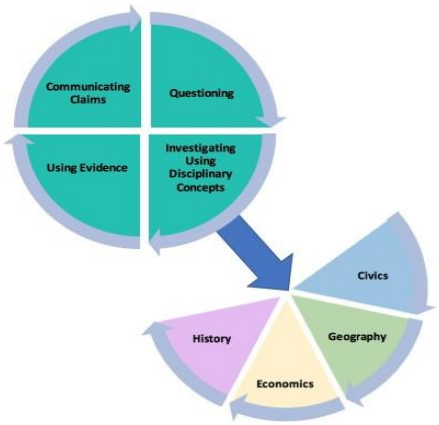
Source: Kentucky Department of Education

This example is for a Kindergarten Civics standard from the concept Civic and Political Institutions.

Pearson

18

KAS Inquiry Overview



The KAS "place an equal importance on both the mastery of important social studies concepts and disciplinary practices . . . . As indicated by the graphic on this slide, concept knowledge cannot be achieved effectively without the practice of inquiry. Neither development of the practices nor development of the knowledge and understanding within the lenses is sufficient on its own."

Pearson

Source: Kentucky Department of Education | 9

KAS Inquiry Practice: Questioning

1. **Compelling Questions**
   - Open-ended with many defensible responses
   - Centered on enduring, significant, unresolved issues
   - Rigorous
   - Intellectually challenging and interesting
   - Inspire investigation within the disciplinary strands
2. **Supporting Questions**
   - Discipline specific
   - Build knowledge for answering a compelling question
   - Lead students to information that is generally accepted within the discipline
3. **Paired Examples**
   - Compelling: Is interaction between different people and cultures beneficial?
   - Supporting: How did trade affect Song China?
   - Compelling: How are people and places affected by rapid migration?
   - Supporting: How have shifting settlement patterns changed the Midland -Odessa region?

Pearson

| 10

Project Overview

Project Overview

1 Grades and Grade Bands	2 Disciplines and Practices	3 Item Types	4 Test Components
Fifth (K-5)	Civics	Multiple choice (MC)	Standalone items
Eighth (6-8)	Economics	Multiple select (MS)	Cluster sets
Eleventh (HS)	Geography	Technology enhanced (TE)	
	History	Short answer (SA)	
	Inquiry	Extended response (ER)	

Pearson

12

Choice

1. **MC**
   - Used as standalone and cluster items
   - Written in the form of a question
   - Four answer options
   - Only one correct answer
2. **MS**
   - Used as standalone and cluster items
   - Written in the form of a question
   - Five answer options
   - Two correct answers
3. **Style**
   - Interrogative first
     - "Which" + noun (rather than "Which of the following")
     - "How"
     - "Why"

Pearson

13

TE Items

1. **Description**
   - Used as standalone and cluster items
   - Interactive
   - May have one or more correct answers
   - Written as statements rather than questions
   - Include assessment of content and directions for solving in the prompt/stem
   - Worth 1 or 2 pt.
2. **Create TEs that assess students in ways that MC items do not.**
3. **Interaction types used for KY SS**
   - Match
   - Match–Table Grid
   - Hot Spot
   - Hot Text
4. **Examples**
   - See the next four slides
   - See additional examples in ABBI>K -PREP>Sandbox and ABBI>End -of-Span>Sandbox

Pearson

14

TE Interaction: Matching –Match

An economics student is investigating the compelling question "Is competition good?" As part of the investigation, the student asks different people to answer the supporting question "How much competition do you want in the market?" Move each response to the correct box to show whether it supports perfect competition or a monopoly.

"I want the price for a good to be stable instead of having to compare prices."

"I like to have many options to choose from when I want a service."

"It makes me angry when I want a good and there is a shortage of the good."

Perfect Competition	Monopoly

- Grade 11
- HS.E.MI.1
- HS.E.I.UE.2
- HS.E.I.UE.2
- DOK 2

Pearson

15

TE Interaction: Matching –Match Table Grid

1. Limit to only one selection for each row or column.
2. Grade 5 DOK 1 Example (aligned to 3.E.MA.1)

Decide whether each of the properties in the table describes a private or a public property. Choose **one** answer for each description.

Description	Private Property	Public Property
The property is paid for with tax money.	○	○
The property can be bought and sold by a person.	○	○
The property is owned by the people in a community.	○	○

Pearson

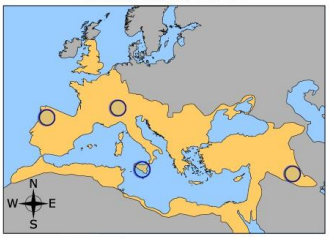
16

TE Interaction: Hot Spot

1. Grade 8
2. 6.G.HE.1
3. DOK 2

The physical environment helped the Roman Empire develop a vast trade network. Select **one** location with a physical environment that made it an ideal trade hub within the Roman Empire.

Roman Empire, 117 CE

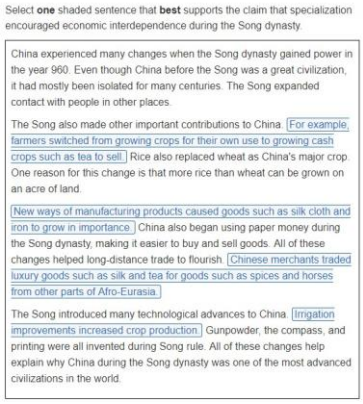


Pearson

17

TE Interaction: Hot Text

1. **Students should be asked to select entire sentences or paragraphs rather than phrases for hot text.**
2. **KY SS Grade 8 sample**
   - 7.E.ST.2 Analyze the impact of specialization upon trade and the cost of goods and services. (Specialization, Trade and Interdependence)
   - 7.I.UE.2 Analyze evidence from multiple perspectives and sources to support claims and refute opposing claims, noting evidentiary limitations to answer compelling and supporting questions. (Using Evidence)
3. **DOK 2**


Select **one** shaded sentence that **best** supports the claim that specialization encouraged economic interdependence during the Song dynasty.

China experienced many changes when the Song dynasty gained power in the year 960. Even though China before the Song was a great civilization, it had mostly been isolated for many centuries. The Song expanded contact with people in other places.

The Song also made other important contributions to China. For example, farmers switched from growing crops for their own use to growing cash crops such as tea to sell. Rice also replaced wheat as China's major crop. One reason for this change is that more rice than wheat can be grown on an acre of land.

New ways of manufacturing products caused goods such as silk cloth and iron to grow in importance. China also began using paper money during the Song dynasty, making it easier to buy and sell goods. All of these changes helped long-distance trade to flourish. Chinese merchants traded luxury goods such as silk and tea for goods such as spices and horses from other parts of Afro-Eurasia.

The Song introduced many technological advances to China. Irrigation improvements increased crop production. Gunpowder, the compass, and printing were all invented during Song rule. All of these changes help explain why China during the Song dynasty was one of the most advanced civilizations in the world.

Pearson

18

Choosing the Right TE Interaction

Item Function	TE Interaction Options
Identifying the location of a place/event/concept on a map or graphic	Hot Spot
Classifying or sorting multiple pieces of information	Match Match Table Grid
Putting events in chronological order	Match
Organizing processes	Match
Providing evidence	Hot text

Pearson

19

SA Items: Description

1. **2-point items used in cluster sets only**
2. **Must align to assigned discipline for the set**
3. **Style**
   - Generic directions
     - Read the question carefully. Then enter your answer in the space provided.
   - A statement that quotes directly from the aligned disciplinary standard.
     - Using your knowledge of [insert language of the standard], [insert action].
   - Specific directions
     - In your response, use evidence from multiple sources to [insert language related to the action]. Explain your answer in **at least** two sentences.
   - NOTE: Review SA practice items across grades for specific examples.
4. **Must have multiple correct responses**
5. **Must require students to synthesize information from more than one source, with KDE preferring use of all the sources**
6. **Requires completion of an exemplar and answer cues**

Pearson

20

SA Example (Grade 5)

1. **KAS Disciplinary Standard**
   – 5.C.CP.3 Describe how the U.S. Constitution upholds popular sovereignty, ensures rule of law and establishes a federal system.
2. **KAS Inquiry Practice**
   – 5.I.CC.2 Construct arguments using claims and evidence from multiple sources on how a founding principle(s) is applicable today.
3. **Directions and Prompt**
   Read the question carefully. Then enter your answer in the space provided.

   Using your knowledge of the U.S. Constitution, evaluate the following claim.

   Claim: The U.S. Constitution upholds the idea of popular sovereignty.

   Use evidence from **at least** two sources to support the claim. Explain your answer in **at least** two sentences.

Pearson 21

SA Example (Grade 8)

1. **KAS Disciplinary Standard**
   – 7.G.HE.1 Examine how physical geography influenced the societies and empires of Afro - Eurasia and the Americas between 600 -1600.
2. **KAS Inquiry Practice**
   – 7.I.UE.1 Use multiple sources to develop claims in response to compelling and supporting questions.
3. **Directions and Prompt**
   Read the question carefully. Then enter your answer in the space provided.

   Using your knowledge of how physical geography influenced empires in Afro - Eurasia, evaluate the following claim.

   Claim: China's physical geography limited interactions between the Song Dynasty and other empires.

   In your response, use evidence from multiple sources to support or refute the claim. Explain your answer in **at least** two sentences.

Pearson 22

SA Example (Grade 11)

1. **KAS Disciplinary Standard**
   – HS.G.HI.2 Analyze how cultural and economic decisions influence the characteristics of various places.
2. **KAS Inquiry Practice**
   – HS.G.I.UE.2 Gather information and evidence from credible sources representing a variety of perspectives relevant to compelling and/or supporting questions in geography.
3. **Directions and Prompt**
   Read the question carefully. Then enter your answer in the space provided.

   Using your knowledge of how economic decisions influence the characteristics of various places, answer the following supporting question.

   Supporting question: How has economic growth been both good and bad for Texas?

   In your response, use evidence from the sources to answer the supporting question. Explain your answer in **at least** two sentences.

Pearson 23

ER Items: Description

1. **4-point items that are used in cluster sets only**
2. **Must align to the assigned discipline**
3. **Style**
   - Generic directions
     - Read the question carefully. Then enter your answer in the space provided.
   - Alignment: KDE expects to see language that shows clear alignment to the set, the disciplinary standard, and/or the inquiry practice.
     - Example of alignment to inquiry: A prompt that begins with "Construct an explanation" or "Construct an argument."
     - Example of alignment to the set: Asking students to respond to the compelling question or to a supporting question. Note that the prompt must identify the question as compelling or supporting.
     - Example of alignment to the disciplinary standard: Incorporating selected terms into the prompt or repeating phrases from the standard
   - Specific directions that reference the expectations of the inquiry practice.
4. **Must have multiple correct responses**
5. **Must require students to synthesize information from more than one source, with KDE preferring use of all the sources**
6. **Requires completion of an exemplar and answer cues**

Pearson

24

ER Example (Grade 5)

1. **KAS Disciplinary Standard**
   - 5.C.PR.1 Evaluate whether various rules and laws promote the general welfare, using historical and contemporary examples.
2. **KAS Inquiry Practice**
   - 5.I.CC.2 Construct arguments using claims and evidence from multiple sources on how a founding principle(s) is applicable today.
3. **Directions**
   - Read the question carefully. Then enter your answer in the space provided.
4. **Prompt**
   - Federalists and Anti-Federalists disagreed that the Constitution created a government that was good for the people. Construct an argument that answers the supporting question "Does the Constitution establish a government that promotes the general welfare?" Support your claim with evidence from multiple sources. Write **at least** two paragraphs.
5. **Note for Grade 5 only**
   - KDE is open to using bullet points after the main prompt to provide scaffolded direction to students.

Pearson

25

ER Example (Grade 8)

1. **KAS Disciplinary Standard**
   - 7.G.HI.2 Examine ways in which one culture can both positively and negatively influence another through cultural diffusion, trade relationships, expansion and exploration.
2. **KAS Inquiry Practice**
   - 7.I.UE.2 Analyze evidence from multiple perspectives and sources to support claims and refute opposing claims, noting evidentiary limitations to answer compelling and supporting questions.
3. **Directions**
   - Read the question carefully. Then enter your answer in the space provided.
4. **Prompt**
   - Construct an argument to answer the compelling question "Is interaction between different people and cultures beneficial?" Use what you have learned about cultural diffusion, trade relationships, and expansion during the Song Dynasty. Use multiple sources to develop a claim in your response and note at least one evidentiary limitation to your response. Write **at least** two paragraphs.

Pearson

26

ER Example (Grade 11)

1. **KAS Disciplinary Standard**
   - HS.G.HE.1 Assess the reciprocal relationship between physical environment and culture within local, national and global scales.
2. **KAS Inquiry Practice**
   - HS.G.I.CC.2 Engage in disciplinary thinking and construct arguments, explanations or public communications relevant to compelling and/or supporting questions in geography.
3. **Directions**
   - Read the question carefully. Then enter your answer in the space provided.
4. **Prompt**
   - Construct an argument to answer the compelling question "How are people and places in Texas affected by rapid migration?" Use what you know about the reciprocal relationship between the physical environment and culture to answer. Use multiple sources to develop a claim in your response. Write **at least** two paragraphs.

Pearson

27

Depth of Knowledge (DOK)

DOK measures the cognitive complexity, not the difficulty, of a task.

1 Recall & Reproduction	2 Working with Skills & Concepts	3 Strategic Thinking	4 Extended Thinking
• Requires only a onestep cognitive process, such as recalling facts or locating information • Limited to a basic demonstration of social studies skills rather than a recall of social studies facts for KY SS	• Requires cognitive processing beyond identification or recall • Some reasoning required • Examples: comparing and contrasting; identifying cause and effect • Tip: In MCs, distractors usually inaccurate	• Requires complex or abstract cognitive processing • Requires reasoning • Examples: connecting ideas and citing supporting evidence; applying a concept to a different context; synthesizing information from multiple stimuli (especially unfamiliar stimuli) • Tip: In MCs, distractors often all true	• Not applicable for KY SS

Pearson

28

Lessons Learned: Dos and Don'ts

DOs and DON'Ts: Bias and Sensitivity

DO	DON'T
• Consider the needs of all populations. • Make maps and other graphics as simple as possible.	• Use idioms and multiple-meaning words that may be unfamiliar to some students. • Use stereotypes, offensive language, and highly controversial subjects. • Ignore the realities of social studies content.

Pearson 30

DOs and DON'Ts: Item Construction

DO	DON'T
• Indicate the key for MC and MS items. • Verify scoring or describe how to solve TE items. • Write plausible incorrect options. • Use plain language that is as clear and concise as possible. • Use parallel language (syntax, content, and style) and length for MC and MS items. • Align items to only one KAS Disciplinary Standard.	• Clue the test taker. • Include unnecessary information. • Use negative stems or prompts ("Which is NOT a reason . . . ?"). • Use absolutes such as *always* or *never* in only one answer option. • Require an unreasonable number of answers to earn credit for TE items. • Ask students to complete diagrams, lists, propose titles, etc.

Pearson 31

DOs & DON'Ts: All Items

DO	DON'T
• Verify scoring. • Use the language of the standards whenever possible. • Dual-align items to inquiry. • Limit items to the time period specified in the aligned disciplinary standard. • Have tight alignment to the standard and practice.	• Assume extensive content knowledge not stated in the assigned KAS Disciplinary Standard. • Develop items that only test recall of content knowledge. • Expect students to make inferences.

Pearson 32

DOs & DON'Ts: Cluster Sets

DO	DON'T
• Use stimuli that allow exploration of the same topic through the "lens" of different disciplines • Utilize different disciplinary standards from within the entire grade band. • Write items that require students to use multiple stimuli to answer. • Let the compelling question guide the development of the set. • Focus on depth rather than breadth.	• Be creative in the type of product expected for SA and ER items. • Create items that are interdependent or that clue each other. • Repeat the same information across different stimuli in a cluster. • Include more than one supporting question per set.

Pearson 133

DOs & DON'Ts: Stimulus Selection

DO	DON'T
• Use variety, including excerpts, bulleted lists, diagrams, political cartoons, photographs, maps, headlines, timelines, and graphs. • Include a stimulus for at least 50% of all items. • Select primary sources or high-interest modern texts that show the agency and perspective of underrepresented groups. • Use public domain images. • Take advantage of using *de minimis* texts.	• Use summarized, encyclopedic "tertiary" sources • Author sources that would be considered encyclopedic • Use lengthy text stimuli that will require excessive scrolling by students. • Use wiki-based sites. • Use .edu sites that are student produced. • Assume that all images from .gov sites are in the public domain

Pearson 134

DOs & DON'Ts: *De Minimis*

DO	DON'T
• Enjoy the flexibility of using texts from non-public domain sources. • Use direct quotations as a standalone stimulus or within the context of an appropriate summary. • Consider putting two or three related short excerpts from different sources together as one source in a cluster set • Summarize *de minimis* text that is above-grade level, and identify as "based on"	• Quote more than 3–4 sentences. • Adapt *de minimis* text.

Pearson 135

DOs and DON'Ts: Topics

DO	DON'T

• Introduce diversity.
• Show the agency of underrepresented groups.
• Include references to Kentucky.
• Align items to the grade -level themes.
  • Gr 3: Global
  • Gr 4–8: Within the context of the time period
• Look for targeted assignments for some standalone items.
	• Overlook cultures in Africa, Asia, Latin America, and Oceania

Pearson

I 36

Grade-Level Themes

Grade	Topic	Date Range
K	Self, school, local community	Past to present
1	Local and state	Past to present
2	North America (Canada, United States, Mexico)	Past to present
3	Africa, the Americas[1], Asia, Europe, Oceania	Present
4	United States	1492 to ~1763
5	United States[2]	~1763 to ~1791
6	River valley and classical civilizations	3500 BCE–600 CE
7	Afro-Eurasia and the Americas	600–1600
8	United States	1600–1877
Civics	Kentucky, United States, and the world	Past to present
Economics	Kentucky, United States, and the world	Past to present
Geography	Kentucky, United States, and the world	Past to present
U.S. History	United States	1877 to the present
World History	World[3]	1300 to the present

K: "local" = KY

Gr 3: United States limited to comparisons to other countries

Pearson

I 37

Resources & Wrap- Up

Item Writer Resources

1. **Sample Items**
   - ABBI
     - Grades 5 and 8: Kentucky>K-PREP>Sandbox
     - Grade 11: Kentucky>End-of-Span>Sandbox
   - Status: Author>Create
2. **Neo: https://neo.pearson.com/groups/kentucky_-social-studies-item-writer-resources**
   - Training PowerPoints
     - Kentucky Summative Assessments—Social Studies
     - ABBI
   - Kentucky Academic Standards and Other Resources
     - KAS
     - Glossary
     - High School Clarifications
   - Other Resources
     - Applying Webb's Depth of Knowledge (DOK) in Social Studies
     - Asset Writer Checklist
     - Public Domain Source List
     - Quick Reference Guide
   - ABBI Job Aids

Pearson 39

Item Writer Responsibilities

1. **Confidentiality**
   - Item writers may not copy, discuss, or disclose the information or materials used during this training or as part of the writing assignment.
   - Email communication should NOT include secure information.
   - Item writers will securely destroy all paper or electronic copies of materials after completion of the assignment.
2. **Ownership**
   - All items developed for Kentucky must not be used elsewhere. Items become the sole property of Pearson/KDE.
3. **Punctuality**
   - Item writers must submit assignments according to the schedule and specifications detailed on the Statement of Work and item-writing assignment.
4. **Originality and Quality**
   - Item writers are expected to submit original, high-quality items that meet the program specifications and Pearson expectations.
5. **Restrictions**
   - Item writers may not accept outside offers to produce materials designed for practicing or familiarizing students with the content of Kentucky Social Studies.

