Kentucky Performance Rating for Educational Progress

2017–18 Technical Manual

Version 2.0

Pearson

K-PREP 2017-2018 Update

The 2017-2018 academic year marked the seventh year of Kentucky Performance Rating of Educational Progress (K-PREP), teaching and assessing students for success beyond K-12 academic instruction. The K-PREP *Technical Manual* contains information on the development, scoring, and maintenance of the K-PREP assessment program. 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. For the 2017-2018 assessment cycle, changes were made within K-PREP from the previous year, thus affecting the description and presentation of information in both the technical manual and yearbook. These changes and their impact on the documents are summarized below.

The Stanford Achievement Test
From its inception, the K-PREP assessment program included the Stanford Achievement Test, *Tenth Edition* (Stanford 10), to report Kentucky student achievement in relation to student achievement across the nation. Additionally, items from SAT10 were used to report student achievement aligned to Kentucky Academic Standards (KAS). For the spring 2018 test administration, a legislative mandate removed the SAT10 portion of the K-PREP assessments. While this reduced the overall test length, the length of the KAS-aligned (standards based) portion of the K-PREP assessments did not change.

Language Mechanics
With the removal of the Stanford 10 component of the K-PREP assessment, the K-PREP Language Mechanics assessment at grades 4 and 6 was also dropped from test administration. This test was composed solely from the Stanford 10 Language test at the corresponding grades, but used scores mapped to Novice, Apprentice, Proficient, and Distinguished to indicate student performance to these performance levels.

Science
In spring 2018, a new Science assessment was administered statewide in grades 4 and 7. After the test administration, a standard setting workshop was held with Kentucky educators to recommend performance standards (i.e., cut scores) to classify student performance into Novice, Apprentice, Proficient, and Distinguished. A separate technical report describes the standard setting process and the recommended performance standards.

Table of Contents

LIST OF TABLES **7**

LIST OF FIGURES **8**

1. BACKGROUND **9**

KENTUCKY INSTRUCTIONAL RESULTS INFORMATION SYSTEM (1992-1998) 9
COMMONWEALTH ACCOUNTABILITY TESTING SYSTEM (1998-2010) 9
UNBRIDLED LEARNING (2010-2016) 10
KENTUCKY'S TRANSITION TO ESSA (2017-PRESENT) 10
ORGANIZATIONS AND GROUPS INVOLVED 11
Kentucky Department of Education *11*
Kentucky Educators *11*
School Curriculum, Assessment, and Accountability Council *12*
National Technical Advisory Panel on Assessment and Accountability *12*
Contractors *12*
KENTUCKY PERFORMANCE RATING FOR EDUCATIONAL PROGRESS ASSESSMENT PROGRAM 13
Reading *13*
Mathematics *14*
Science *14*
Social Studies *14*
Language Mechanics *14*
On-Demand Writing *14*

2. TEST DEVELOPMENT **15**

K-PREP CONTENT AND KENTUCKY ACADEMIC STANDARDS ALIGNMENT 15
ITEM DEVELOPMENT 16
Item Specifications *16*
Item Writing *17*
Content Advisory Committees *18*
Bias and Sensitivity Review *18*
Item Editing *21*
SCORING GUIDES 21
FORMS DEVELOPMENT 21
Test Design and Blueprints *21*
Form Content Alignment *25*
Statistical Guidelines *25*
Field-testing *25*
TEST BOOKLET DESIGN 26
BRAILLE AND LARGE PRINT TEST MATERIALS 27

3. TEST ADMINISTRATION **28**

TEST ADMINISTRATION WINDOW 28
TEST MAKE-UP PROCEDURES 28
ELIGIBILITY REQUIREMENTS AND EXEMPTIONS 28
ACCOMMODATIONS 29
TEST ADMINISTRATION PROCEDURES 29
District Assessment Coordinators *29*
District and Building Assessment Coordinators' Manual *30*
Test Administrators' Manual *30*
Interpretive Guide *30*

Test Security 30

4. REPORTS 31

Appropriate Uses for Scores and Reports 31
Individual Student Report 31
Kentucky Performance Report 31
Description of Scores 31
Raw Score 31
Scale Score 31
Student Performance Level 32
Lexiles and Quantiles 32
Description of Reports 32
Student Report 32
School Listing Report 32
Kentucky Performance Report 32
Cautions for Score Interpretations and Use 33
Understanding Measurement Error 33
Interpreting Scores at Extreme Ends of the Distribution 33
Limitations When Comparing Scale Scores at Reporting Group Levels 33
Inappropriateness of Comparing Scale Scores Between Content Tests 34
Program Evaluation 34