Pearson 40

Logistics

1. **Email communication should NOT include secure information that relates to the content of items.**
   - Discussion of secure information should be by phone.
2. **SOWs**
   - Training time is exact.
   - Item counts indicate the maximum number of items writers will be asked to submit.
   - Start and end dates are not the same as assignment due dates.
3. **Optional training opportunities will be provided as needed.**

Pearson 41

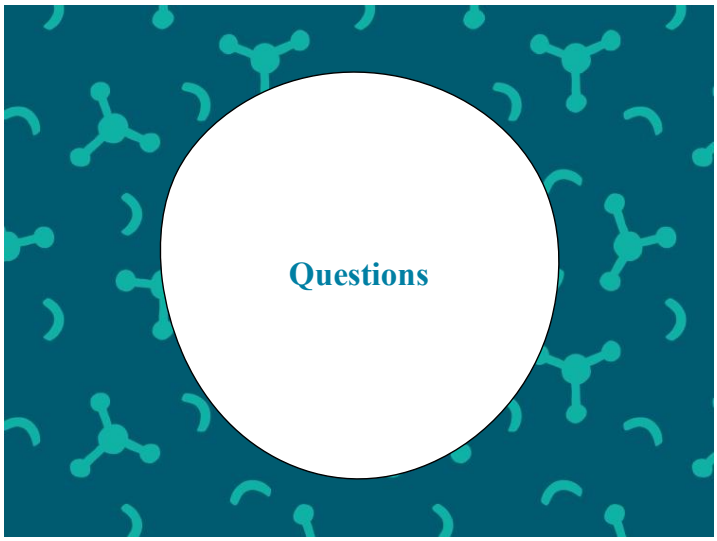
Pearson Contact Information

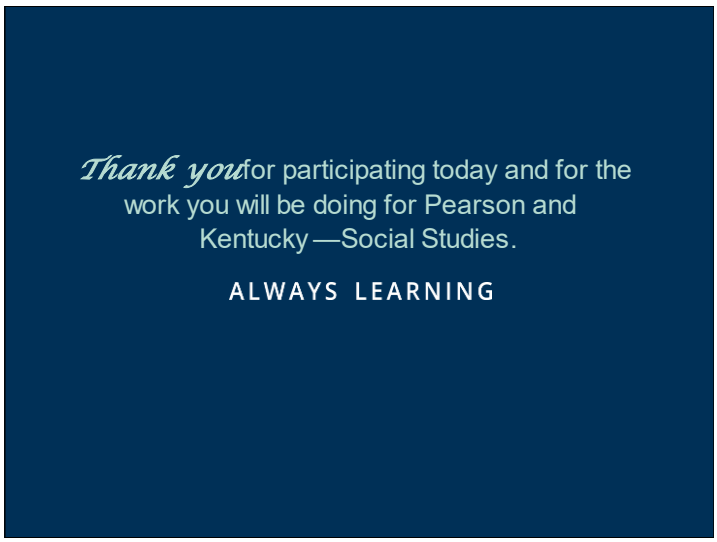
1. **Sharon Staples**
   - sharon.staples@pearson.com
   - 319-229-5212
2. **Ariel Juarez**
   - ariel.juarez@pearson.com
   - 210-526-1579
3. **Lydia Mantis**
   - lydia.mantis@pearson.com
4. **Michael Bardgett**
   - michael.bardgett@pearson.com

Please include all four team members on emails. Avoid mentioning secure information in emails.

Pearson

42





Appendix D. Item Development Review Criteria Checklist

Item Review Criteria Checklist
Does the item…
☐ Align to the standards and item test specifications
☐ Have one and only one clearly correct answer
☐ Have a stem that gives the student a full sense of what the item is asking
☐ Use incorrect response options that are plausible, reasonable misconceptions and errors
☐ Use response options that relate to the stem in the same way
☐ Avoid having one response option that is markedly different from the others
☐ Avoid clues to students, such as absolutes or words repeated in both the stem and options
☐ Measure the specified portion of the curriculum and/or test specifications
☐ Conform to KY item style specifications
☐ Test worthwhile concepts or information
☐ Reflect good and current teaching practices
☐ Avoid wordiness
☐ Reflect content in a manner that is free from bias against any person or group
☐ Allow for equal access among all populations of interest
Does the rubric (if any) for the item…
☐ Contain a clear definition of each score level
☐ Lend itself to clear differentiation between score levels
Is the stimulus/art (if any) for the item including passages…
☐ Required to answer the item
☐ Likely to be interesting to students
☐ Clearly and correctly labeled
☐ Providing sufficient additional information to answer the item
☐ Appropriate for the grade level and student population
☐ At the appropriate reading level
☐ Presenting grade-appropriate graphics and information load

Appendix E. Item and Passage Writer Source Requirements

Item and Passage Writer Source Requirements

- **Goal:** Encouraging item writers to use quality source material to create higher-quality items and passages resulting in less/quicker review time during development.

How do we decide what should be accepted or rejected?

- In addition to the criteria already used by Content to decide if a submission is acceptable, the use and citing of sources should be considered. Rejection of an item or passage may be determined based on the questions:
    1. Are sources listed for all facts and data that are used in the item?
    2. Are the sources authoritative and appropriate to the topic?
    3. Are citations and working links provided for sources from the open web?
    4. Are PDFs or scans provided for any print sources used, and for sources that come from proprietary databases that might not be universally accessible?

If the answer to any of the above is no, then the item should be rejected and/or

the writer should be asked to revise and resubmit.

What is a "fact"?

- Item writers should use and provide sources for all facts that they include in their items, and a broad definition should be applied to "fact." Types of facts and information that should be sourced by the writer include, but are not necessarily limited to, the following:

    1. A statement of fact in the item stem or in scoring rubrics. *(e.g., the average male elephant weighs 9,900 pounds)*
    2. Any data presented in a chart, graph, etc.
    3. Information presented in any art, photographs, diagrams, maps etc.
    4. Biographical data such as birth and death dates, names, etc.
    5. Quotes from notable publications or individuals.
    6. Non-English terms, medical terms, chemical names, etc. used in the item stem. *(e.g., the wood of* Acacia nilotica *was used by ancient Egyptians to make statues and furniture)*
    7. Qualitative evaluations like "most" or "best" should be backed up by a source showing that the assertion is reasonable.
    8. Real-life situations and scenarios. As above, this type of information should be backed up by a source showing the scenarios are reasonable. *(e.g., an average adult can swim 200 meters in 4 minutes)*

- It is possible that an item may not need source information. Types of information that *may not* require sourcing by the item writer may include:

    1. Non-factual real world scenarios (e.g., *Maria and Susan took a walk around a lake. They saw 10 different types of trees.)*

2. Generic (e.g., *a table at a pizza restaurant could seat 8 people* vs. *a table at a pizza restaurant measured 25 inches high and had a diameter of 3 feet*)
3. Fictional
4. Custom dimensions, prices, etc. (e.g., *A local beekeeper sold her jars of honey for $5.00 each* vs. *the average price for a 15 ounce jar of honey is $4.25)*

If a writer has included facts in an item but has not provided a source, the item should be rejected.

What is an authoritative source for item writing?

- Item writers should use authoritative sources for any facts they include in their items. Content Specialists should evaluate the authority of the sources cited by the writer as part of the accept/reject decision process. An authoritative source is:

    1. Authored by an expert in the field.
    2. Reputable (e.g., *Encyclopedia Britannica* or the CIA World Factbook).
    3. Has sources listed or cited to back up claims.
    4. Current.
    5. Objective.
    6. Not user-authored.
    7. Not a personal Website, blog, "hobby" site, or a student project site.
    8. Well-written, and free of grammatical and typographical errors.

- Web-based sources should be produced by authoritative organizations or by qualified individuals through reputable institutions. Suitable web-based sources may include:

    1. Government sites (.gov).
    2. Educational institution sites (.edu but NOT student pages or projects).
    3. Specialty sites (e.g., the American Heart Association or the Arbor Day Foundation).
    4. Texts or articles accessed via Google books, Google scholar or similar sites.
    5. Databases and encyclopedias accessed via public or academic libraries.

- Sources used to write items or passages should be appropriate to the topic. For example, a tour company website would be an appropriate place to find information for the price of a bus tour around Paris. The tour company website would *not* be an appropriate source for information about the length of the Seine River, or the height of the Eiffel Tower - the writer should use an authoritative source such as an encyclopedia, gazetteer, or the official Eiffel Tower website.

If a writer has based the facts in an item on sources that are not authoritative and are not appropriate, the item should be rejected.

Keep in mind…

- Whenever possible item writers should use a primary source for facts and data.
    1. A primary source is the original source of the information or data.
    2. A secondary source is any place where primary source data has been republished.

    Example: United States population information should be pulled from the Census Bureau (the primary source), not from a book where the author writes about the population information (a secondary source), even if the Census Bureau is cited in the book.

- Whether or not a writer needs a second source to bolster or confirm a fact he or she is including in an item is very dependent on the nature of the information. Some facts and data may have only one truly authoritative, recognized source.

    Example: The U.S. Energy Information Administration is the one source for U.S. fuel consumption and production statistics.

Red Flags!

- Writers should provide sources for all facts and descriptive information in the items. If an item includes 4 or 5 facts on different topics, and yet only one source is cited on the item template, that is a good clue that the writer has not provided all the sources and the item might need to be rejected.

- The use of the following sites as sources for facts should cause an *automatic reject* of the item.

    Wiki sites:

    Wikipedia (www.wikipedia.org) and other "wiki" type of websites (will usually have the word *wiki* in the URL (e.g. chemwiki.ucdavis.edu)) – These websites should be avoided as sources because *any one* can add, edit, and delete information on the wiki; contributors are not required to provide any proof that they are authorities in the subject on which they are commenting.

    Q&A sites:

    Answers.com (www.answers.com), Ask.com (www.ask.com), All Experts (www.allexperts.com), Yahoo! Answers (www.answers.yahoo.com), etc. – These sites contain answers, provided by complete strangers, usually using online aliases and with no credentials provided, to questions people have posted.

    Expert sites:

    About.com (www.about.com) and eHow.com (www.ehow.com) – The authors who maintain the individual topic pages are rarely experts in the subject and are posting what they have read from other sources, usually without citing them.

    Essay sites:

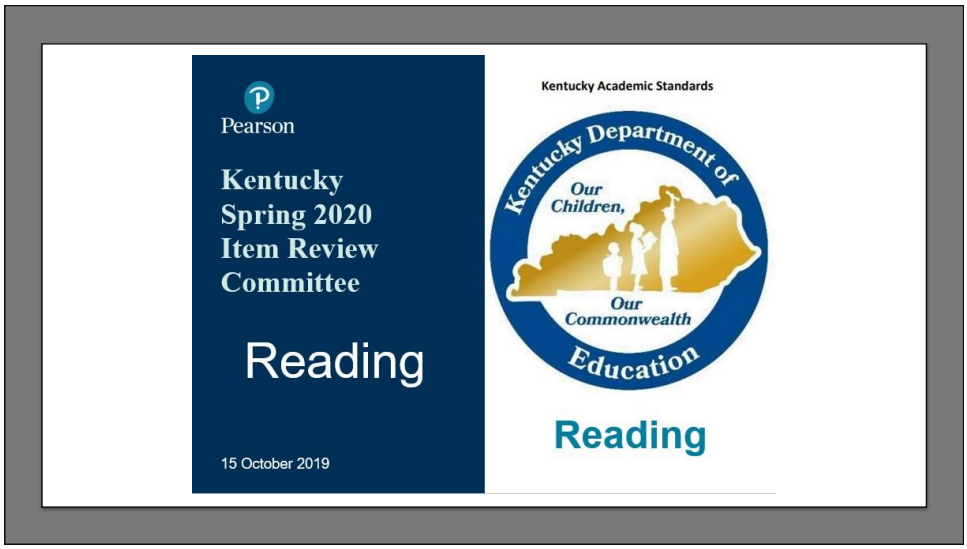
Thinkquest (http://www.thinkquest.org/pls/html/think.library) and Livestrong (www.livestrong.com) – Similar to the Expert sites above, these sites contain essays written by authors without credentials and, in the case of Thinkquest, school children.

Blogs:

Personal blogs from websites such as www.blogspot.com, www.wordpress.com, www.blogs.com, www.blogger.com, www.livejournal.com, etc. – Blogger credentials are rarely available and blogs often are not objective.

For a training on how to attach a source document in ABBI, please see the training posted on Neo.

Appendix F. Reading Item Content Review Training



Agenda

- "Housekeeping"
- Welcome and Introduction

I. Assessment Overview

Components of the Reading Assessment
Evidence -Centered Test Design
Standards
Item Types

II. Item Review Committee Meetings

Reviewer Role
Review Process, Materials
Item Review Guiding Questions and Criteria

III. ABBI Training

"Housekeeping"

Non-Disclosure/Security

- Process vs. Specifics
- Materials
- Cell Phones

Schedule

Grade	Tuesday	Wednesday
3 - 5	8:30 am - 5:00 pm	8:30 am - 5:00 pm
6 - 8	8:30 am - 5:00 pm	8:30 am - 5:00 pm
10	8:30 am - 5:00 pm	8:30 am - 5:00 pm

Breaks and lunch will be determined in each room

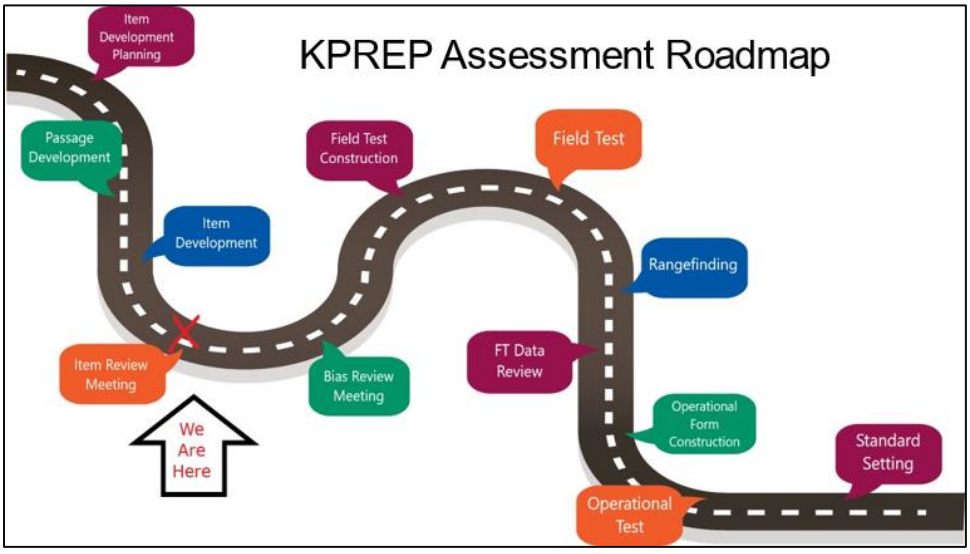
Welcome and Introductions

Reviewer Role

The role of each reviewer is to offer your professional perspective on all items in your assigned item group. Most of the work will be self-paced and individual, but there will be opportunities for discussion as well.

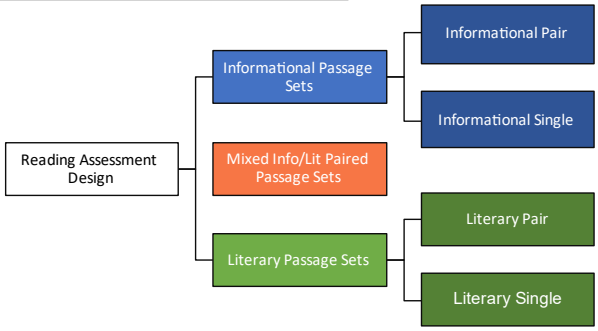
- Be focused
- Provide detailed feedback for each item as needed
- Ask clarifying questions as needed
- Participate in discussions
- Respect the opinions of all involved

Assessment Overview



Assessment Overview

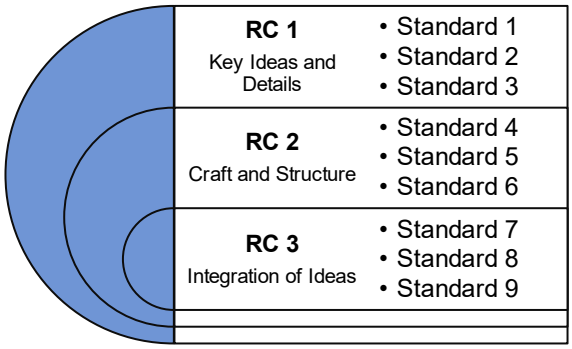
Components of the Reading Assessment



Assessment Overview

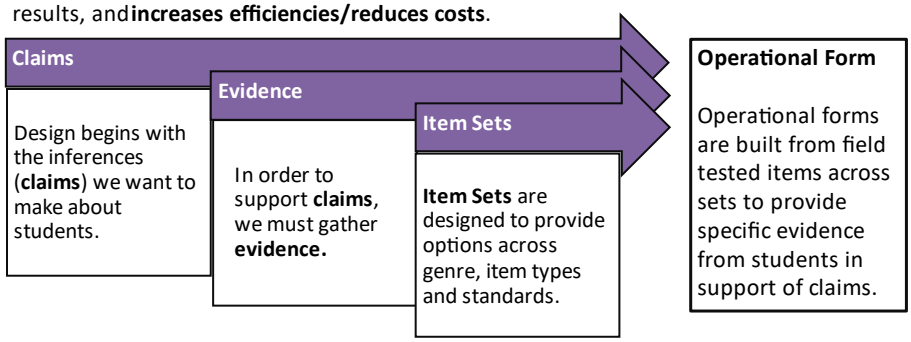
KAS Reporting Categories across Standards

Reading Item Development plans and assessment blueprints are developed around this structure with items separated into three reporting categories.



Evidence-Centered Design (ECD)

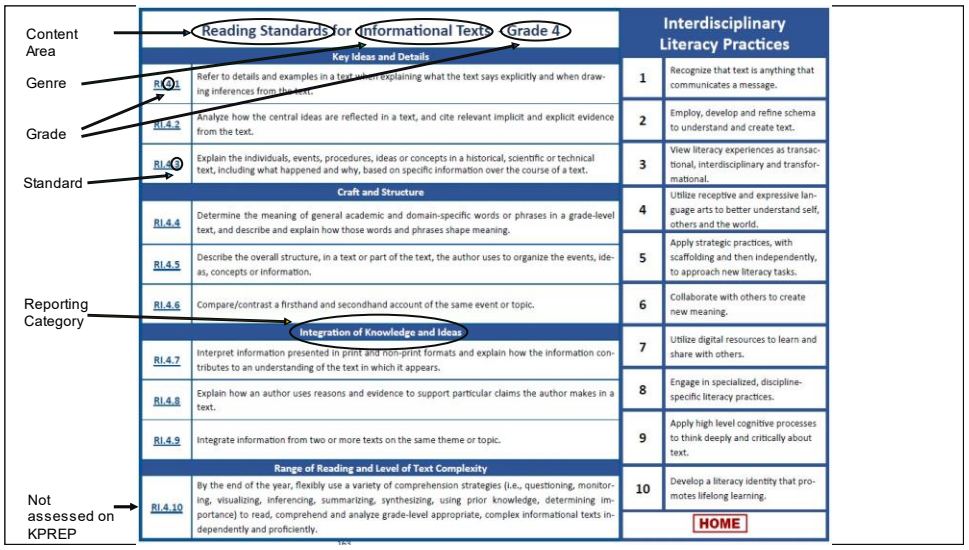
ECD is a deliberate and systematic approach to assessment development that **establishes the validity** of the assessments, **increases the comparability** of year-to-year results, and **increases efficiencies/reduces costs**.

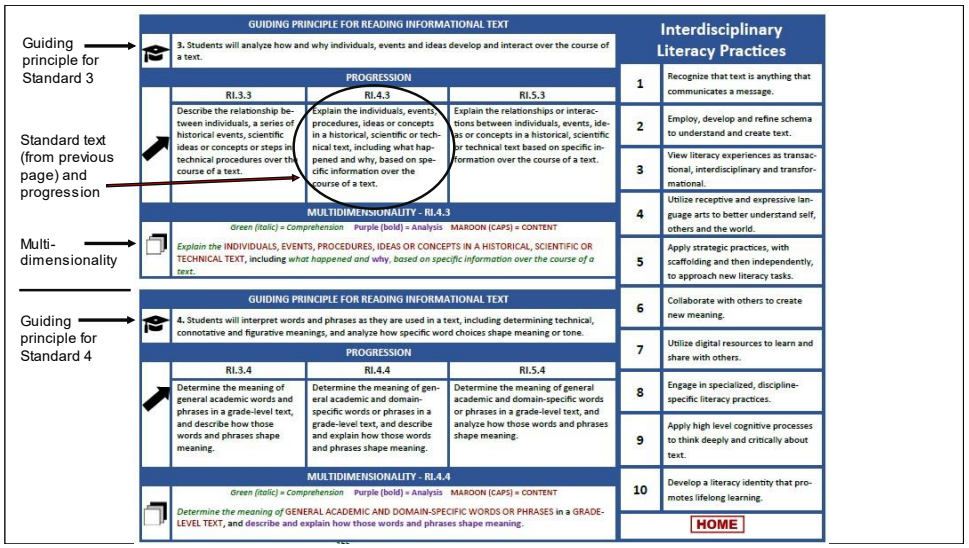


Standards

KAS: What are the Reading Standards?

- Describe what a student needs to be able to do to show mastery
- Targeted to both literary and informational passages
- Provide for a range of teaching and assessment options
- Multi-faceted allowing for some standards to be assessed across several items





Kentucky Item Types

- **Multiple Choice Items (MC)**
- **Multiple Select Items (MS)**
- **Technology-Enhanced Items (TE)**
- **Short Answer Items (SA)**
- **Extended Response Items (ER)**

Item Types

Multiple Choice (MC) Items

1 point

Directions: Read both passages and answer the following questions.

from Lewis and Clark's Journey of Discovery

by Judith Edwards

Originally published in 1999

In 1804, President Thomas Jefferson tasked Meriwether Lewis and William Clark with leading an expedition to explore the territory acquired in the Louisiana Purchase of 1803. The goals of the expedition were to find a way across the western part of the continent to the Pacific Ocean, to make contact with the Native American tribes there, and to map the new territory. This excerpt describes the expedition's quest to locate the land

How did Sacagawea's presence **most** influence the expedition?

- A. Her familiarity with the territory helped to guide the expedition.
- B. Her experience on similar expeditions assured the success of the expedition.
- C. Her understanding of the Shoshone language helped with communication about the expedition.
- D. Her reassurances that the group was close to the Shoshone camp provided comfort during the expedition.

Item Types

Multiple Select (MS) Items

2 points

Directions: Read both passages and answer the following questions.

from Lewis and Clark's Journey of Discovery

by Judith Edwards

Originally published in 1999

In 1804, President Thomas Jefferson tasked Meriwether Lewis and William Clark with leading an expedition to explore the territory acquired in the Louisiana Purchase of 1803. The goals of the expedition were to find a way across the western part of the continent to the Pacific Ocean, to make contact with the Native American tribes there, and to map the new territory. This excerpt describes the expedition's quest to locate the land

Which pieces of evidence from the passage **best** support the inference that Lewis and Clark urgently needed to find the Shoshone camp? Select **two** correct answers.

- A. "The men were using their tow lines and poles constantly." (paragraph 1)
- B. "The cliffs were twelve hundred feet high. . . ." (paragraph 1)
- C. ". . . game was becoming scarce." (paragraph 2)
- D. ". . . a beaver apparently gnawed on the green willow. . . ." (paragraph 2)
- E. "Twenty-one days had passed since the expedition left. . . ." (paragraph 2)

Item Types

Technology Enhanced (TE) Item 1

Directions: Read both passages. Then answer the following questions.

from Streams to the River, River to the Sea

by Scott O'Dell

This novel about the Lewis and Clark expedition is told from the perspective of Sacagawea, who was born into the Shoshone tribe but who has lived with the Mandan tribe for many years. Here, she narrates their search for the land of her people, which she has not seen since her childhood.

1 We reached the place above the falls that Captain Clark had marked with stakes and little flags. Here the canoes were put in the water, much to our delight, for the portage had been hard on everyone.

2 Clothes and food and all the provisions were loaded into the canoes. The men got out their ropes and poles and we went on toward the Shining

Directions: Move each answer into the correct box in the table.

Move **each** setting detail into the correct box to match it with the description that **best** shows its influence on the plot of the passage.

reminds Sacagawea of her home

prompts concern that the Shoshone are on alert

inspires Sacagawea to move faster

forces Sacagawea to travel by land

Setting Detail	Influence
"Only the mountains with snow on them,' . . ." (paragraph 7)	
". . . bits of thin smoke drift up from a grove of pine trees." (paragraph 10)	
". . . the print of a man's moccasin, a ring of cold ashes, wisps of smoke, a pointed quill . . ." (paragraph 18)	
". . . round, blue stones that covered the river bottom . . ." (paragraph 25)	

Item Types

Technology Enhanced (TE) Item 2

from Lewis and Clark's Journey of Discovery

from Streams to the River, River to the Sea

Directions: Read both passages and answer the following questions.

from Lewis and Clark's Journey of Discovery

by Judith Edwards

Originally published in 1999

In 1804, President Thomas Jefferson tasked Meriwether Lewis and William Clark with leading an expedition to explore the territory acquired in the Louisiana Purchase of 1803. The goals of the expedition were to find a way across the western part of the continent to the Pacific Ocean, to make contact with the Native American tribes there, and to map the new territory. This excerpt describes the expedition's quest to locate the land of the Shoshone tribe along the Missouri River, guided by a Shoshone woman named Sacagawea.

1 Navigating the river was increasingly difficult. The men were using their tow lines and poles constantly. On July 19 the party

Directions: Select all the correct answers.

Identify one plot element of each passage and two plot elements of both passages to show how the authors emphasize different information in their accounts.

Plot Element	from *Lewis and Clarke's Journey of Discovery*	from *Streams to the River, River to the Sea*	Both
Relates challenges of traveling down the river	○	○	○
Explains the reason for the camp's location	○	○	○
Describes the discovery of the Shoshone summer camp	○	○	○
Demonstrates a consequence of the party separating	○	○	○

Item Types

Technology Enhanced (TE) Item 3

Directions: Read both passages. Then answer the following questions.

from Frankenstein: Prodigal Son

by Dean Koontz

Erika is a human-like being created by Victor Frankenstein, who plans to replace all naturally-born humans with an artificial, immortal species called the New Race. After she thinks she sees something moving in her room at night, she goes to the house library to read.

1 Comfortable in her robe, ensconced in a wing-back chair, Erika spent the night and the morning with no company but books, and even took her breakfast in the library.

2 Reading for pleasure, lingering over

Directions: Select all the choices that correctly answer the question.

Which phrases in paragraphs 5 and 6 **best** provide context for the meaning of "to eschew emotion"? Select **two** correct answers.

5 Erika understood the concept of love and found it appealing, but she didn't know if she would ever *feel* it. The New Race was supposed to value reason, to eschew emotion, to reject superstition.

6 She had heard Victor say that love was superstition. One of the Old Race, he'd made himself New. He claimed that perfect clarity of mind was a pleasure greater than any mere sentiment.

Item Types

Technology Enhanced (TE) Item 4

4 Jack pines grow back fast. In a few years each jack pine will be about as tall as a kitchen table, and the burned patch will look like a miniature forest. Growing quickly together in the sun, the jack pines will crowd out all other trees.

Directions: Complete the paragraph by selecting the correct phrase from the drop-down menus.

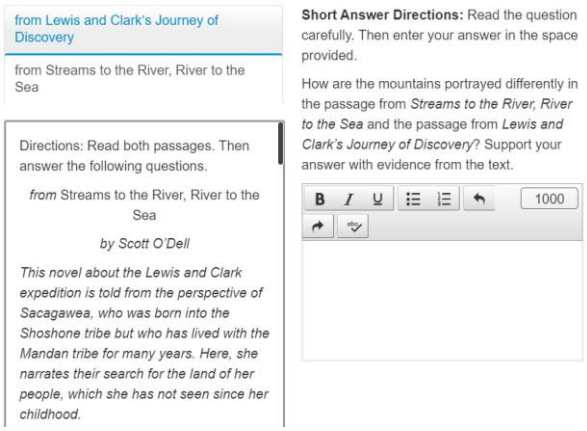
Complete the paragraph that explains the purpose of the first photograph.

The first photograph helps the reader understand Choose... .

This photograph also makes it clear that the hidden seeds need assistance in order to reach the soil. This helps the reader understand the role Choose... play in helping to renew the forest.

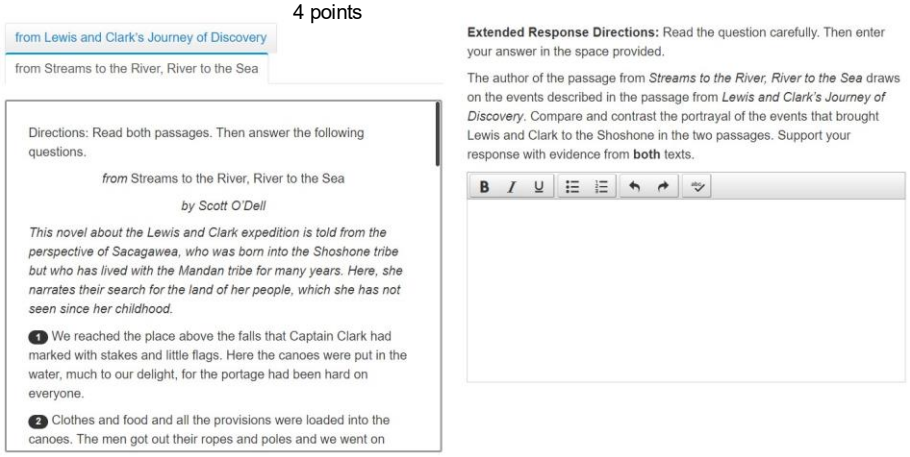
Item Types — Short Answer (SA) Items

2 points



Item Types — Extended Response (ER) Items

4 points



Emphasis on Item Simplification

- This program is currently in the process of creating an item bank at all grade levels
- We are focused on increasing the number of both accessible and complex items, targeting cognitive levels 2 and 3
- Item simplification includes:
  - straightforward language in stems and answer choices
  - concise ER and SA prompts; reducing wordiness
  - reducing the number of interactions in TEs when appropriate
  - MS items limited to five options with two keys

Content Review: Review Process

The role of the Content Reviewer is to provide expert content review of items within assigned passage sets.

- Review item sets assigned to you using Item Review Criteria
- Assign Item Status
  - Accept— Recommend the item be approved as it is
  - Accept with Edits— Recommend the item be approved with edits suggested for improvement:
    - Could be a content edit, edit to standard alignment, edit to functionality, etc.
  - Reject— Recommend the item NOT be approved; fatal flaws prevent any ability to revise

Content Review: Role of the Reviewer

Please note what is NOT the role of the Content Review committee

- Bias/Sensitivity Item Review committees will review all items next week using bias/sensitivity guidelines; that is not the responsibility of this committee
  - Reviewers may note bias -related concerns for a passage or items, but review focus must be on content of the items themselves
- Texts cannot be rejected/revised at this stage
  - Reviewers may note egregious errors/typos within passages
  - Reviewers may note concerns with passage content, but review focus must be on items themselves

Item Review: Materials

The following documents will be available to reviewers:

- ELA Item Reviewer Training PowerPoint
- Guiding Questions/Item Review Criteria
- Kentucky Standards Document
- Technology Enhanced Item Scoring Guides
- SA and ER Scoring Rubrics

Item Review: Process

Committee Item Review Process

1. Determine Item Review Assigned Group (A -F).
2. Navigate in ABBI to grade level and filter by item sequence (A -F).
3. Sort by item sequence.
4. Begin with first item in the group.
5. Read passage, then review items using review checklist.
6. Vote on each item in ABBI.
7. Enter comments (if any) to identify issues and/or offer recommendations for resolution.
8. Facilitator will review votes and comments in live time and discuss trends with the group as needed.

Item Review Criteria/Guiding Questions

1. Standard Alignment:
   - Does the item allow for students to demonstrate mastery of the aligned standard?
2. Content Appropriateness:
   - Is the content of the item clear, concise, and appropriate for the intended grade level?
3. Key and answer options:
   - Is the keyed answer the only correct option?
   - Are distractors plausible and mutually exclusive?
4. Item construction and functionality:
   - Is the item constructed with appropriate grammar and syntax across all elements?
   - If the item has a technology-based stimulus or requires a technology-based response, is the technology design effective and grade-level appropriate?
   - Does the item function correctly?

Criterion 1: Alignment to the Standards

Items should:

- Reflect the language of the standard as appropriate
- Assess only one standard
- Align to part or all of a standard

Note: It may require multiple items to assess the full standard

Criterion 1: Alignment to the Standards (Vocabulary)

Vocabulary items should:

- Allow for context to help determine meaning
- Focus on language meaning and impact, not simple definitions

Aligned to standard	Unaligned to Standard
How does the author's use of the word "cadence" impact the meaning of the passage?	What does the word "cadence" mean as it is used in paragraph 6?

Criterion 2: Content Appropriateness

Items should:

- Reflect the reading level for the tested grade
- Require appropriately complex thinking and problem solving
- Assess topics and concepts that adhere to grade level learning

31

Criterion 2: Content Appropriateness

Language and complexity must be appropriate for the tested grade level.

Appropriate for elementary level	Too complex for elementary level
Which detail from the passage **best** supports the idea that Jamela's family and friends were frightened when they could not find her?	Which quotation **best** implies that Erika has begun to feel conflicted about Victor's plans for a revolution?

Criterion 3: Key and Answer Options

- Answer options are parallel and equally plausible
- Distractors are independent from the others
- Only one option is correct for MC items

Parallel Item Examples

How does the phrase "One bright morning" in paragraph 3 shift the tone of the passage?

- A. From cautionary to intense
- B. From anxious to hopeful
- C. From serious to familiar
- D. From tragic to playful

How does the beaver mentioned in paragraph 2 influence events in the passage?

- A. By preventing Clark from receiving Lewis's warning
- B. By providing a sign that the Shoshone camp was nearby
- C. By downing a tree that made navigating the river more difficult
- D. By encouraging Clark's party to wait for Lewis where game was plentiful

Criterion 3: Key and Answer Options

Item lacking parallelism

What is the impact of the phrase "enforced idleness" as it is used in paragraph 12 on meaning in the speech?

- A. It reinforces the idea that efforts to solve the problem of unemployment have not be exhausted
- B. It removes the responsibility for unemployment from those who would prefer to work
- C. It offers a remedy to the emotional problems associated with unemployment
- D. Access to public relief is denied

Item with options that are the opposite of one another

How do the rhetorical questions in paragraphs 7 and 8 **best** advance President Johnson's purpose?

- A. By leading the audience to reject the importance of connecting people with nature
- B. By leading the audience to consider the value of connecting people with nature
- C. By leading the audience to doubt there is danger in permitting children to venture into the wilderness
- D. By leading the audience to understand the danger of allowing children to venture into the wilderness

Criterion 3: Key and Answer Options

Items should avoid **internal** clueing or miscues:

- answer options should NOT repeat or echo a word used in the stem

Items should avoid **external** clueing or miscues:

- items should not be answerable using other items in the set
- other items in the set should not mislead students toward selecting the wrong answer option for any given item

Criteria 4: Item construction and functionality

Technology-based items:

- use of technological format must be justifiable and relevant; should not duplicate the logic/structure of an MC item
- allow for a variety of technology-enhanced student responses with a limited subset of correct responses

Items that address graphics should:

- Be aligned to specific standards that support such an analysis
- Analyze how the graphics support the purpose of the passage

Criteria 4: Item construction and functionality

All items:

- Are conceptually, grammatically, and syntactically consistent between the stem and answer choices, and among answer choices
- Function and score correctly in ABBI

Criteria 4: Item construction and functionality

Effective use of TE capabilities to select multiple options

Identify **one** plot element of **each** passage and **two** plot elements of **both** passages to show how Koonz used *Frankenstein* as a source for *Prodigal Son*.

	from *Frankenstein*	from *Prodigal Son*	Both
The main character learns the value of emotion.	○	○	○
Curiosity leads to a search for answers.	○	○	○
The main character makes a scientific discovery.	○	○	○
A scientific process allows creating new life.	○	○	○

Fails to add more value than an MC item

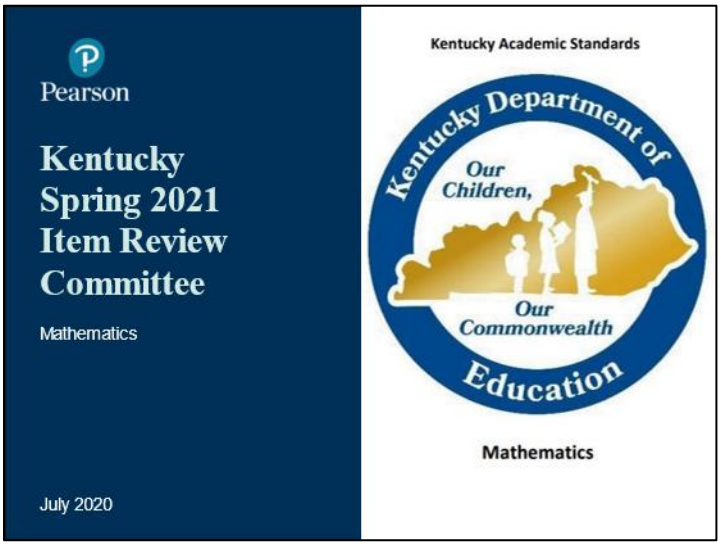
Select one option to indicate which character demonstrated courage in a challenging situation.

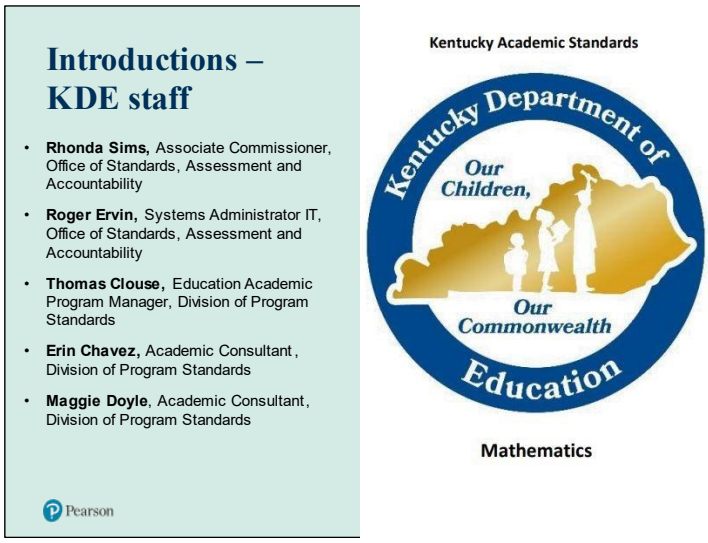
	Demonstrated Courage
Charles Martin	○
Abigail Rose	○
Kaleigh Sue	○
Miles Griffyn	○

Next Steps

- Item Review Group Assignments
- ABBI Training
- Begin Review

Appendix G. Mathematics Item Content Review Training





Introductions – Content Development staff

- **Adrian Rivera**, Pearson, Test Development Manager
- **Jennifer Ramirez**, Pearson, Math Content Lead
- **Jiselle Jones**, Pearson, Math Content

Pearson

Kentucky's Vision for Students

"Each and every student is empowered and equipped to pursue a successful future."

14

Math Item Review Schedule

Two Week Independent Review Window from July 24th – August 6th with Office Hour Options

Date	Time
Friday, July 24th (1 hour)	11 am – 12 pm EST (10 – 11 am CST)
Monday, July 27th	3 – 3:30 pm EST (2 – 2:30 CST)
Wednesday, July 29th	11 – 11:30 am EST (10 – 10:30 am CST)
Friday, July 31st	3 – 3:30 pm EST (2 – 2:30 CST)
Tuesday, August 4th	11 – 11:30 am EST (10 – 10:30 am CST)

Pearson

15

Meeting Security

Non-Disclosure

- When you accepted the invitation for the meeting, you signed a Non - Disclosure Agreement that specifies you will not share or discuss the content of the items you will be reviewing with anyone outside your meeting.
- Sharing any specifics about the items you will be reviewing is not permitted.
- Examples of information that should not be shared include:
  - The standards and the number of items that were developed to each
  - The contexts/situations used in the items
  - The phrasing of the questions and the format of items
  - The correct answers and rubrics

Environment

- Please remain in a secure, private work area during your review and during any meeting times.
- Work areas should be in location where your computer screen will not be visible.
- During any meeting times, please make sure you are in an area where conversations about item specifics will not be audible to anyone except you.

Pearson 16

Meeting Security

Materials

- Taking screen shots or printing any passages or items is not permitted.
- Non-secure materials will be available to print or download to your computer desktop if/as you deem necessary.
- If you print any of these materials, please keep them in a secure place during the review period and must be shredded at the conclusion of the review.
- If you download any of these materials to your computer, please delete them from your desktop at the conclusion of the review.
- Any notes or scratch paper used during your review must be shredded at the conclusion of the review.

What can you share with non-meeting participants?

- Information about the test development process.
- General descriptions/impressions of your meeting.

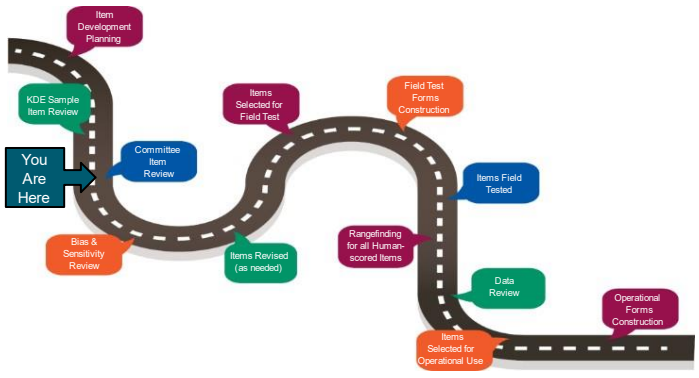
Pearson 17

Table of Contents

18

Item Development

The Journey of an Item



Pearson

10

Math Item Types

Item Type	Point Values	ABBI Element	Scoring Method
Multiple choice (MC)	0 or 1	• Choice	• Machine scored
Technology Enhanced (TE) or Fill-in-the-blank (FIB)	0 or 1	• Variety of Elements	• Machine scored
Multiple Select (MS)	0, 1, or 2	• Choice	• Machine scored
Short Answer (SA)	0, 1, or 2	• Technology Enhanced Parts • Equation Editor	• Machine scored • AI scored • Human scored
Extended Response (ER)	0, 1, 2, 3, or 4	• Technology Enhanced Parts • Equation Editor	• Machine scored • AI scored • Human scored

Pearson

11

Cluster Sets

- Allows for multiple item types (assessing different standards) to share a common context.
- Each item part is independent of the other item part(s).
- A variety of item types can be used in the cluster set.
  - Examples:
    MC, MC, MC
    MC, MS, SA
    MS, MC, ER
- Cluster items can be identified by their UINs.