5. PERFORMANCE STANDARDS 35

Performance Level Descriptions and College/Career Readiness 35
K-PREP and College/Career Readiness 35
On-Demand Writing 36
Science and Social Studies 38

6. ITEM ANALYSES 40

Item Mean Scores 40
Item-Test Score Correlations 40
Differential Item Functioning 41
Item Response Theory 43
On-Demand Writing Item Analysis 46

7. SCALING 47

Rationale 47
Measurement Models 47
Process 48
Overview 48
Quality Control 49
Scaled Scores 49
Transformation of Raw Scores 49
Considerations and Limitations 51
Results 53
Lexiles and Quantiles 53

8. EQUATING 54

Rationale 54
Process 54
Linking Items 54
Analysis 57
Field-test Item Calibration 58

9. RELIABILITY 59

Definition of Reliability 59
Estimating Reliability 60
Test-Retest Reliability Estimation 60
Alternate Forms Reliability Estimation 60
Internal Consistency Reliability Estimation 60
Domain Reliability Estimation 61
Standard Error of Measurement 61
Use of the Standard Error of Measurement 61
Conditional Standard Error of Measurement 62
Scoring Reliability for Open-Ended Items 62
Reader Agreement 62
Score Resolutions 63
Reliability of Performance Level Categorization 63
Accuracy and Consistency 64
Calculating Accuracy 64
Calculating Consistency 65
Calculating Kappa 65

10. VALIDITY 67

Argument-Based Approach to Validity 67
Scoring 68
Generalization 68
Extrapolation 68
Implication 69
Validity Argument Evidence for the Kentucky Assessment 69
Scoring 69
Generalization 71
Extrapolation 72
Implication 72
Summary of Validity Evidence 73

11. PERFORMANCE SCORING 74

Rubric Creation 74
Rangefinding 75
Scoring Process 75
Recruitment 75
Training 76
Quality Control 77
Security 78

12. QUALITY CONTROL PROCEDURES 79

Test Construction 79
Non-Scannable Documents 79
Data Preparation 79
Production Control 80
Scanning and Editing 80
Performance Scoring 81
Equating 81
Scoring and Reporting 82

13. ON DEMAND WRITING 83

BASE SCALE 83
EQUATING 84
SCALED SCORES 85
GLOSSARY OF TERMS 88
REFERENCES 90
APPENDIX A. READING ITEM WRITER TRAINING 92
APPENDIX B. MATHEMATICS ITEM WRITER TRAINING 102
APPENDIX C. ITEM DEVELOPMENT REVIEW CHECKLIST 108
APPENDIX D. READING CONTENT COMMITTEE REVIEW 110
APPENDIX E. MATHEMATICS CONTENT COMMITTEE REVIEW 115
APPENDIX F. ITEM CONTENT COMMITTEE REVIEW CHECKLIST 120
APPENDIX G. ITEM BIAS COMMITTEE REVIEW 122
APPENDIX H. ITEM BIAS COMMITTEE REVIEW CHECKLIST 127
APPENDIX I. READING PASSAGE BIAS COMMITTEE REVIEW 129
APPENDIX J. READING PASSAGE BIAS COMMITTEE REVIEW CHECKLIST 134
APPENDIX K. ODW ITEM WRITER TRAINING 136
APPENDIX L. ODW CONTENT COMMITTEE REVIEW 143
APPENDIX M. ODW PROMPT REVIEW CHECKLIST 146
APPENDIX N. ODW BIAS COMMITTEE REVIEW 150
APPENDIX O. ODW BIAS REVIEW CHECKLIST 155
APPENDIX P. SCIENCE BIAS REVIEW CHECKLIST 157
APPENDIX Q. ON-DEMAND WRITING SCORING RUBRIC 175