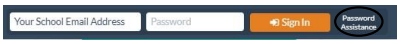
Clusters	Grade 6
Stimulus	MA0620C1_00
Item 1	MA0620C1_01
Item 2	MA0620C1_02
Item 3	MA0620C1_03
Item X	MA0620C1_XX

Pearson

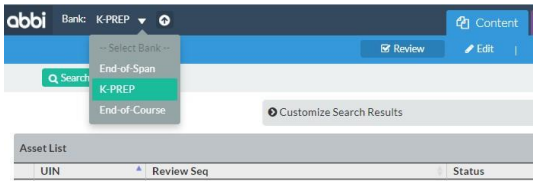
12

ABBI - Login Instructions

- Go to https://abbi.pearson.com
- Use your school email address as your username.
- Click on Password Assistance and follow the steps to set up your password.



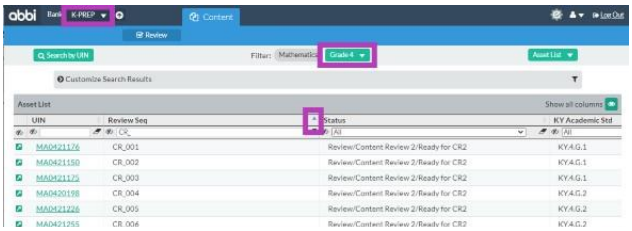
- Note: ABBI works best with Google Chrome and Microsoft Edge.
- If you are reviewing grades 3-8, you will be working in the K-PREP bank. Grade 10 will be in the End-of-Span bank.



Pearson

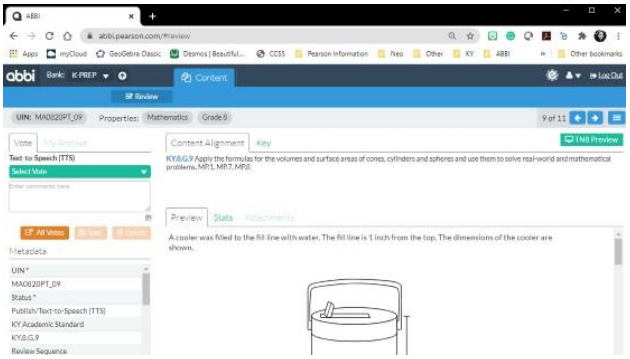
ABBI – Navigating through the items

- ABBI Main Asset List
- All items available for review in a grade will be listed here.
- The items are Sequenced by standard in the Review Seq Column
  - Click on the arrow shown below by Review Seq to sequence the items in order starting with 001.
- Click on the UIN of each item to review the item.



Pearson

ABBI – Navigating through the items



Committee Item Review Steps

Kentucky Academic Standards for Mathematics

Statistics and Probability	
Standards for Mathematical Practice	
MP.1. Make sense of problems and persevere in solving them.	MP.5. Use appropriate tools strategically.
MP.2. Reason abstractly and quantitatively.	MP.6. Attend to precision.
MP.3. Construct viable arguments and critique the reasoning of others.	MP.7. Look for and make use of structure.
MP.4. Model with mathematics.	MP.8. Look for and express regularity in repeated reasoning.
Cluster: Develop understanding of statistical variability.	
Standards	**Clarifications**
KY.6.SP.1 Recognize a statistical question as one that anticipates variability in the data related to the question and accounts for it in t answers. MP.1, MP.3, MP.6	For example, "How old am I?" is not a statistical question, but "How old are the students in my school?" is a statistical question because one anticipates a variety of values with associated variability in students' ages. Coherence KY.5.MD.2→KY.6.SP.1→KY.7.SP
KY.6.SP.2 Understand that a set of numerical data collected to answ a statistical question has a distribution which can be described by its center, spread and overall shape. MP.2, MP.6, MP.7	Students distinguish between graphical representations which are skewed or approximately symmetric; use a measure of center to describe a set of data. Coherence KY.5.MD.2→KY.6.SP.2→KY.7.SP
KY.6.SP.3 Recognize that a measure of center for a numerical data s summarizes all of its values with a single number to describe a typi value, while a measure of variation describes how the values in the distribution vary. MP.2, MP.5, MP.6	Emphasis is on the sensitivity of measures of center to changes in t data, such as mean is generally much more likely to be pulled towa an extreme value than the median. Additionally, measures of varia (range, interquartile range) describe the data by giving a sense of t spread of data points. Coherence KY.6.SP.3→KY.7.SP
Attending to the Standards for Mathematical Practice	
Students recognize a question such as "What did I eat for breakfast?" is not a statistical question, whereas "What is the popular breakfast in my school?" will elicit data they can measure precisely (MP.6) and draw conclusions based on that data (MP.3). After collecting data, by creating a distribution of that data, students recognize data generally follows a structure and can be described in terms of structure (MP.7). By accurately calculating the mean (or any other statistical measure), students are now more precise in describing data, for example, describe the rainfall for the month as "about average" to "the rainfall this month is slightly higher than that of the last 10 years and within the interquartile range for that data" (MP.6)	

Link to Kentucky Academic Standards

Overview of Review Steps

- Step 1 • Work each item independently
- Step 2 • Score the item
- Step 3 • Verify alignment to the Kentucky Academic Standards (KAS)
- Step 4 • Verify alignment to Target of the Standard
- Step 5 • Verify Mathematical Practice(s) Alignment
- Step 6 • Verify Cognitive Complexity
- Step 7 • Review item, rubrics, and rationales for errors or concerns
- Step 8 • Vote in ABBI

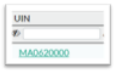
Pearson

18

Step 1 – Work Each Item Independently

Each participant has been assigned a grade level.
- The items are sequenced by KAS Standard then item type.
- You can find the items available to you the ABBI Asset List.

- Select the UIN of the item you wish to review.



- The preview screen shows a preview of the item, any rubrics that the item has, and the metadata (e.g., key, calculator selection, cognitive complexity, mathematical practices, etc.)
- To see the item how a student will see it, select TN8 Preview.

- Work the item.

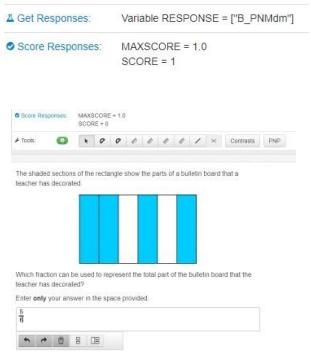
  Note: All scratch paper will need to be securely shredded at the end of your review.

Pearson

19

Step 2 – Score the Item

- Select your answer in TN8 Preview.
- Select Score Responses.
- MAXSCORE reflects the total number of points possible.
- Score reflects the total number of points scored as correct.
  - Only machine-scored items will reflect a number other than "0" for the SCORE.
  - Equation Editor items will not score in ABBI.
- Verify that the correct response matches the listed key, rationale or rubric(s) for the item.



Pearson

20

Step 3 – Verify Alignment to KAS

Use the online version of the Kentucky Academic Standards to verify the following:

- Item aligns to the KAS indicated.
- Item is written to the appropriate target for the standard (conceptual understanding, procedural skill/fluency, application).
- Use the Coherence within the KAS to examine connections to the same topic in previous grades to ensure the task is crafted to elicit a more sophisticated level of understanding than would have been acceptable in the previous grade?
- The numbers/number types and types of representation (whether the area model, shapes, graphs, functions, etc.) match those called for by the targeted standard and those appropriate for the grade level.

Pearson

21

Step 3 – Verify Alignment to KAS continued...

- Unlike classroom assessment items, items used on the Kentucky State Assessment can only report out (align) to one Kentucky grade level mathematics standard.
- Some standards are more robust than other standards, so it may not always be possible to assess all parts of a standard in a single item. As the bank gets healthier, the intent is to have a bundle of items that collectively assess all parts of the standard.

Standard of Mathematical Content	Clarifications & Coherence	Attending to the Standards for Mathematical Practice
• Defines what students should understand and be able to do. • When possible, the full intent of a standard is assessed. • For the more robust standards, the items aligned to the standard collectively meet the full intent of the standard. • Look to see if there is a coherent connection to the same topic in a previous grade or to another grade-level standard.	• Communicates the expectations more clearly and concisely to teachers, parents, students and stakeholders through examples and illustrations • Provides guidance on how that content standard connects to others within and across grade levels	• Defines how students engage in mathematical thinking. • Items provide meaningful opportunities for students to engage in the standards for mathematical practices

Pearson

22

Step 4 – Verify Alignment to Target of Standard

- Consider: If the standard is **conceptual understanding**, does the task require more than knowing isolated facts and methods? Are students asked to make sense of why a mathematical idea is important and the kinds of contexts in which it is useful?
- Consider: If the standard is **procedural skill/fluency**, does the task require students to apply procedures accurately, efficiently, flexibly and appropriately? Does the task focus students' attention on the use of procedures for the purpose of developing a deeper level of understanding of mathematical concepts or ideas? If general procedures may be followed, can they be followed mindlessly or are students asked to engage with the conceptual ideas that underlie the procedures to complete the task successfully?
- Consider: If the standard is **application**, does the task offer students the opportunity to solve problems in a relevant and meaningful way? Are students asked to select an efficient method to find a solution and develop critical thinking skills? Are students asked to actively examine task constraints that may limit possible solutions and strategies?

Pearson

23

Step 5 – Verify Mathematical Practice(s)

- Does the item give the student an opportunity to engage with at least one mathematical practice at the appropriate level of depth required by the standard?

  Note: Each cluster within the KAS for Mathematics has these bookmarked to the descriptions in the front matter of the standards document AND has an Attending to the Standards for Mathematical Practice component.
- Verify that the item assesses the Mathematical Practices selected.

Standards for Mathematical Practice	
MP.1. Make sense of problems and persevere in solving them. MP.2. Reason abstractly and quantitatively. MP.3. Construct viable arguments and critique the reasoning of others. MP.4. Model with mathematics.	MP.5. Use appropriate tools strategically. MP.6. Attend to precision. MP.7. Look for and make use of structure. MP.8. Look for and express regularity in repeated reasoning.

Attending to the Standards for Mathematical Practice

Students compare the value of the digits based on where they are in a number (**MP.7**). They reason 10 tens equal 100, 70 tens equal 700 and this can be illustrated with base 10 blocks or other visuals (**MP.2**). Students look across series of problems to notice a pattern when multiplying by 10, 100 or 1000 (**MP.8**) and justify why patterns exist (why 36 x 100 = 3600), rather than superficially noting 'you add zeros,' they explain or show there are actually 36 *hundreds*, so 3600 (**MP.3**). Students use similar reasoning to compare decimal values, explaining tenths are larger than hundredths and therefore, they look to first see which values have more tenths before looking at how many hundredths it has (**MP.2, MP.7**). Students use tools such as number lines and base 10 blocks to see place value relationships with decimals in order to compare and to round (**MP.5**).

Pearson

I 24

Step 6 - Verify Cognitive Complexity

Table 2: Levels of Complexity

	Level 1	Level 2	Level 3
Procedural Complexity:[13]	Solving the problem entails little procedural[14] demand or procedural demand is below grade level.	Solving the problem entails common or grade-level procedure(s) with friendly numbers.	Solving the problem requires common or grade-level procedure(s) with unfriendly numbers,[15] an unconventional combination of procedures, or requires unusual perseverance or organizational skills in the execution of the procedure(s).
Conceptual Complexity:[16]	Solving the problem requires students to recall or recognize a grade-level concept. The student does not need to relate concepts or demonstrate a line of reasoning.	Students may need to relate multiple grade-level concepts or different types, create multiple representations or solutions, or connect concepts with procedures or strategies. The student must do some reasoning, but may not need to demonstrate a line of reasoning.	Solving the problem requires students to relate multiple grade-level concepts and to evidence reasoning, planning, analysis, judgment, and/or creative thought OR work with a sophisticated (nontypical) line of reasoning.
Application Complexity:	Solving the problem entails an application of mathematics, but the required mathematics is either directly indicated or obvious.	Solving the problem entails an application of mathematics and requires an interpretation of the context to determine the procedure or concept (may include extraneous information). The mathematics is not immediately obvious. Solving the problem requires students to decide what to do.	In addition to an interpretation of the context, solving the problem requires recognizing important features, and formulating, computing, and interpreting results as part of a modeling process.

Source: achieve.org; table 2
https://www.achieve.org/files/Cognitive%20Complexity%20Mathematics%20Assessment_FINAL_0.pdf

Pearson

I 25

Step 7 – Review Item, Rubrics, and Rationales for Errors or Concerns

- Item is conceptually, grammatically, and syntactically consistent between the stem and all answer choices.
- Has answer choices that are plausible and attractive to the student who has not mastered the objective or skill.
  - Multiple Choice, Multiple Select, and Inline Choice will contain rationales that can be seen in ABBI (not in the TN8 Preview) by hovering your cursor over each option.
  - All other items contain rubrics which can be seen on the Preview Screen
- Item does not provide cues (intentionally or unintentionally) for how to approach finding a solution.
- KAS Focus (found in the metadata) matches the item and accurately describes which part or parts of the standard are being assessed.

Pearson

I 26

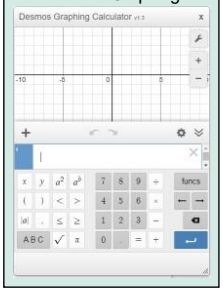
Step 7 – Other considerations

- Calculator - There are three calculator selections available in ABBI.
  - Yes – A calculator should be used.
  - No – A calculator should not be used.
  - Z – Neutral.

Grades 3-5
Desmos Four Function

Grades 6-8
Desmos Scientific

Grade 10
Desmos Graphing



Pearson

27

Step 7 – Other considerations

- As you review, please keep in mind that a variety of items is needed for each of the following:

Target of the Standard	Cognitive Complexity	Relevance
• Conceptual understanding • Procedural skill & fluency • Application	• Low • Medium • High	• Items give students an authentic opportunity to connect content standards to real-world issues and/or contexts.

Pearson

28

Step 8 – Vote in ABBI

From the dropdown menu select your vote for the item.	
Accept	No concerns
Accept with edits	Minor edits required
Accept with reconciliation	Alignment/Scoring/Context Concerns
Reject	Acomplete rewrite is required

In the Comment note any of the following concerns:

- Standard of Mathematical Content alignment
- Alignment to the target of the standard
- Cognitive complexity
- Standard for Mathematical Practice alignment
- Relevance
- Keys/Rubrics
- Errors
- Precision
- Grammar
- Use of technology

Select "Save" to submit your vote and comments.

Pearson

29

Procedural - low

Grade 4 - KY.4.NBT.2.a

The speed of light is about 186,282 miles per second. Which number name can be used to represent 186,282?

A. one hundred eighty-six thousand, two hundred eighty-two

B. one hundred thousand, eighty-six hundred eighty-two

C. eighteen thousand, sixty-two hundred eighty-two

D. eighteen hundred, sixty-two thousand eighty-two

Pearson

Procedural - Medium

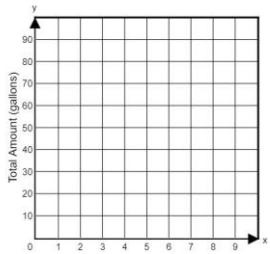
Grade 7 - KY.7.RP.2.b

Part A

Pete puts water into an empty pool. After 8 minutes there are 40 gallons of water in the pool. Use the coordinate plane provided to plot two points that can be used to represent the relationship between time, x, and the total amount of water, y, in the pool.

Select the places on the coordinate plane to plot the points.

Water in a Pool

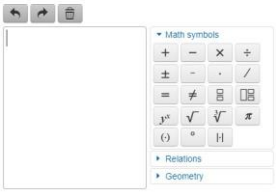


Part B

Using your points from Part A:

- What is the constant of proportionality in terms of the context? Show your work or explain how you determined your answer.
- What is the total amount of water, in gallons, in the pool after 47 minutes? Show your work or explain how you determined your answer.

Enter your answers and your work or explanations in the space provided.



Pearson

Procedural - High

High School - KY.HS.G.6

In triangle *PSW*, the measure of $\angle P$ is $(3x^2 - 12x + 55)^\circ$, the measure of $\angle S$ is $(2x + 25)^\circ$, and the measure of $\angle W$ is $(^-2x^2 - 4x + 149)^\circ$. Which statements about triangle *PSW* are true?

Select **two** correct answers.

- ☐ A. Side $\overline{PS}$ has the longest length of the three sides.
- ☐ B. Side $\overline{SW}$ has the shortest length of the three sides.
- ☐ C. The measure of $\angle W$ is less than the measure of $\angle S$.
- ☐ D. The length of side $\overline{PS}$ is greater than the length of side $\overline{PW}$.
- ☐ E. The length of side $\overline{SW}$ is greater than the length of side $\overline{PW}$.

Pearson

Conceptual – Low

High School - KY.HS.A.20.a

A system of equations is shown.

$$\begin{cases} 3x - 2y = 7 \\ 4x - y = 13 \end{cases}$$

Which equation could replace $4x - y = 13$ without changing the solution to the system?

- ○ A. $^-12x - 3y = 39$
- ○ B. $^-12x - 4y = 52$
- ○ C. $^-12x + 3y = ^-39$
- ○ D. $^-12x + 4y = ^-52$

Pearson

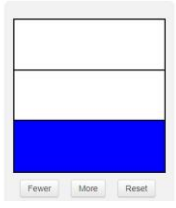
Conceptual – Medium

Grade 3 - KY.3.G.2

Part A

Divide the rectangle shown into three equal areas by using the More and Fewer buttons.

Then shade the rectangle to represent a unit fraction of the whole by selecting the part or parts.



Part B

Explain which unit fraction can be used to represent the area of the rectangle shaded in Part A.

Enter your explanation in the space provided.

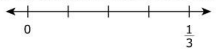
Pearson

Application – Medium

Grade 5 - KY.5.NF.7.c

The number line shows how 4 friends equally shared $\frac{1}{3}$ of a cup of peanuts.

How the Friends Shared the Peanuts



- How much of the peanuts, in cups, did each friend get?
- Explain how the number line can be used to justify your answer.

Enter your answer and your explanation in the space provided.

Math symbols

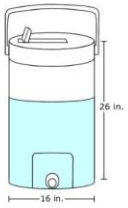
Pearson

Application – High

Grade 8 - KY.8.G.9

Question Information | Formulas

A cooler was filled to the fill line with water. The fill line is 1 inch from the top. The dimensions of the cooler are shown.



After 6 hours, the cooler was $\frac{1}{4}$ full of water.

- What was the average amount of water used each hour?
- Show your work or explain how you determined your answer.

Enter your answer and your work or explanation in the space provided.

Math symbols

Relations

Geometry

Pearson

Application – High

Grade 8 - KY.8.G.9

Question Information | Formulas

Figure	Volume	Surface Area
Cone	$V = \frac{1}{3}\pi r^2 h$	$SA = \pi\left(r + \sqrt{(r^2 + h^2)}\right)$
Cylinder	$V = \pi r^2 h$	$SA = 2\pi rh + 2\pi r^2$
Sphere	$V = \frac{4}{3}\pi r^2 h$	$SA = 4\pi r^2$

After 6 hours, the cooler was $\frac{1}{4}$ full of water.

- What was the average amount of water used each hour?
- Show your work or explain how you determined your answer.

Enter your answer and your work or explanation in the space provided.

Math symbols

Relations

Geometry

Pearson

Contact information

- Please send any general or ABBI questions that **do not contain identifying or specific information about items** to jennifer.ramirez1@pearson.com
- Any questions specifically about items can be brought to one of the office hour sessions offered.

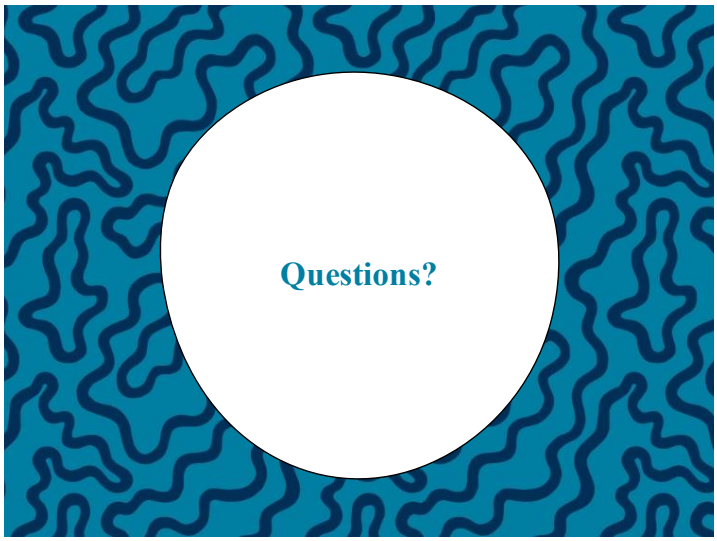
Office Hour Options	
Friday, July 24th (1 hour)	11 am – 12 pm EST (10 – 11 am CST)
Monday, July 27th	3 – 3:30 pm EST (2 – 2:30 CST)
Wednesday, July 29th	11 – 11:30 am EST (10 – 10:30 am CST)
Friday, July31st	3 – 3:30 pm EST (2 – 2:30 CST)
Tuesday, August 4th	11 – 11:30 am EST (10 – 10:30 am CST)

Shared Folder

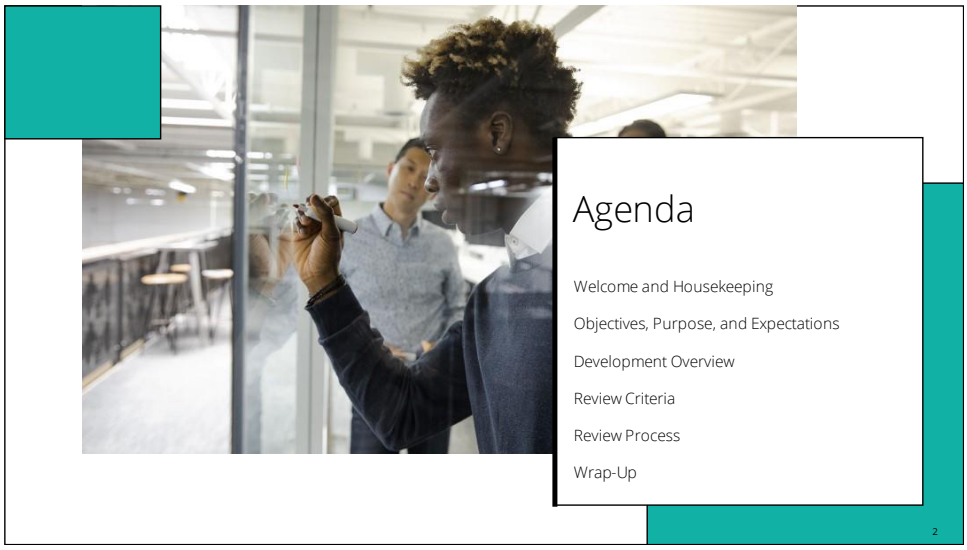
- Following the training, a Microsoft SharePoint Folder will be shared with you.
- This folder contains the materials that we are covering today.
- Please reach out if you are unable to access this folder.

Thank you for participating in this review!

Pearson



Appendix H. Social Studies Item Content Review Training



Welcome and Housekeeping

3

Welcome to Participants

Kentucky Department of Education

- Heather Ransom, Academic Program Consultant
- Lauren Gallicchio, Academic Program Consultant

Pearson

- Adrian Rivera, Senior Test Development Manager
- Sharon Staples, Principal Assessment Specialist

Kentucky Educators

4

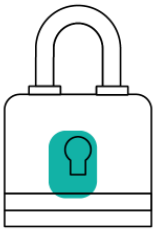
Housekeeping

Confidentiality and Nondisclosure Agreement

- Participants must maintain the security of the assets, documents, and materials being reviewed.
- Participants may not copy, discuss, or disclose in any manner specific information or materials used during this meeting, while reviewing assets, or after the review committee has concluded.
- All materials for the assessment program are the property of the State of Kentucky.

Honorarium

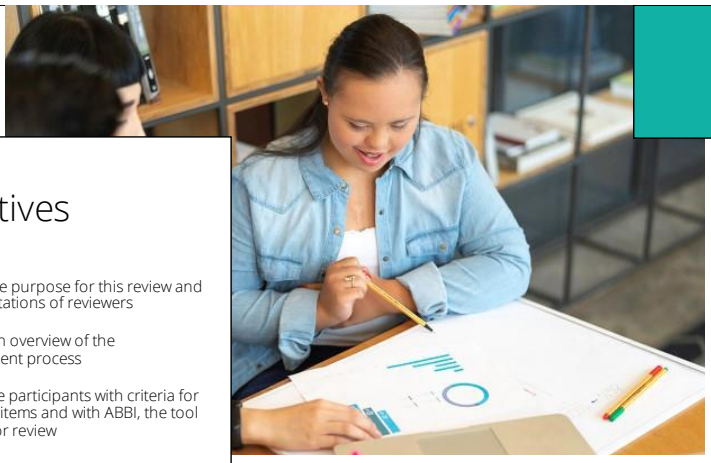
- Pearson provides a link after the review ends
- Processing can occur only after submission of requested information



5

Objectives, Purpose, and Expectations

6



Objectives

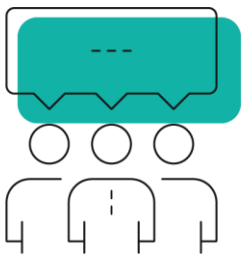
1. Explain the purpose for this review and the expectations of reviewers
2. Provide an overview of the development process
3. Familiarize participants with criteria for reviewing items and with ABBI, the tool needed for review

7

Purpose and Focus

Purpose

An opportunity to advise Pearson and KDE on content concerns in assets



Focus

- Disciplinary standards
- Inquiry practices
- DOK
- Accuracy

8

Expectations

- Choose a secure environment where others cannot see your computer screen
- Choose an environment that is free of distractions
- Avoid multi-tasking
- Complete reviews outside of regular school hours
- Use a secure method of communication when contacting Pearson or KDE
- Refrain from discussing assets except during designated times within the scheduled review window

9

Development Overview

10

Kentucky Statute

Per 158.6453, beginning in fiscal year 2017 –2018, and every six (6) years thereafter, the Kentucky Department of Education (KDE) shall implement a process for reviewing Kentucky's academic standards and the alignment of corresponding assessments.

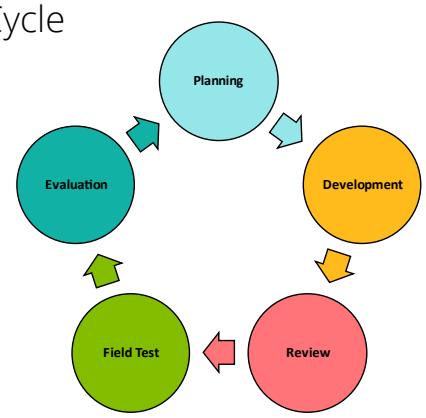
How does this statute affect this review work?

All standards are eligible for assessment item development.

The assessments administered at Grades 5, 8, and 11 are grade-span tests:

Grade 5 test items assess Kindergarten through Grade 5 standards
Grade 8 test items assess Grade 6 through Grade 8 standards
Grade 11 test items assess civics, economics, geography, U.S. history, and world history standards.

Item Life Cycle



12

Disciplinary Content and Inquiry Practices

The KAS "place an equal importance on both the mastery of important social studies concepts and disciplinary practices. . . . As indicated by the graphic on this slide, concept knowledge cannot be achieved effectively without the practice of inquiry. Neither development of the practices nor development of the knowledge and understanding within the lenses is sufficient on its own."

Source: Kentucky Department of Education

13

Blueprint

Domain[1]	Grade 5	Grade 8	Grade 11
Civics	25%	25%	25%
Economics	25%	25%	25%
Geography	25%	25%	25%
History	25%	25%	25%[2]

[1] A minimum of 50% of items for each domain will be dualigned to the inquiry practices of Questioning, Using Evidence, or Communicating Conclusions.

[2] Grade 11 History includes U.S. and world history.

14

Test Components

Standalone Items

- Items that are self -contained rather than part of a set
- All aligned to a single disciplinary standard; some also aligned to an inquiry practice
- Item Types

Multiple Choice (MC):
Four options, one correct answer, 1 pt.
Multiple Select (MS):
Five options, two correct answers, 2 pt.
Technology Enhanced (TE):
Interactive with one or more correct answers, 1 or 2 pt.

Cluster Sets

- Sets of items with shared stimuli
- Single and dual-aligned items
- Sets typically aligned to more than one discipline
- Item Types

MC
MS
TE
Short Answer (SA):
Open-ended, multiple correct answers, 2 pt.
Extended Response (ER):
Open-ended, multiple correct answers, 4 pt.

15

Assessment Goals

KAS	• Aligning items to the intent of the KAS • Providing balanced representation of all disciplines
DEI	• Representing multiple perspectives • Using authentic voices or experts in the field
DOK	• Using a range of cognitive complexity • DOK 1, 2, and 3

16

Review Criteria

17

Item Review Criteria

- **Alignment**
  - Does the item align to the identified disciplinary standard (and inquiry practice)?
  - Does the item measure the intent of the aligned standard (and practice)?
  - Is the DOK accurate and appropriate?
- **Content**
  - Is clear, concise, and grade-level appropriate language used?
  - Is the content grade-level appropriate?
  - Do K–5 items appropriately test the content and theme of the aligned grade-level?
  - Is the stimulus, if present, appropriate?
  - Are all MC and MS distractors plausible?
  - Does the item clue the correct answer or, if it is a cluster item, does it clue another item in the cluster set?
- **Accuracy**
  - Is the accurate?
  - Is the scoring information accurate?
  - Does the asset have spelling or grammatical errors?

NOTE: Review for bias and sensitivity occurs at a different time.

18

Depth of Knowledge (DOK)

1 **Recall and Reproduction**
- Single-step process
- Recall facts
- Locating information
- Performing a basic skill

2 **Working with Skills & Concepts**
- Some reasoning
- Comparing/ contrasting
- Cause and effect

3 **Strategic Thinking**
- Complex or abstract reasoning
- Connecting ideas
- Citing supporting evidence
- Transferring learning
- Synthesizing information

19

DOK 1 Example

This photograph is from the Great Depression.

Source: Public Domain

What name did the public give to communities such as this?

A. Hoovervilles

B. Pittsburghs

C. Potter's Fields

D. Roosevelt Towns

20

DOK 2 Example

This photograph is from the Great Depression.

Source: Public Domain

Why was the term " Hoovervilles" commonly used to describe communities such as this?

A. The public had a negative perception of congressional actions to provide affordable housing.

B. The public perceived the economic policies of the president as ineffective.

C. The public had a negative perception of state plans to raise interest rates.

D. The public perceived programs proposed by economics professors as favoring the rich.

21

DOK 3 Example

This photograph is from the Great Depression.

Source: Public Domain

Why were communities such as this commonly known as Hoovervilles during the 1930s?

A. To imply that the economy should recover on its own

B. To encourage Democrats to develop an economy recovery program

C. To imply that Republicans were responsible for the economic downturn

D. To encourage charitable organizations to fix the economy

22

Review Process

23

Review Resources

- Item Review Criteria Checklist
  - Provided via email
- Kentucky Academic Standards
  - In ABB
  - Available at https://education.ky.gov/curriculum/standards/kyacadstand/Documents/Kentucky_Academic_Standards_for_Social_Studies_2019.pdf
- Glossary of Terms for the Kentucky Academic Standards
  - Available at https://education.ky.gov/curriculum/standards/kyacadstand/Documents/KAS_for_Social_Studies_Glossary_of_Terms.pdf
- High School Disciplinary Clarifications
  - Available at https://education.ky.gov/curriculum/standards/kyacadstand/Documents/High_School_Disciplinary_Clarifications.pdf
- DOK in Social Studies
  - Resources courtesy of Dr. Karin Hess
  - Cognitive Rigor Matrix https://01fd434f-c1b0-45d9-899e-3654cb7c37d5.filesusr.com/ugd/5e86bd_54ed69c317bc4e14a06be7a1b5be7590.pdf
  - Applying Webb's Depth -of-Knowledge (DOK) Levels in Social Studies: https://www.nciea.org/publications/DOKsocialstudies_KH08.pdf

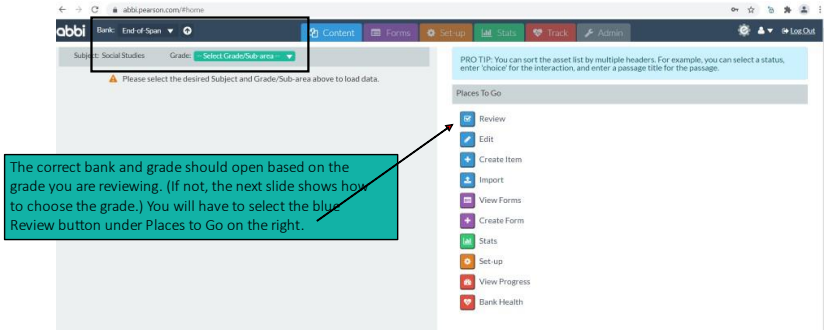
24

ABBI: Logging In

- Using Chrome or Firefox as your browser, access the ABBI site: https://abbi.pearson.com/.
- Log in with the user name and password supplied by Pearson.

25

ABBI: Initial Log-In Page



The correct bank and grade should open based on the grade you are reviewing. (If not, the next slide shows how to choose the grade.) You will have to select the blue Review button under Places to Go on the right.

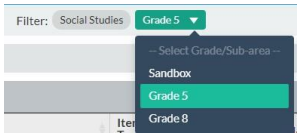
26

ABBI: Asset List

After logging in, use the drop -down menu to select the appropriate grade (5, 8, or 11).

Grades 5 and 8 are in the K -PREP bank.

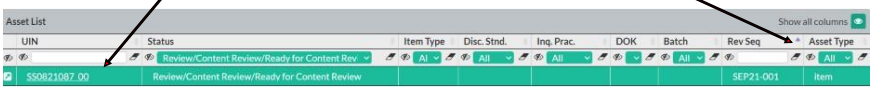
Grade 11 is in the End -of-Span bank.



27

ABBI: Sorting

1. Status: Review>Content Review>Ready for Content Review
2. Rev Seq: Select the "up" triangle to sort into the correct review order.
3. Select the first UIN to navigate to the review screen.



28

ABBI: Navigating

Each asset has navigation tools in the top right -hand corner. The arrows can be used to move to the next asset or to go back to the previous asset.

The list icon returns you to the Asset List page.

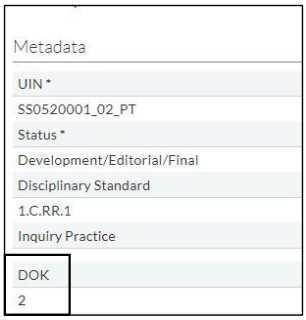
29

ABBI: KAS Standards and Practices

- Every item (except cluster set directions) has an aligned disciplinary standard shown at the top 1/3 of the screen.
- Approximately 50% of the items also have an aligned inquiry practice shown below the disciplinary standard.
- Cluster sources will not have either.
- You will leave a comment when you believe that an item is not aligned.

30

ABBI: DOK Alignment



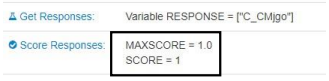
- Every item (except cluster set directions) has an assigned DOK indicated in the Metadata on the left side of the ABBI screen.
- Cluster sources will not have DOK information.
- You will leave a comment when you believe that an item is not correctly aligned to a DOK.

31

ABBI: Scoring

Use the green TN8 Preview button in the upper right to view the item in an environment similar to what students will see. If needed, enable pop -ups in order to view.

Solve MC, MS, and TE items. To verify that the scoring is accurate, select Score Responses at the top of the page. If the correct answer has been selected, then the SCORE value will equal the MAXSCORE value.

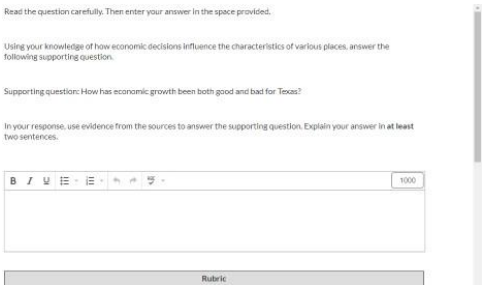


After review is complete, exit the TN8 Preview to return to ABBI.

32

ABBI: Rubrics

SA and ER items have rubrics that include an exemplar and answer cues. Rubrics are located within ABBI rather than the TN8 Preview. To view, use the scroll bar on the right side to scroll down to the rubric.



33

ABBI: Recording Votes

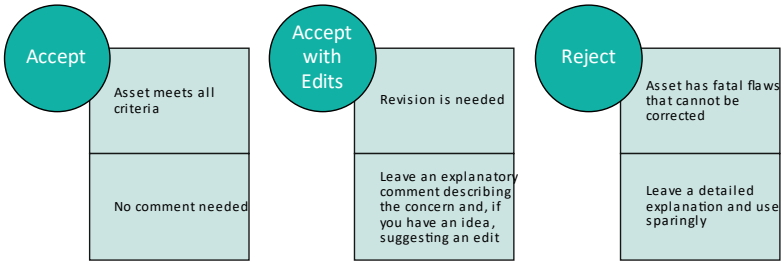
Each asset has a field that allows committee members to record a vote. This field is in the upper left of the ABBI page.

Please record a vote on every asset. When finished, select Save.

The criteria for voting is on the next slide.

34

ABBI: Voting Options



Do not use Accept with Reconciliation.

35

ABBI: Standalone Item Example

This source is from an interview with someone who grew up in Illinois in the late 1800s. Select **one** shaded sentence that **best** shows how cities changed during this time period.

We lived on a farm, and even telephones were curiosities to myself and the country boys of my age. . . .

Reminds me of a trip to the "city" once when I was about a dozen years old. My father and a neighbor . . . had to go up to the Big Town, which was Chicago, on some sort of business. . . . I suppose I'd been extra [earnest] at doing chores, weeding potatoes, killing worms on the tomato plants, or something . . . and Father rewarded me by taking me along. . . .

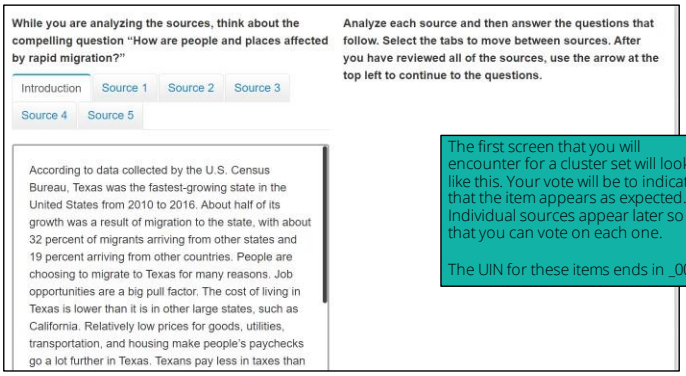
You can imagine what a time I had seeing things I'd never seen before, in fact had only dreamed about or heard about. . . .

But when I saw my first trolley car slipping along Cottage Grove Avenue in Chicago . . . slipping along without horses or engine or apparent motive power . . . well it was just too . . . much for me. I didn't know what to think.

—Interview with Harry Reece, Library of Congress, www.loc.gov (accessed March 2, 2020)

36

ABBI: First Cluster Screen in TN8 Preview

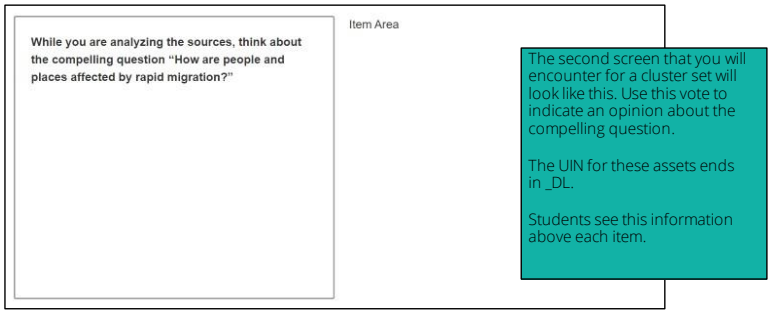


The first screen that you will encounter for a cluster set will look like this. Your vote will be to indicate that the item appears as expected. Individual sources appear later so that you can vote on each one.

The UIN for these items ends in _00.

37

ABBI: Second Cluster Screen in TN8 Preview



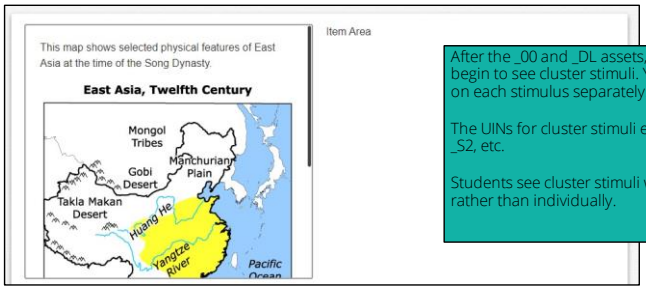
The second screen that you will encounter for a cluster set will look like this. Use this vote to indicate an opinion about the compelling question.

The UIN for these assets ends in _DL.

Students see this information above each item.