List of Tables

Table 2.1 K-PREP Bias and Content Review Meeting Participation Summary 20
Table 2.2 K-PREP Reading Test Blueprint 22
Table 2.3 K-PREP Mathematics Test Blueprint 22
Table 2.4 K-PREP Science Test Blueprint 23
Table 2.5 K-PREP Social Studies Test Blueprint 23
Table 2.6 2012 K-PREP On-Demand Writing Test Blueprint 23
Table 2.7 2013 K-PREP On-Demand Writing Test Blueprint 23
Table 2.8 2014 K-PREP On-Demand Writing Test Blueprint 24
Table 2.9 2015 K-PREP On-Demand Writing Test Blueprint 24
Table 2.10 2016 K-PREP On-Demand Writing Test Blueprint 24
Table 2.11 2017 K-PREP On-Demand Writing Test Blueprint 24
Table 2.12 2018 K-PREP On-Demand Writing Test Blueprint 24
Table 2.13 K-PREP Test Booklet by Grade 27
Table 5.1 Reading and Mathematics Final Cut Points and Impact Data 36
Table 5.2 ODW Final Performance Level Cut Points 37
Table 5.3 ODW Final Round Impact Data 38
Table 5.4 2011 Science and Social Studies Performance Level Distribution (KCCT) .38
Table 5.5 2012 Science and Social Studies Cut Points and Impact Data (K-PREP) ...38
Table 6.1 Item 2x2 Contingency Table for the *k*th Score Level 42
Table 6.2 Criteria for Item Fit Statistics 45
Table 7.1 Proficient Cut Points for Derived Scaled Scores 50
Table 7.2 Raw Score to Scale Score Conversion 52
Table 7.3 Scaled Scores by Performance Level 52
Table 8.1 2018 Linking Items by Item Type 55
Table 8.2 2018 Reading Linking Items by Test Blueprint 56
Table 8.3 2018 Mathematics Linking Items by Test Blueprint 57
Table 8.4 Unstable Linking Items 58
Table 9.1 Example Accuracy Classification Table 64
Table 9.2 Example Accuracy Classification Table for Proficient Cutpoint 65
Table 9.3 Example Consistency Classification Table 65
Table 13.1 Raw Score Cut Points and 2015 Raw Score Impact 84
Table 13.2 Derived ODW Cut Points on Theta Metric 84
Table 13.3 2016 ODW Equating Constants 85
Table 13.4 ODW Scaled Scores by Performance Level 87

List of Figures

Figure 6.1 Graph of 1PL Model 43
Figure 6.2 Graph of Partial Credit Model for Three-point Item 44
Figure 6.3 Observed and Expected Performance on Item of Average Difficulty.....45
Figure 6.4 Observed and Expected Performance on Difficult Item 46

1. Background

Over the last twenty years, Kentucky's assessment program has evolved to such an extent that it is now one of the country's leading assessment program in preparing students for future success. The assessment program has utilized resources within Kentucky as well as 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, and they continue to make significant contributions today. This chapter provides a history of the Kentucky assessment program and the contributors whom have guided its progression.

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. Under KIRIS, the various test item types were administered in three distinct assessment components: a traditional assessment (multiple-choice and open-ended questions), performance event (performance task involving individual and group problem solving skills), and portfolio assessment (student-chosen collection of work). Student performance within KIRIS was divided into four achievement categories: novice, apprentice, proficient, and distinguished.

Commonwealth Accountability Testing System (1998-2010)

Beginning in 1999, the content 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 as well as creating 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 the areas of reading, language arts and mathematics.

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, which 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 is college and career readiness for all Kentucky students, which itself has been defined by the goals put forth by the Partnership for Assessment of Readiness for College and Careers 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 covers 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, sponsored by the National Governors Association (NGA) and the Council of Chief State School Officers (CCSSO), while the standards for science and social studies remain from the previous curriculum standards framework.

The Kentucky Performance Rating for Educational Progress (K-PREP) is the collection of tests created and administered to assess 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 (CRT) portion of K-PREP is built using test content written specifically for Kentucky's assessment. Student performance from the CRT portion is divided in the four achievement categories used in the previous testing systems: novice, apprentice, proficient, and distinguished (see chapter 5, "Performance Standards," for a description of how these achievement levels were defined). 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 (see chapter 4, "Reports"). Beginning in 2018, Stanford 10 is no longer a component of the K-PREP assessments.