38

ABBI: Cluster Stimulus Example

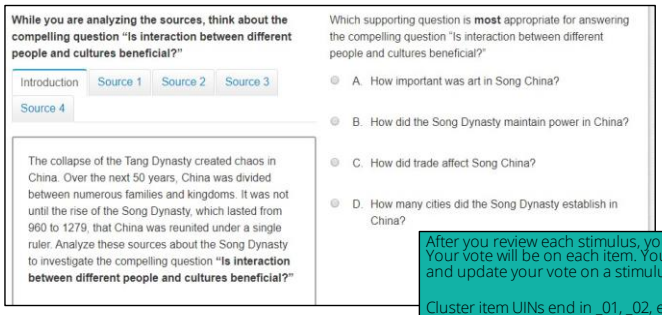


After the _00 and _DL assets, you will begin to see cluster stimuli. You will vote on each stimulus separately.

The UINs for cluster stimuli end in _IN, _S1, _S2, etc.

Students see cluster stimuli with items rather than individually.

ABBI: Cluster Item Example



After you review each stimulus, you will see items. Your vote will be on each item. You may go back and update your vote on a stimulus if needed.

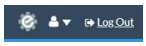
Cluster item UINs end in _01, _02, etc.

You are reviewing all items developed for a cluster set. Students will only see selected items.

40

ABBI: Logging Out

Use the Log Out option in the upper right of the screen to log out of ABBI.



Please log out rather than simply closing the browser window.

41

Wrap Up

42

Reminders

- Only discuss item content using a secure method.
- Important dates
  - Optional office hour: Thursday, September 9, 4:30–5:30 p.m. EDT / 3:30–4:30 p.m. CDT
  - Optional office hour: Monday, September 13, 4:30–5:30 p.m. EDT / 3:30–4:30 p.m. CDT
  - Optional office hour: Friday, September 17, 4:30–5:30 p.m. EDT / 3:30–4:30 p.m. CDT
  - Review complete: Monday, September 20, 9 a.m. EDT / 8 a.m. CDT
- Watch for an email after 9/20 that includes a link to information needed for payment.

43

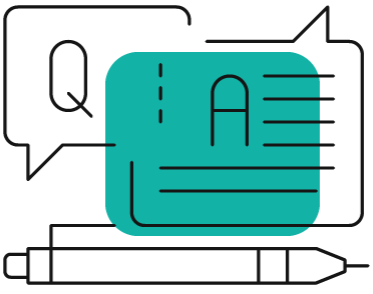
Questions?

Sharon Staples

sharon.staples@pearson.com

319-229-5212 (office)

706-421-6681 (cell)



44


Pearson
45

Appendix I. Item Content Review Checklist

Check to ensure that the content of each item:

- ☐ Is targeted to assess only one standard (unless specifications indicate otherwise).
- ☐ Deals with material that is important in testing the targeted standard.
- ☐ Uses grade-appropriate content.
- ☐ Is presented at a reading level suitable for the grade level being tested.
- ☐ Uses appropriate thinking skills (application, analysis, conclusions, extending).
- ☐ Has a stem that facilitates answering the question or completing the statement without looking at the answer choices.
- ☐ Has a stem that does not present clues to the correct answer choice.
- ☐ Has answer choices that are plausible and attractive to the student who has not mastered the objective or skill.
- ☐ Has mutually exclusive distractors.
- ☐ Has one and only one correct answer choice.
- ☐ Is conceptually, grammatically, and syntactically consistent between the stem and answer choices, and among the answer choices.
- ☐ Functions and scores correctly.

Appendix J. Mathematics and ELA Item Bias Review Training

Pearson

Kentucky Summative Assessments

Bias and Sensitivity Review

August 2020

Kentucky Department of Education — Our Children, Our Commonwealth

Photograph in the Carol M. Highsmith Archive, Library of Congress, Prints and Photographs Division

Pearson

12

Agenda

Purpose and Expectations

1. **Purpose**
   - an opportunity to advise Pearson and KDE on items by reviewing for bias and sensitivity concerns
2. **Expectations**
   - Reviews occur in a secure environment where others cannot see your computer screen.
   - Reviewers select an environment that is free of distractions.
   - Reviewers avoid multi -tasking.
   - Reviews occur outside of regular school hours.
   - Reviewers use secure methods of communication when contacting Pearson or KDE.
   - Reviewers refrain from discussing the content of specific assets except during designated times within the scheduled review window.

Pearson 17

Rationale for Review

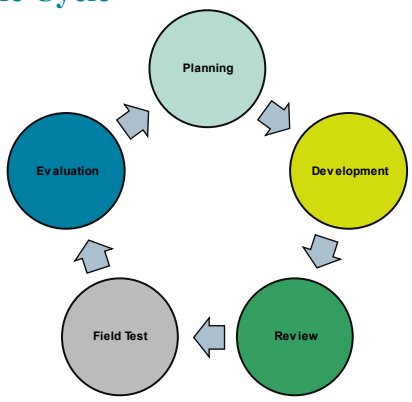
Fairness and sensitivity cannot be properly addressed as an afterthought. It must be confronted throughout the interconnected phases of the testing process, from test design and development to administration, scoring, interpretation and use.

—National Research Council, 1999

Pearson 18

Development Overview

Item Life Cycle



Pearson

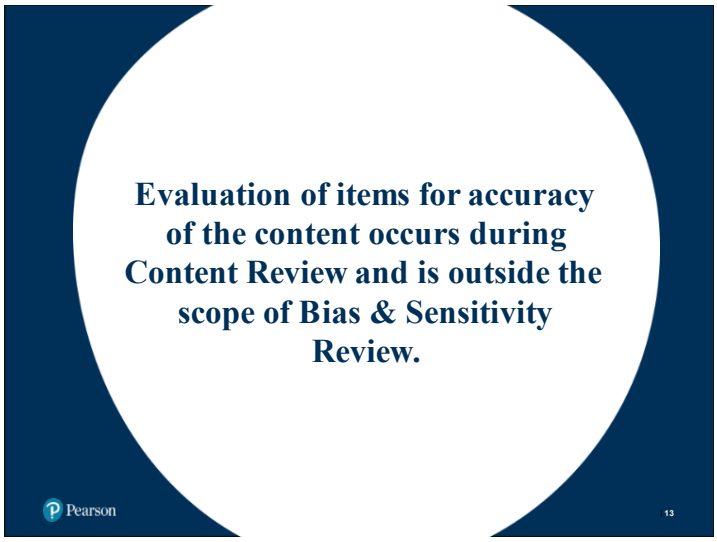
10

Test Components: Description

1. **Standalone Items**
   - Definition: Items that are self -contained rather than part of a set
   - Math reviewers will see mostly standalone items
2. **Cluster Sets**
   - Definition: Sets of independent items with shared stimuli
   - Reviewers in reading, editing and mechanics, and math will see sets

Pearson

11



Bias: Construct Relevance & Language Appropriateness

1. Is the content, context, and vocabulary grade appropriate?
2. Does the asset use low frequency and/or ambiguous vocabulary?
3. Is the asset free of idioms that would disadvantage English Learners?
4. Does the item contain language and/or a dialect that is not commonly used across the state or that has different connotations in different parts of the state? (e.g., "shopping cart" versus "buggy")
5. Does the asset rely on prior knowledge of extraneous information, such as the fact that the player with the lowest score wins in golf?

Pearson 14

Bias: Groups

1. Does the asset discriminate against or give an advantage to students of certain ethnic, racial, religious, or political backgrounds?
2. Does the asset discriminate against English Learners or students with special needs?
3. Does the asset favor one gender over another?
4. Are graphics used in an asset adaptable to Braille and large print?
5. Does the asset respectfully portray represented groups rather than perpetuating stereotypes?
6. Does the asset have a context associated with certain socioeconomic groups?
   – luxury automobiles
   – ski lodges

Pearson 15

Bias: Geography

1. Does the setting for the asset provide an unfair advantage for students of a certain region?
2. Does the setting for the asset provide an unfair advantage for students in an urban or rural area? (e.g., grain elevators)

Pearson 16

Sensitivity

1. Is the asset likely to elicit undue emotion in students?
2. Assets should
   – Appropriately portray life's tragedies
   – Avoid controversial topics, unless required by the standard
   – Avoid recent catastrophic occurrences

Pearson 17

ABBI: Logging In

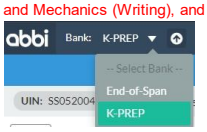
- Using Chrome as your browser, access the ABBI site: https://abbi.pearson.com/.
- At the log in screen, select "Password Assistance."
- Enter your email address. This will email you a link to set a password of your choosing.
- Log into ABBI at https:abbi.pearson.com/ using your email and newly created password.

Pearson

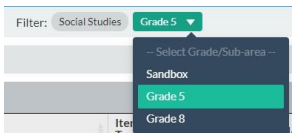
19

ABBI: Asset List (K-PREP bank)

After logging in, use the drop -down menu in the upper left of the ABBI screen to select the K-PREP bank.
The K-PREP bank houses all items in Grades 3-8 for Reading, Editing and Mechanics (Writing), and Math.



Then use the drop -down menus in the top middle of the page to select the appropriate subject/grade (based on specific review assignment).

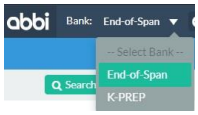


Pearson

20

ABBI: Asset List (End-of-Span)

The End-of-Span bank houses all items in Grades 10 Reading and Math, and Grade 11 Editing and Mechanics (Writing).
To access these grades, select the End-of-Span bank.



Then select your subject/grades from the drop -downs in the middle of your screen.

Pearson

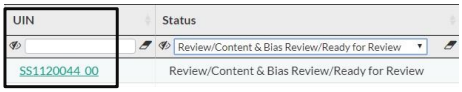
21

ABBI: Sorting

Rev Seq: Type in "BR" in box and press ENTER.
Select the "up" arrow to sort into the correct review order.
Reading and Editing & Mechanics screenshot:

Math screenshot:

Select the first UIN to navigate to the review screen.



Pearson

22

ABBI: Navigating

Each item has navigation tools in the top right -hand corner. The arrows can be used to move to the next item or to go back to the previous.

The list icon returns you to Asset List page.

Pearson

23

ABBI: Viewing the Item

Use the green TN8 Preview button in the upper right to view the item in an environment similar to what students will see. If needed, enable pop-ups in order to view.

TN8 Preview

After review is complete, exit the TN8 Preview to return to ABBI.

Pearson

24

ABBI: Recording Votes

Each asset has a field that allows committee members to record a vote. This field is in the upper left of the ABBI page.

Reviewers should vote on each asset, adding comments as needed. When finished, select Save.

Pearson

25

ABBI: Voting Options

1 Accept

- **Basis for vote:** The asset meets all review criteria and no changes are needed.
- **Comment:** No comment is needed.

2 Accept with Edits

- **Basis for vote:** The reviewer believes that a revision can fix the bias or sensitivity concern.
- **Comment:** The reviewer describes the bias and sensitivity concern and, if possible, provides specific suggested language for the revision.

3 Accept with Reconciliation

NOT USED for Bias and Sensitivity Review

4 Reject

- **Basis for vote:** The reviewer believes that the asset has fatal flaws that cannot be corrected.
- **Comment:** The reviewer provides a detailed comment explaining the rationale for rejecting the item.
- **NOTE:** This option should rarely be used. In order to provide the largest possible bank of items for Kentucky, the preference is always to revise flawed items.

Pearson

26

ABBI: Standalone Item Example

The speed of light is about 186,282 miles per second. Which number name can be used to represent 186,282?

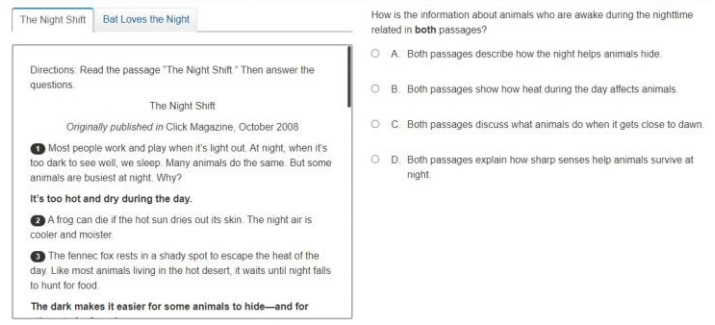
- A. one hundred eighty-six thousand, two hundred eighty-two
- B. one hundred thousand, eighty-six hundred eighty-two
- C. eighteen thousand, sixty-two hundred eighty-two
- D. eighteen hundred, sixty-two thousand eighty-two

Pearson

27

ABBI: Cluster Item Example –

Cluster items show the aligned stimuli on the left and the item on the right. Note that their may be multiple tabs on the left showing paired passages for answering the set of questions.

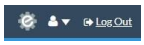


Pearson

28

ABBI: Logging Out

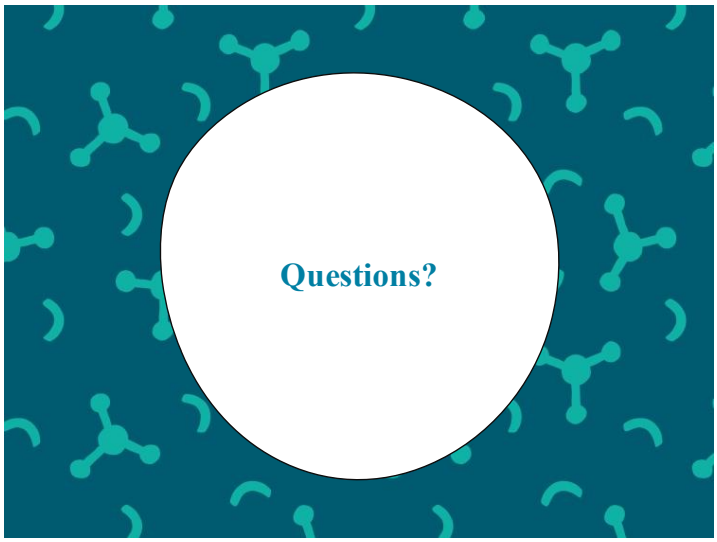
Use the log out option in the upper right of the screen to log out of ABBI.



Please log out rather than simply closing the browser window.

Pearson

29



Pearson Contact Information

- Please include all three contacts on all emails.
- Avoid mentioning secure information in emails.

Adrian Rivera	Chip Lowe	Jennifer Ramirez
Adrian.rivera@pearson.com	Chip.lowe@pearson.com	Jennifer.ramirez1@pearson.com

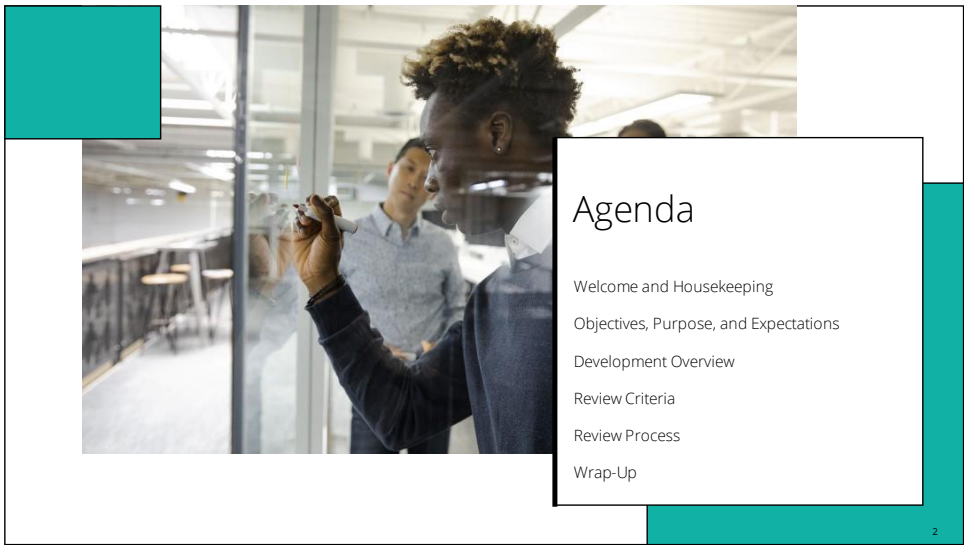
Pearson

131

Thank you for participating today and for your contributions to the Kentucky Summative Assessments.

ALWAYS LEARNING

Appendix K. Social Studies Item Bias Review Training



Welcome and Housekeeping

3

Welcome to Participants

Kentucky Department of Education

- Heather Ransom, Academic Program Consultant
- Lauren Gallicchio, Academic Program Consultant

Pearson

- Adrian Rivera, Senior Test Development Manager
- Sharon Staples, Principal Assessment Specialist

Kentucky Educators

4

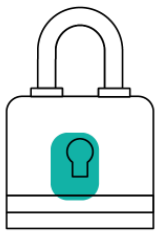
Housekeeping

Confidentiality and Nondisclosure Agreement

- Participants must maintain the security of the assets, documents, and materials being reviewed.
- Participants may not copy, discuss, or disclose in any manner specific information or materials used during this meeting, while reviewing assets, or after the review committee has concluded.
- All materials for the assessment program are the property of the State of Kentucky.

Honorarium

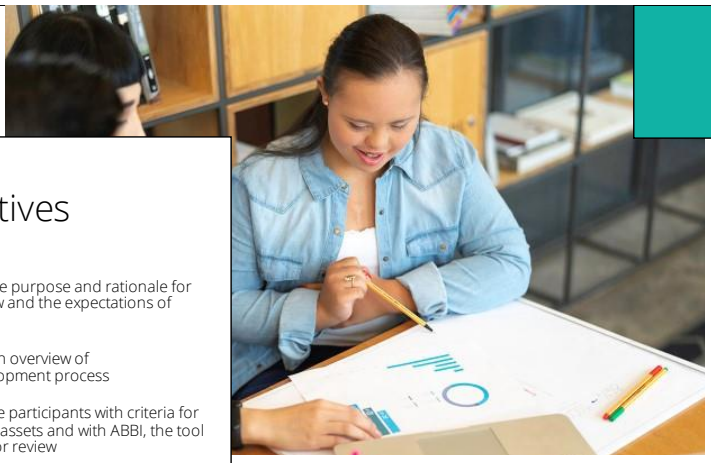
- Pearson provides a link after the review ends
- Processing can occur only after submission of requested information



5

Objectives, Purpose, and Expectations

6



Objectives

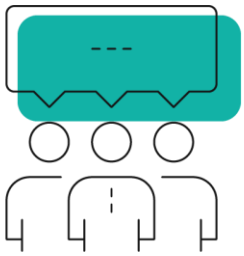
1. Explain the purpose and rationale for this review and the expectations of reviewers
2. Provide an overview of the development process
3. Familiarize participants with criteria for reviewing assets and with ABBI, the tool needed for review

7

Purpose and Rationale

Purpose

An opportunity to advise Pearson and KDE on bias and sensitivity concerns in assets



Rationale

"Fairness and sensitivity cannot be properly addressed as an afterthought. It must be confronted throughout the interconnected phases of the testing process, from test design and development to administration, scoring, interpretation and use."

—National Research Council, 1999

8

Expectations

- Choose a secure environment where others cannot see your computer screen
- Choose an environment that is free of distractions
- Avoid multi-tasking
- Complete reviews outside of regular school hours
- Use a secure method of communication when contacting Pearson or KDE
- Refrain from discussing assets except during designated times within the scheduled review window

9

Development Overview

10

Kentucky Statute

Per 158.6453, beginning in fiscal year 2017 –2018, and every six (6) years thereafter, the Kentucky Department of Education (KDE) shall implement a process for reviewing Kentucky's academic standards and the alignment of corresponding assessments.

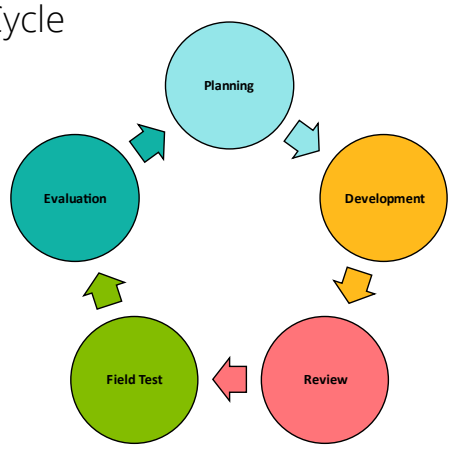
How does this statute affect this review work?

All standards are eligible for assessment- item development.

The assessments administered at Grades 5, 8, and 11 are grade-span tests:

Grade 5 test items assess Kindergarten through Grade 5 standards
Grade 8 test items assess Grade 6 through Grade 8 standards
Grade 11 test items assess civics, economics, geography, U.S. history, and world history standards.

Item Life Cycle



12

Review Criteria

14

Bias: Construct Relevance and Language Appropriateness

- Are the content, context, and vocabulary grade appropriate?
  - **NOTE: Evaluation for accuracy occurs during Content Review and is outside the scope of Bias & Sensitivity Review.**
- Is low frequency or ambiguous vocabulary used?
- Are idioms that would disadvantage English Learners used?
- Is regional language that is not common throughout the state used?
- Does the asset rely on prior knowledge of extraneous, non -social studies content?

15

Bias: Groups

- Does the asset discriminate against or give an advantage to students of certain ethnic, racial, religious, or political backgrounds?
- Does the asset discriminate against English Learners or students with special needs?
- Does the asset favor one gender over another?
- Are graphics adaptable to Braille and large print?
- Does the asset respectfully portray represented groups rather than perpetuating stereotypes?
- Does the asset have a context associated with certain socioeconomic groups?
- Does the setting of the asset unfairly advantage students of a certain region?