Kentucky's Transition to ESSA (2017-Present)

As Kentuckians engaged in the development of a new accountability system under the Every Student Succeeds Act (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 board's vision that each and every student is empowered and equipped with the knowledge, skills and dispositions to pursue a successful future; 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 each and every student; and the department's underlying values of equity, achievement and integrity, provide coherence with the state's new accountability system.

Under ESSA and Senate Bill 1, Kentucky is required to meaningfully differentiate between schools through its accountability system in an effort 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 Kentucky Board of Education approved a new accountability system to be implemented beginning with the 2018-19 school year. Therefore, the 2017-18 school year is a transition year.

In 2017-2018, as in the past, Kentucky public school students in grades 3 through 8 completed Kentucky Performance Rating for Educational Progress (K-PREP) tests in five content areas. Students take reading and mathematics assessments annually in grades 3 through 8. Other subjects are assessed once per grade level with science assessed in grades 4 and 7; and social studies and on-demand writing assessed in grades 5 and 8. At the high school level, the state is moving to summative tests in English, mathematics, science and social studies developed by Kentucky teachers and aligned to the Kentucky Academic Standards. A new social studies summative assessment will be developed once new social studies standards are approved.

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 K-PREP program.

Kentucky Department of Education

The Kentucky Department of Education (KDE), located in Frankfort, Kentucky, leads the design, implementation, and reporting of the accountability model and its components. KDE consists of smaller organizations that provide specific guidance to K-PREP. The Office of Assessment and Accountability (OAA) works directly on K-PREP with intra-office support from the Division of Accountability Data and Analysis (data and statistics) and the Division of Assessment Support (communications). In addition, members of the Office of Next Generation Learners provide content support on the K-PREP tests, reviewing and providing feedback on the construction of test forms.

Kentucky Educators

Educators play the next most significant role in the design and maintenance of large-scale assessment programs in the Commonwealth, second only to KDE itself. During the initial development stages of an assessment program, educators are solicited to provide input on assessment design, including the best methods for assessing particular 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 assessment development. As discussed in the next chapter, "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 K-PREP test forms. Here, educators directly affect test content, removing items from consideration or proposing changes to items to make them more appropriate for testing.

In addition to item review meetings, educators participate in other meetings held throughout an assessment program. During the summer of 2012, Kentucky educators were assembled in Lexington, Kentucky, to recommend performance standards for the reading, mathematics and writing tests. Educators used their expertise to provide input on achievement level definitions and cut points for the K-PREP tests. These *standard setting* meetings are discussed in more detail in chapter 5.

School Curriculum, Assessment, and Accountability Council

Kentucky Revised Statues (KRS) 158.6452 requires that a School Curriculum, Assessment, and Accountability Council (SCAAC) is created to study, review, and make recommendations concerning Kentucky's system of setting academic standards, assessing learning, identifying academic competencies and deficiencies of individual students, holding schools accountable for learning, and assisting schools to improve their performance. The council shall advise the Kentucky Board of Education and the Legislative Research Commission on issues related to the development of and communication of the academic expectations and core content for assessment, the development and accountability program, recognition of high performing schools, imposition of sanctions, and assistance for schools to improve their performance under KRS 158.6453, 158.6455, 158.782, and 158.805.

National Technical Advisory Panel on Assessment and Accountability

Kentucky Revised Statues (KRS) 158.6453 and 158.6455 require that the National Technical Advisory Panel on Assessment and Accountability (NTAPAA) is consulted on any proposed additions and changes to the Kentucky assessment and accountability system. NTAPAA is composed of measurement experts who possess years of experience in large-scale testing and state accountability programs; it is an assemblage of persons with diverse backgrounds who can respond to the many facets of measurement design and implementation. When requested, NTAPAA and KDE convene, along with other organizations (see Contractors), to discuss measurement and/or accountability issues as determined by KDE.

Contractors

Human Resources Research Organization

Human Resources Research Organization (HumRRO), a measurement solutions provider based in Louisville, Kentucky, has a long-standing involvement with the Kentucky assessment program. During its involvement, HumRRO has conducted several alignment and validation studies for presentation to NTAPAA as well as for state and national conferences. Also, HumRRO provides quality control verification, replicating measurement analyses performed by prime contractors of state assessment programs, including Kentucky. Chapter 7, "Scaling," provides more detail regarding HumRRO's involvement in the measurement analyses conducted on K-PREP by Pearson.

MetaMetrics
MetaMetrics®, based in Durham, North Carolina, provides measurement solutions that link assessment results to real-world instruction. The most visible of these solutions are the Lexile® and Quantile® measures that link student performance on assessments to content material at complexity (or difficulty) levels near student ability. Linking assessment results to instruction in this fashion gives users access to content material that will foster development toward the increasing cognitive demands required at subsequent grade levels. Chapter 4, "Reports," and chapter 7, "Scaling," provide descriptions of these measurement frameworks.

Pearson
Pearson's U.S. educational assessment headquarters are located in Iowa City, with additional offices in Austin and San Antonio, which together provide a full range of assessment and measurement services to states and districts throughout the U.S. As the prime contractor for K-PREP, Pearson works with KDE—through its management of project schedules and deliverables, communications, and client meetings – to develop valid and reliable assessments that measure in a fair manner the educational progress of Kentucky students. By means of this report and the accompanying documentation, Pearson will describe in sufficient detail all aspects of the development and delivery of K-PREP, from item generation to psychometric analysis to score interpretation.

ILSSA - University of Kentucky
The ILSSA group is composed of staff at the University of Kentucky dedicated to designing and implementing large-scale assessments for students with significant cognitive disabilities. ILSSA has been the leader of Kentucky's alternate assessment program since its inception in 1990. ILSSA has developed a separate Alternate Kentucky Performance Rating for Education Progress (Alternate K-PREP) Technical Manual for the Alternate K-PREP assessment program.

Kentucky Performance Rating for Educational Progress Assessment Program

The new assessment program in Kentucky, a result of Senate Bill 1, was designed to prepare students for the demands of the 21st century. These demands were rooted in the Common Core State Standards, which were adopted for K-PREP in reading, mathematics and writing, and the core content for science and social studies adopted from the previous curriculum framework. This section provides a brief description of the content areas assessed through K-PREP. Chapter 2 outlines the test blueprint for each test.

Reading

The Reading tests focus on three main skills: reading comprehension, language use and vocabulary. Students are expected to develop reading comprehension skills through increasing text complexity from one grade to the next and by making connections across multiple texts. Also, students should develop a craft of appropriate language use as well as the ability to understand words and phrases and their relationships, especially when acquiring new vocabulary. More information on content standards for Reading can be found at
KDE website (ELA).

Mathematics

The Mathematics tests at grades 3-5 assess knowledge and foundations in whole numbers, operations, fractions, decimals as well as geometry. The tests at grades 6-8 build upon knowledge assessed at the lower grades and include algebra and probability and statistics. More information on content standards for Mathematics can be found at KDE website (Math).

Science

For the K-PREP Science tests (2012-2015), the standards were organized around seven "Big Ideas" important to the discipline: structure and transformation of matter, motion and forces, the Earth and the Universe, unity and diversity, biological change, energy transformations, and interdependence. This organization of concepts is the same across grades to allow multiple opportunities to learn these scientific concepts. New standards for science were recently adopted and future technical manuals will contain more information regarding the new standards.

In 2015, Kentucky adopted a new set of Science academic standards that featured assessable performance expectations of what students should know and be able to do with foundations of science and engineering practices, core disciplinary ideas, and cross cutting concepts. More information on these new content standards can be found at KDE website (Science). In spring 2018, new Science assessments were administered in grades 4 and 7.

Social Studies

In the Social Studies tests, students are expected to develop the ability to make informed decisions as citizens of a culturally diverse, democratic society in an interdependent world. The Social Studies tests assess concepts organized into five "Big Ideas": government and civics, cultures and societies, economics, geography, and historical perspective. This organization of concepts is the same across grades to allow multiple opportunities for developing an understanding of the Big Ideas. More information on content standards for Social Studies can be found at KDE website (Social Studies).

Language Mechanics

From 2012 to 2017, the Stanford 10 language subtest was used as the K-PREP "Language Mechanics" test to measure word- and sentence-level skills and whole text skills in mechanics and expression. More information on content standards Language within English Language Arts can be found at KDE website (ELA). This test is no longer used within K-PREP.