16

Sensitivity

- Is the asset likely to elicit undue emotion in students?
- Does the asset appropriately portray life's tragedies?
- Does the asset elicit association with recent catastrophic occurrences?
- Does the asset avoid controversial topics?

NOTE: Social studies assets may address controversial or emotional topics. This committee should flag these assets only if their treatment of such material disadvantages any student population.

17

Review Process

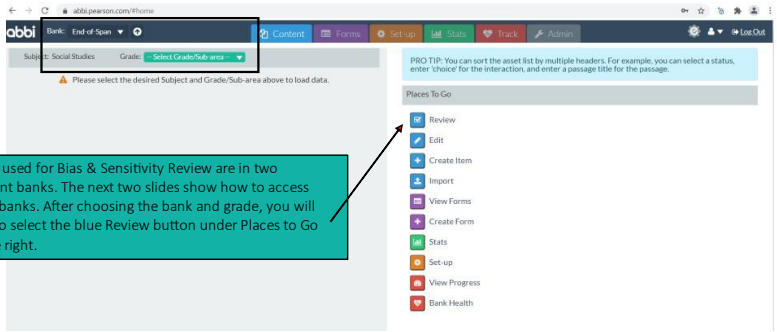
18

ABBI: Logging In

- Using Chrome or Firefox as your browser, access the ABBI site: https://abbi.pearson.com/
- Enter your username and password.

19

ABBI: Initial Log-In Page

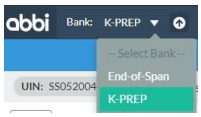


Assets used for Bias & Sensitivity Review are in two different banks. The next two slides show how to access those banks. After choosing the bank and grade, you will have to select the blue Review button under Places to Go on the right.

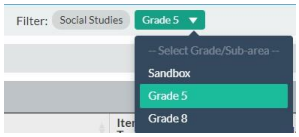
20

ABBI: Asset List (Grades 5 and 8)

After logging in, use the drop -down menu in the upper left of the ABBI screen to select the K -PREP bank.



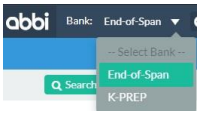
Then use the drop -down menu in the top middle of the page to select the grade (5 or 8).



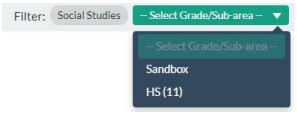
21

ABBI: Asset List (Grade 11)

To access grade 11, select the End -of-Span bank.



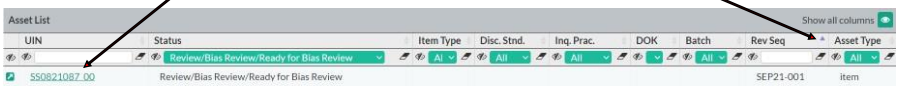
Then select HS (11).



22

ABBI: Sorting

1. Status: Review>Bias Review>Ready for Bias Review
2. Rev Seq: Select the "up" triangle to sort into the correct review order.
3. Select the first UIN to navigate to the review screen.



23

ABBI: Navigating

Each asset has navigation tools in the top right -hand corner. The arrows can be used to move to the next asset or to go back to the previous asset.

The list icon returns you to the Asset List page.

24

ABBI: Viewing the Asset

Use the green TN8 Preview button in the upper right to view the asset in an environment similar to what students will see. If needed, enable pop -ups in order to view.

TN8 Preview

After review is complete, exit the TN8 Preview to return to ABBI.

25

ABBI: Recording Votes

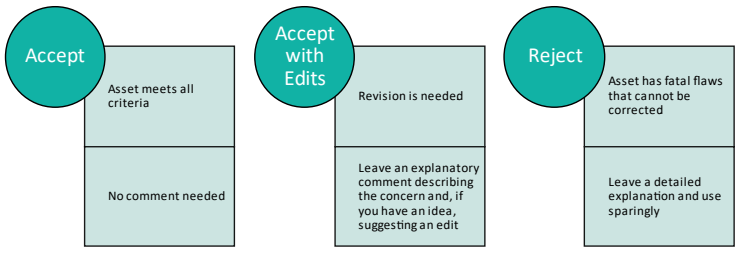
Each asset has a field that allows committee members to record a vote. This field is in the upper left of the ABBI page.

Please record a vote on every asset. When finished, select Save.

The criteria for voting is on the next slide.

26

ABBI: Voting Options



Do not use Accept with Reconciliation.

27

ABBI: Standalone Item Example

This source is from an interview with someone who grew up in Illinois in the late 1800s. Select **one** shaded sentence that **best** shows how cities changed during this time period.

We lived on a farm, and even telephones were curiosities to myself and the country boys of my age. . . .

Reminds me of a trip to the "city" once when I was about a dozen years old. My father and a neighbor . . . had to go up to the Big Town, which was Chicago, on some sort of business. . . . I suppose I'd been extra [earnest] at doing chores, weeding potatoes, killing worms on the tomato plants, or something . . . and Father rewarded me by taking me along. . . .

You can imagine what a time I had seeing things I'd never seen before, in fact had only dreamed about or heard about. . . .

But when I saw my first trolley car slipping along Cottage Grove Avenue in Chicago . . . slipping along without horses or engine or apparent motive power . . . well it was just too . . . much for me. I didn't know what to think.

—Interview with Harry Reece, Library of Congress, www.loc.gov (accessed March 2, 2020)

28

ABBI: First Cluster Screen in TN8 Preview

While you are analyzing the sources, think about the compelling question "How are people and places affected by rapid migration?"

Analyze each source and then answer the questions that follow. Select the tabs to move between sources. After you have reviewed all of the sources, use the arrow at the top left to continue to the questions.

Introduction | Source 1 | Source 2 | Source 3 | Source 4 | Source 5

According to data collected by the U.S. Census Bureau, Texas was the fastest-growing state in the United States from 2010 to 2016. About half of its growth was a result of migration to the state, with about 32 percent of migrants arriving from other states and 19 percent arriving from other countries. People are choosing to migrate to Texas for many reasons. Job opportunities are a big pull factor. The cost of living in Texas is lower than it is in other large states, such as California. Relatively low prices for goods, utilities, transportation, and housing make people's paychecks go a lot further in Texas. Texans pay less in taxes than

The first screen that you will encounter for a cluster set will look like this. Your vote will be to indicate that the item appears as expected. Individual sources appear later so that you can vote on each one.

The UIN for these assets ends in _00.

29

ABBI: Second Cluster Screen in TN8 Preview

While you are analyzing the sources, think about the compelling question "How are people and places affected by rapid migration?"

Item Area

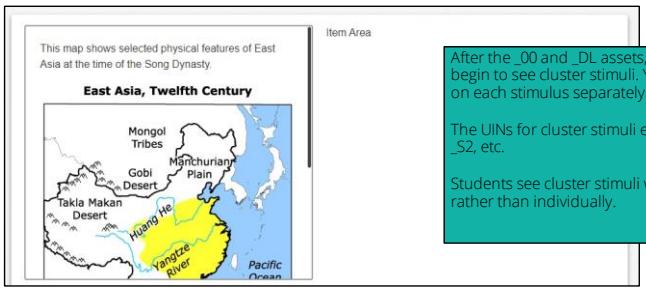
The second screen that you will encounter for a cluster set will look like this. Use this vote to indicate an opinion about the compelling question.

The UIN for these assets ends in _DL.

Students see this information above each set of stimuli.

30

ABBI: Cluster Stimulus Example

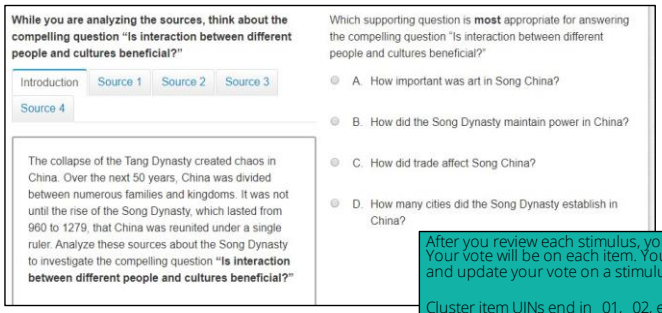


After the _00 and _DL assets, you will begin to see cluster stimuli. You will vote on each stimulus separately.

The UINs for cluster stimuli end in _IN, _S1, _S2, etc.

Students see cluster stimuli with items rather than individually.

ABBI: Cluster Item Example



After you review each stimulus, you will see items. Your vote will be on each item. You may go back and update your vote on a stimulus if needed.

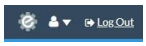
Cluster item UINs end in _01, _02, etc.

You are reviewing all items developed for a cluster set. Students will only see selected items.

32

ABBI: Logging Out

Use the Log Out option in the upper right of the screen to log out of ABBI.



Please log out rather than simply closing the browser window.

33

Wrap Up

34

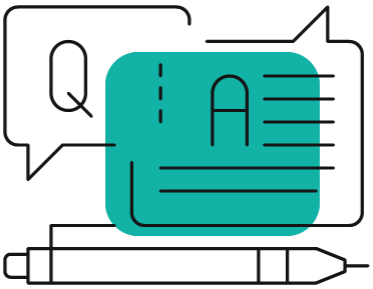
Questions?

Sharon Staples

sharon.staples@pearson.com

319-229-5212 (office)

706-421-6681 (cell)



35

Pearson

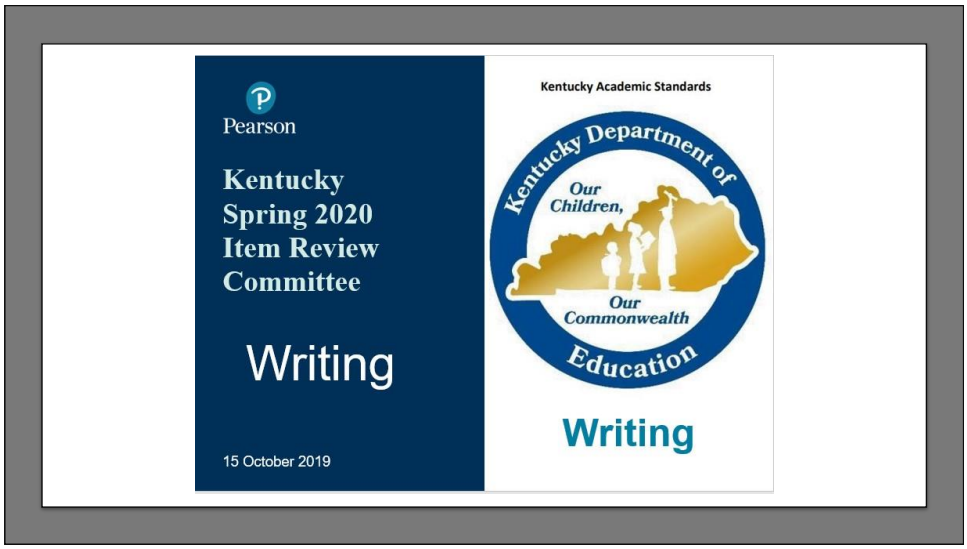
36

Appendix L. Item and Passage Bias Review Checklist

Look for items and passages that

- reflect favoritism toward a gender or ethnic group;
- are potentially offensive, inappropriate, or negative toward any group;
- discriminate in any way against individuals with disabilities;
- have reference to religion that shows favoritism or promotion;
- contain any controversial or emotionally charged subject matter;
- have underlying assumptions not shared across ethnic, racial, and gender groups, socioeconomic levels, and geographic areas;
- contain language and/or dialect that is not commonly used across the state or has different connotations in various parts of the state;
- have an inappropriate tone;
- use low frequency and/or ambiguous vocabulary; and
- are disadvantageous to English learners.

Appendix M. On-Demand Writing Item Content Review Training



Agenda

- "Housekeeping"
- Welcome and Introduction

I. Assessment Overview

Components of the Writing Assessment
Evidence -Centered Test Design
Standards
Item Types

II. Item Review Committee Meetings

Reviewer Role
Review Process, Materials
Item Review Guiding Questions and Criteria

III. ABBI Training

"Housekeeping"

Non-Disclosure/Security

- Process vs. Specifics
- Materials
- Cell Phones

Schedule

Grade	Tuesday	Wednesday
5, 8, 11	8:30 am - 5:00 pm	8:30 am - 5:00 pm

Breaks and lunch will be determined in the room

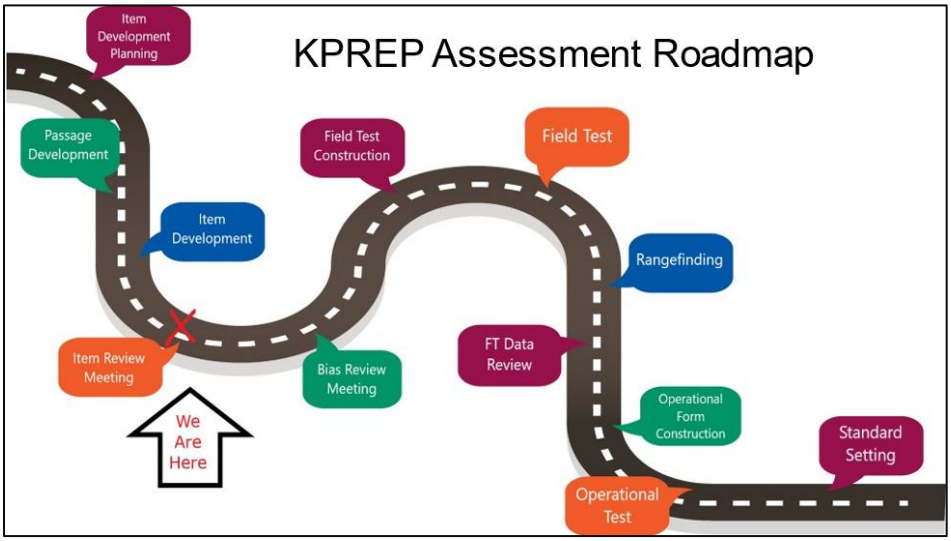
Welcome and Introductions

Reviewer Role

The role of each reviewer is to offer your professional perspective on all items in your assigned item group. Most of the work will be self-paced and individual, but there will also be opportunities for discussion as well.

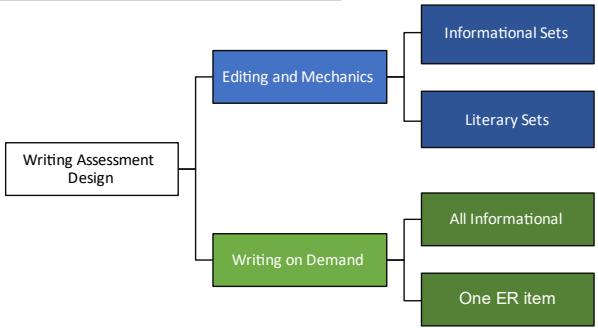
- Be focused
- Provide detailed feedback for each item as needed
- Ask clarifying questions as needed
- Participate in discussions
- Respect the opinions of all involved

Assessment Overview



Assessment Overview

Components of the Writing Assessment



Evidence-Centered Design (ECD)

ECD is a deliberate and systematic approach to assessment development that **establishes the validity** of the assessments, **increases the comparability** of year-to-year results, and **increases efficiencies/reduces costs**.

Claims

Design begins with the inferences (**claims**) we want to make about students.

Evidence

In order to support **claims**, we must gather **evidence.**

Item Sets

Item Sets are designed to provide options across genre, item types and standards.

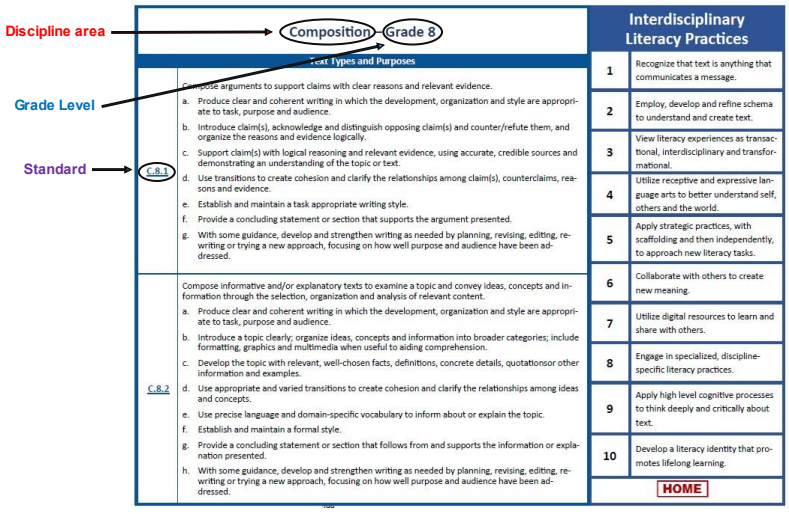
Operational Form

Operational forms are built from field tested items across sets to provide specific evidence from students in support of claims.

Standards

KAS: What are the Writing Standards?

- Describe what a student needs to be able to do to show mastery
- Support Language and Composition claims
- Targeted to both literary and informational passages
- Provide for a range of teaching and assessment options



Kentucky Item Types

- **Multiple Choice Items (MC)**
- **Multiple Select Items (MS)**
- **Short Answer Items (SA)**
- **Extended Response Items (ER)**

Item Types

Multiple Choice (MC) Items

Directions: Read the passage and answer the following questions.

NASA Unveils Sustainable Campaign to Return to Moon, on to Mars

In December of 2017, President Donald Trump signed Space Policy Directive-1, in which the president will direct (1) NASA "to lead an innovative and sustainable program of exploration with commercial and international partners to enable human expansion across the solar system and to bring back to Earth new knowledge and opportunities."

In answer to that bold call, and consistent with the NASA Transition Authorization Act of 2017, NASA recently submitted to

What is the best revision for underlined phrase 1?

- A. NO CHANGE
- B. directed
- C. was going to direct
- D. would have directed

Item Types

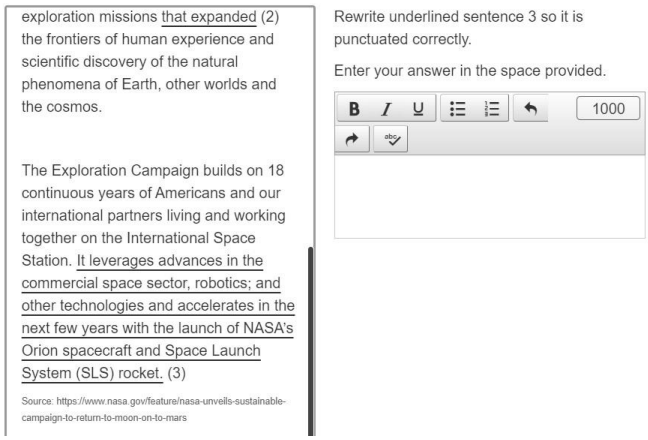
Multiple Select (MS) Items

In answer to that bold call, and consistent with the NASA Transition Authorization Act of 2017, NASA recently submitted to Congress a plan to revitalize and add direction to NASA's enduring purpose. The National Space Exploration Campaign calls for human and robotic exploration missions that expanded (2) the frontiers of human experience and scientific discovery of the natural phenomena of Earth, other worlds and the cosmos.

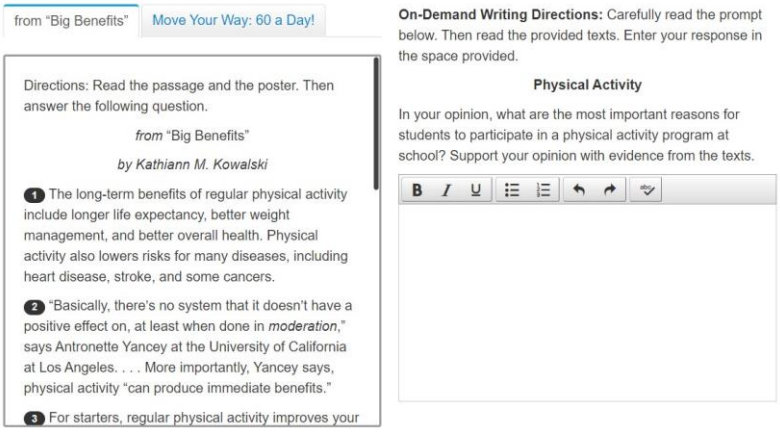
Which are the **best** choices for underlined phrase 2? Select **two** correct answers.

- A. NO CHANGE
- B. to expand
- C. which expanded
- D. that will expand
- E. which were expanding

Item Types Short Answer (SA) Items



Item Types Extended Response (ER) Items



Emphasis on Item Simplification

- This program is currently in the process of creating an item bank at all grade levels
- We are focused on increasing the number of both accessible and complex items, targeting performance levels 2 and 3
- Item simplification includes:
  - straightforward language in stems and answer choices
  - concise ER and SA prompts; reducing wordiness
  - MS items limited to five options with two keys

Content Review: Role of the Reviewer

The role of the Content Reviewer is to provide expert content review of items within assigned passage sets.

- Review item sets assigned to you using Item Review Criteria
- Assign Item Status
  - Accept— Recommend the item be approved as it is
  - Accept with Edits— Recommend the item be approved with edits suggested for improvement:
    - Could be a content edit, edit to standards, edit to functionality, etc.
  - Reject— Recommend the item NOT be approved; fatal flaws prevent any ability to revise

Content Review: Role of the Reviewer

Please note what is NOT the role of the Content Review committee

- Bias/Sensitivity Item Review committees will review all items next week using bias/sensitivity guidelines; that is not the responsibility of this committee
- Texts cannot be rejected/revised at this stage
  - Reviewers may note egregious errors/typos within passages
  - Reviewers may note concerns with passage content, but review focus must be on items themselves

Item Review: Materials

The following documents will be available to reviewers:

- ELA Item Reviewer Training PowerPoint
- Guiding Questions/Item Review Criteria
- Kentucky Standards Document
- SA and ER Scoring Rubrics

Item Review: Process

Committee Item Review Process

1. Determine Item Review Assigned Group (A -B)
2. Navigate in ABBI to grade level and filter by item sequence (A -B)
3. Sort by item sequence
4. Begin with first item in the group
5. Read passage, then review items using review checklist
6. Vote on each item in ABBI
7. Enter comments (if any) to identify issues and/or offer recommendations for resolution
8. Facilitator will review votes and comments in live time and discuss trends with the group as needed

Item Review Criteria/Guiding Questions

1. Standard Alignment:
   - Does the item allow for students to demonstrate mastery of the aligned standard(s)?
2. Content Appropriateness:
   - Is the content of the item clear, concise, and appropriate for the intended grade level?
3. Key and answer options:
   - Is the keyed answer the only correct option?
   - Are distractors plausible and mutually exclusive?
4. Item construction and functionality:
   - Is the item constructed with appropriate grammar and syntax across all elements?
   - Does the item function and score correctly?

Criterion 1: Alignment to the Standards

Items should:

- Align to a significant part orall of a standard
- Reflect the language of the standard as appropriate
- Assess only one standard
- Note: It may require multiple items to assess the full standard

Criterion 2: Content Appropriateness

Items should:

- Reflect the reading level for the tested grade
- Require appropriately complex thinking and problem solving
- Assess topics and concepts that adhere to grade-level learning

24

Criterion 3: Key and Answer Options

Items should avoid **internal** clueing or miscues:

- answer options should NOT repeat or echo a word used in the stem

Items should avoid **external** clueing or miscues:

- items should not be answerable using other items in the set
- other items in the set should not mislead students toward selecting the wrong answer option for any given item

Criteria 4: Item construction and functionality

All items:

- Are conceptually, grammatically, and syntactically consistent between the stem and answer choices, and among answer choices
- Function and score correctly in ABBI

Next Steps

- Item Review Group Assignments
- ABBI Training
- Begin Review

Appendix N. On-Demand Writing Content Review Checklist

1. Is the topic or subject matter grade appropriate?
2. Does the writing situation for a stand alone prompt provide the necessary background the student needs to complete the writing task?
3. Do the writing directions identify the purpose of the writing task, the format and type of response, and the audience to or for whom it is being written?
4. With the passage-based prompts, is the passage or the paired passage set complete enough for the writing task required?
5. Does the prompt guide the student to an appropriate and original response?
6. Is the prompt accessible to all students?
7. Does the prompt deter any possible inappropriate paths for student response that might cause an alert when scored?
8. Is the prompt high-interest and does it motivate students to want to write?
9. Is the prompt free of bias or sensitivity issues?
10. Is the passage or situation written in a clear and direct manner?

Appendix O. On-Demand Writing Bias Review Checklist

Look for passages/prompts that:

- reflect favoritism toward a gender or ethnic group
- are potentially offensive, inappropriate, or negative toward any group
- discriminate in any way against individuals with disabilities
- have reference to religion that shows favoritism or promotion
- contain any controversial or emotionally charged subject matter
- have underlying assumptions not shared across ethnic, racial, and gender groups, socioeconomic levels, and geographic areas
- contain language and/or dialect that is not commonly used across the state or has different connotations in various parts of the state
- have an inappropriate tone

Appendix P. On-Demand Writing Scoring Rubrics

KAS Opinion Rubric--5th Grade On-Demand Writing

Guiding Principle C1: Students will compose arguments to support claims in an analysis of substantive topics or texts, using valid reasoning and relevant and sufficient evidence.

Note: *In 5th grade, students compose opinion pieces, using writing and digital resources, on topics or texts, supporting the writer's perspective with reasons and information. (C.5.1) The shift to composing arguments begins in 6th grade.*

Scoring Element	Score Point 1	Score Point 2	Score Point 3	Score Point 4
Clarity and Coherence	States an opinion that may **lack** focus or be **unclear**. **Misses many or all** demands of the prompt.	States a **general** opinion that addresses the prompt, but may have **lapses** in focus. **Attempts** to address **some** demands of the prompt.	Introduces and maintains a **clear and coherent** opinion. Addresses **all** demands of the prompt.	Introduces and maintains a **clear, credible and coherent** opinion. **Thoroughly** addresses all demands of the prompt.
Support	Includes **minimal or no purposeful** support of opinion with reasons. Provides **incomplete, inaccurate and/or irrelevant** explanation of reasons. Provides **minimal or unrelated** facts and details to support the reasons.	**Attempts** to support opinion with reasons. Provides **vague and/or general** explanation of reasons. Provides **vague and/or general** facts and details to support the reasons.	Supports opinion with **logical** reasons. Provides **clear** explanation of reasons. Provides facts and details that **clearly** support the reasons.	**Thoroughly** supports opinion with logical reasons. Provides **carefully selected** explanation of reasons to **strengthen the opinion.** Provides reasons that are **thoughtfully** linked to facts and details to support the opinion.
Sourcing	Uses **one or none** of the provided sources or **ineffectively** uses a minimum of two provided sources to support the opinion. Cites **little or no** evidence. **Little or no** use of quoting, summarizing and/or paraphrasing of facts and details.	Uses a minimum of two provided sources to **attempt** to support the opinion. **Inconsistently** cites evidence. **Attempts** to quote, summarize and/or paraphrase facts and details.	**Accurately and effectively** uses a minimum of two provided sources to support the opinion. **Effectively** cites evidence by quoting, summarizing and/or paraphrasing facts and details.	Accurately and **skillfully** uses a minimum of two provided sources to support the opinion. **Consistently and thoroughly** cites evidence by quoting, summarizing and/or paraphrasing facts and details.
Organization	Creates **minimal or no overall** structure. **Ineffectively** organizes an opinion with reasons that are supported by facts and details. Makes **minimal or no attempt** to use transitions to connect the opinion, reasons and evidence. Provides a **weak** conclusion section or **lacks** a conclusion section to support the opinion.	**Attempts** to create a structure for the opinion. Organizes introduction of the topic and states an opinion with reasons that are supported by facts and details, but contains **some lapses that disrupt the cohesion or are inappropriate.** **Attempts** to use transitions to connect the opinion, reasons and evidence, but they are **simple and infrequent**. Provides a conclusion section **in an attempt** to support the opinion.	Creates and **maintains a clear** structure to develop the opinion. **Logically** organizes introduction of the topic and states an opinion with reasons that are **logically** ordered and supported by facts and details. Uses **effective** transitions to connect the opinion, reasons and evidence. Provides a **logical** conclusion section to support the opinion.	Creates and maintains a **sophisticated** structure to develop the opinion. **Skillfully** organizes introduction of the topic and states an opinion with reasons that are logically ordered and supported by facts and details. **Consistently** uses a **variety** of transitions to create a **strong** connection between the opinion, reasons and evidence. Provides a **thorough** conclusion to support the opinion.
Language/ Conventions	**Lacks or uses an inappropriate** formal tone or voice. **Lacks** the development of task appropriate writing. Uses **simple or inappropriate** word choice.	Uses a **weak formal tone or voice and/or** has **lapses** in appropriate tone or voice. **Attempts** to develop task appropriate writing. **Attempts** appropriate word choice.	**Establishes and maintains** an appropriate formal tone or voice. **Establishes and maintains** task appropriate writing. **Effectively** uses appropriate word choice.	**Consistently** establishes and maintains a **sophisticated** formal tone or voice. **Consistently** establishes and maintains **sophisticated**, task appropriate writing. **Consistently** uses **effective** and **varied** word choice.

Scoring Element	Score Point 1	Score Point 2	Score Point 3	Score Point 4
	Makes **significant** errors in the conventions of Standard English grammar, usage, spelling, capitalization and punctuation which **interfere** with understanding the writing.	Makes **frequent** errors in the conventions of Standard English grammar, usage, spelling, capitalization and punctuation which **may interfere** with understanding the writing.	**Effectively** uses the conventions of Standard English grammar, usage, spelling, capitalization and punctuation with **minor** errors that **do not interfere** with understanding the writing.	**Skillfully** uses the conventions of Standard English grammar, usage, spelling, capitalization and punctuation with **few**, minor errors that do not interfere with understanding the writing.

KAS Argumentation Rubric–8th Grade On-Demand Writing

Guiding Principle C1: Students will compose arguments to support claims in an analysis of substantive topics or texts, using valid reasoning and relevant and sufficient evidence.

Scoring Elements	Score Point 1	Score Point 2	Score Point 3	Score Point 4
Clarity and Coherence	Makes claim(s) that may **lack** focus or be **unclear.** **Misses many or all** demands of the prompt.	Makes **general** claim(s) that address the prompt, but may have **lapses** in focus. **Attempts** to address **some** demands of the prompt.	Introduces and maintains **clear and coherent** claim(s). Addresses **all** demands of the prompt.	Introduces and maintains clear, **credible** and coherent claim(s). **Thoroughly** addresses all demands of the prompt.
Counterclaims	Makes an **ineffective attempt** or makes **no attempt** to acknowledge opposing claim(s). Makes an **ineffective attempt** or makes **no attempt** to counter and/or refute opposing claim(s).	**Attempts** to acknowledge opposing claim(s), but lacks insight, interpretation or clarification. **Attempts** to counter and/or refute opposing claim(s).	**Acknowledges and distinguishes** opposing claim(s) with insight, interpretation or clarification. **Counters and refutes** opposing claim(s).	**Skillfully** acknowledges and distinguishes opposing claim(s) with insight, interpretation or clarification. **Thoroughly** counters and refutes opposing claim(s) with **carefully selected evidence.**
Support	Includes **minimal or no purposeful** support of claim(s) with evidence. Provides **incomplete, inaccurate and/or irrelevant** explanations of evidence and ideas. Provides **minimal or unrelated** reasoning to support claim(s).	**Attempts** to support claim(s) with evidence. Provides **vague and/or general** explanations of evidence and ideas. Provides **vague and/or general** reasoning to support claim(s).	Supports claim(s) with **logical reasons and relevant evidence.** Provides **logical** explanations of evidence and ideas. Provides reasoning that **clearly links evidence** to support claim(s).	**Thoroughly** supports claim(s) with logical reasons and **carefully selected**, relevant evidence that **strengthens the argument.** Provides **thorough and effective** explanations of evidence and ideas. Provides varied reasoning which **thoughtfully** links evidence to support claim(s).
Sourcing	Uses **one or none** of the provided sources or **ineffectively** uses a minimum of two provided sources to support the claim(s) and/or opposing claim(s). Cites **little or no** evidence. **Little or no** use of quotes and/or paraphrasing of details, examples and ideas.	Uses a minimum of two provided sources to **attempt** to support the claim(s) and/or opposing claim(s). **Inconsistently** cites evidence. **Attempts** to quote and/or paraphrase details, examples and ideas.	**Accurately and effectively** uses a minimum of two provided sources to support the claim(s) and/or opposing claim(s). **Effectively** cites evidence by quoting and/or paraphrasing details, examples and ideas.	Accurately and **skillfully** uses a minimum of two provided sources to support the claim(s) and/or opposing claim(s). **Consistently and thoroughly** cites evidence by quoting and/or paraphrasing details, examples and ideas.
Organization	Builds **minimal or no overall** structure for the argument. **Ineffectively** organizes claim(s), counterclaims, evidence and reasoning, creating a lack of cohesion. Makes a **minimal attempt** or makes **no attempt** to use transitions to link claim(s), counterclaims, reasons and evidence. Provides a **weak conclusion or lacks a conclusion** to support the argument.	**Attempts** to build a structure for the argument. **Attempts** to organize claim(s), counterclaims, evidence and reasoning, but contains **some lapses that disrupt the cohesion or are inappropriate for the context**. **Attempts** to use transitions to link claim(s), counterclaims, reasons and evidence, but they are **simple and infrequent.** Provides a **basic** conclusion or **concluding statement in an attempt** to support the argument.	Builds and **maintains a clear** structure to develop the argument. **Logically** organizes claim(s), counterclaims, evidence and reasoning. Uses **effective** transitions to create cohesion and clarify the relationships among claim(s), counterclaims, reasons and evidence. Provides a **logical** conclusion to support the argument presented.	Builds and maintains a **sophisticated** structure to develop the argument. **Skillfully** organizes claim(s), counterclaims, evidence and reasoning to **strengthen the argument**. **Consistently** uses a **variety** of transitions as well as varied sentence structures to create a **strong** cohesion and clarify the relationships among claim(s), counterclaims, reasons and evidence. Provides a **thorough** conclusion to support the argument presented.
Language/ Conventions	**Lacks or uses an inappropriate** formal tone or voice. **Lacks** a task appropriate writing style. Uses **simple or inappropriate** word	**Uses a weak formal tone or voice and/or has lapses** in appropriate formal tone or voice. **Attempts** to establish a task appropriate	**Establishes and maintains** a formal tone or voice. **Establishes and maintains** a task appropriate writing style.	**Consistently** establishes and maintains a **sophisticated** formal tone or voice. **Consistently** establishes and maintains a **sophisticated,** task appropriate writing

Scoring Elements	Score Point 1	Score Point 2	Score Point 3	Score Point 4
	choice. **Makes significant** errors in the conventions of Standard English grammar, usage, spelling, capitalization and punctuation which **interfere** with understanding the writing.	writing style. **Attempts** to use appropriate word choice. Makes **frequent** errors in using the conventions of Standard English grammar, usage, spelling, capitalization and punctuation which **may interfere** with understanding the writing.	**Effectively** uses appropriate word choice. **Effectively** uses the conventions of Standard English grammar, usage, spelling, capitalization and punctuation with **minor** errors that **do not interfere** with understanding the writing.	style. **Consistently** uses **effective** and **varied** word choice. **Skillfully** uses the conventions of Standard English grammar, usage, spelling, capitalization and punctuation with **few**, minor errors that do not interfere with understanding the writing.

KAS Argumentation Rubric–11th Grade On-Demand Writing

Guiding Principle C1: Students will compose arguments to support claims in an analysis of substantive topics or texts, using valid reasoning and relevant and sufficient evidence.

Scoring Elements	Score Point 1	Score Point 2	Score Point 3	Score Point 4
Clarity and Coherence	Makes claim(s) that may **lack** focus or be **unclear.** **Misses many or all** demands of the prompt.	Makes **general** claim(s) that address the prompt, but may have **lapses** in focus. **Attempts** to address **some** demands of the prompt.	Introduces and maintains **precise and knowledgeable claim(s)** and establishes the significance of those claim(s). Addresses **all** demands of the prompt.	**Thoroughly** introduces and maintains precise, knowledgeable claim(s) and **clearly** establishes the significance of the claim(s). **Thoroughly** addresses all demands of the prompt.
Counterclaims	Makes an **ineffective attempt or** makes **no attempt** to acknowledge opposing claims. Makes an **ineffective attempt or** makes **no attempt** to counter and/or refute opposing claims.	**Attempts** to acknowledge opposing claims, but **lacks** insight, interpretation or clarification. **Attempts** to counter and/or refute opposing claims.	**Acknowledges and distinguishes** claim(s) from alternate or opposing claims with insight, interpretation or clarification. **Counters and refutes** opposing claims.	**Skillfully** acknowledges and distinguishes claim(s) from alternate or opposing claims with insight, interpretation or clarification. **Thoroughly** counters and refutes opposing claims with **carefully selected evidence.**
Support	Includes **minimal or no purposeful** support of claim(s) and/or opposing claims with evidence. Provides **incomplete, inaccurate and/or irrelevant** explanations of evidence and ideas. Provides **minimal or unrelated** reasoning to support claim(s).	**Attempts** to support claim(s) and/or opposing claims with evidence. Provides **vague and/or general** explanations of evidence and ideas. Provides **vague and/or general** reasoning to support claim(s).	Develops claim(s) and/or opposing claims fairly and thoroughly with **logical reasoning and relevant evidence.** Provides **the most relevant** evidence to support claim(s) and opposing claims. Provides reasoning that **points out the strengths and limitations of claim(s)** and opposing claims.	**Fairly and thoroughly develops and supports** claim(s) and/or opposing claims with **insightful** reasoning and **carefully selected,** relevant evidence that **strengthens the argument.** Provides **thorough and effective explanations** of the most relevant evidence and ideas. Provides **complex** reasoning to **clarify** the strengths, limitations and/or **nuances** of claim(s) and opposing claims.
Sourcing	Uses **one or none** of the provided sources or **ineffectively** uses a minimum of two provided sources to support the claim(s) and/or opposing claims. Cites **little or no** evidence. **Little or no** use of quotes and/or paraphrasing of details, examples and ideas.	Uses a minimum of two provided sources to **attempt** to support the claim(s) and/or opposing claims. **Inconsistently** cites evidence. **Attempts** to quote and/or paraphrase details, examples and ideas.	**Accurately and effectively** uses a minimum of two provided sources to support the claim(s) and/or opposing claims. **Effectively** cites evidence by quoting and/or paraphrasing details, examples and ideas.	Accurately and **skillfully** uses a minimum of two provided sources to support the claim(s) and/or opposing claims. **Consistently and thoroughly** cites evidence by quoting and/or paraphrasing details, examples and ideas.

Scoring Elements	Score Point 1	Score Point 2	Score Point 3	Score Point 4
Organization	Builds **minimal or no overall** structure for the argument. **Ineffectively** organizes claim(s), counterclaims, reasons and evidence, creating a lack of cohesion. Makes a **minimal attempt or** makes **no attempt** to use words, phrases and clauses to link sections of the text, claim(s), opposing claims, reasons and evidence. Provides a **weak conclusion or lacks a conclusion** to support the argument presented.	**Attempts** to build a structure for the argument. **Attempts to organize** claim(s), counterclaims, reasons and evidence, but contains **some lapses that disrupt the cohesion or are inappropriate for the context**. **Attempts** to use words, phrases and clauses to link sections of the text, claim(s), opposing claims, reasons and evidence, but they are **simple and infrequent.** Provides a **basic** conclusion or **concluding statement in an attempt** to support the argument presented.	Builds and **maintains a clear** structure to develop the argument. **Logically sequences** claim(s), counterclaims, reasons and evidence. Uses **effective** words, phrases and clauses as well as **varied syntax** to **link the major sections of the text, create cohesion and clarify the relationships** between claim(s) and reasons, between reasons and evidence, and between claim(s) and opposing claims. Provides a **logical** concluding statement or section that follows from and supports the argument presented.	Builds and maintains a **sophisticated** structure to develop the argument. **Skillfully** sequences claim(s), counterclaims, reasons and evidence to **strengthen the argument.** **Consistently** uses **a variety** of effective words, phrases and clauses as well as varied syntax to create a **strong** cohesion and clarify the relationships between claim(s) and reasons, between reasons and evidence, and between claim(s) and opposing claims. Provides a logical, **thorough** concluding statement or section that follows from and **clearly solidifies** the argument presented.
Language/ Conventions	**Lacks or uses an inappropriate** formal tone or voice. **Lacks** a task appropriate writing style. Uses **simple or inappropriate** word choice. **Makes significant** errors in the conventions of Standard English grammar, usage, spelling, capitalization and punctuation which **interfere** with understanding the writing.	**Uses a weak formal tone or voice and/or has lapses** in appropriate formal tone or voice. **Attempts** to establish a task appropriate writing style. **Attempts** to use appropriate word choice. Makes **frequent** errors in using the conventions of Standard English grammar, usage, spelling, capitalization and punctuation which **may interfere** with understanding the writing.	**Establishes and maintains** a formal tone or voice. **Establishes and maintains** a task appropriate writing style. **Effectively** uses appropriate word choice. **Effectively** uses the conventions of Standard English grammar, usage, spelling, capitalization and punctuation with **minor** errors that **do not interfere** with understanding the writing.	**Consistently** establishes and maintains a **sophisticated** formal tone or voice. **Consistently** establishes and maintains a **sophisticated,** task appropriate writing style. **Consistently** uses **effective** and **varied** word choice. **Skillfully** uses the conventions of Standard English grammar, usage, spelling, capitalization and punctuation with **few,** minor errors that do not interfere with understanding the writing.