On-Demand Writing

On-Demand Writing (hereafter, writing) assesses writing skills through goals set forth through the Kentucky Academic Standards. There are goals specific to writing genre (e.g., narrative, informative/exploratory, and argumentative) and goals for writing conventions (e.g., organization and style). Students respond to two types of prompt stimuli: a short stimulus outlining a situation and an extended stimulus that includes a reading passage. Writing ability is determined by performance across both types of stimuli. The scoring rubric used for the writing test is provided in the Appendix P of this manual and can be found online at KDE website (K-PREP).More information on content standards of Writing within English Language Arts can be found at
KDE website (ELA).

2.Test Development

The construction of test forms for K-PREP is a coordinated effort between KDE and the testing contractor, adhering to guidelines that promote fair and ethical testing practices. However, 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, when obtained, using data to determine the statistical quality of the content. Several factors are considered when designing the K-PREP test forms. This chapter provides a description of the test development process of K-PREP, including item development, content and statistical guidelines considered, and test booklet design.

K-PREP Content and Kentucky Academic Standards Alignment

One emphasis during K-PREP content development—item and passage development—is alignment to the Kentucky Academic Standards (KAS). The K-PREP testing contractor began 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 K-PREP blueprints). This allowed the testing contractor to create a robust item pool for the K-PREP assessments that appropriately represents the KAS. The testing contractor also uses an item bank application that maintains the blueprint requirements to guide the content development process and promote adequate coverage of KAS for all future administrations of the K-PREP. The K-PREP test blueprints can be found in the technical manual.

For K-PREP content development, Kentucky's testing contractor designs item writer training material that includes references and discussions to the Kentucky Academic Standards; the KAS are included with key aspects highlighted for training purposes. Training on KAS for content development is essential to address interpretations of the standards so that all K-PREP assessment content is developed to the same guidelines. Item writer training material is reviewed and discussed thoroughly between KDE and the testing contractor, 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: 1) content development requirements may change year to year, and 2) interpretations pertaining to assessing KAS may change, dictated by national perspectives.

During item writer training, the testing contractor presents the Kentucky Academic Standards to the trainees, pointing out key aspects to consider when developing content. These key aspects include specific decomposition of standards into concrete domain targets (e.g., point of view and relationship between texts, in Reading). The goal of this portion of training is to underscore the breadth of content necessary for assessing Kentucky's students on skills within the KAS framework. In addition, the trainees are provided with exemplars to guide their content development.

The testing contractor conducts internal reviews of content submitted by the contracted item writers. These initial reviews focus on appropriateness as well as specificity in assessing KAS. The testing contractor engages with the item writers to

discuss item alignment and suggested content revisions, as necessary. The testing contractor has the authority to, and may, align items to KAS differently than what was intended by the item writers. Items may be rejected by the testing contractor due to poor alignment to KAS. The assessment content, alignments, and reviews by the testing contractor are prepared for review by KDE.

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

During item review workshops, participants review each piece of assessment content for its KAS alignment, in addition to reviews for 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 the testing contractor may discuss recommended changes with regards to previous decisions in KAS alignment. Changes in KAS alignment from 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 assessment content for accuracy prior to use by the testing contractor in building assessment forms. KDE has the final authority on KAS alignment of assessment content.

Item Development

The testing contractor for K-PREP developed item content for Reading, Mathematics, and Writing subject areas. The goal of item development for these subject areas was to build upon item banks for assessing the *Kentucky Academic Standards*.

Item Specifications

To develop appropriate content for large-scale testing, individuals tasked with writing test content—items and passages—must follow specific guidelines. These guidelines can be general to subject-area specific and give the item writers the parameters for creating content appropriate and suitable for assessing achievement. General guidelines for item writing include:

- 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 also be aligned to *Depth of Knowledge* levels, to the extent that an adequate range of skill level is represented. In addition, guidelines of item writing by subject area are used to cover the specific aspects of the particular subject area. For example, for 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. For example, multiple-choice answer options for Mathematics items should either in be ascending or descending order when containing numerical values. Item type and format guidelines are used as well to promote consistency and appropriateness of items' presentation, task, and, in the case of multiple-choice items, answer options.

Furthermore, the accessibility of items for all intended test takers is specified through guidelines of *universal design*. These guidelines 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. Appendices A through O of this manual contain various materials used within the item development process, including presentations for workshops and item review checklists discussed in the next few sections. The materials in these appendices reflect previous years of item development work for K-PREP. The processes highlighted through these materials are the objects of importance, rather than the actual years.

Item Writing

Item Writers/Training

Subject matter experts from the field of education are recruited to develop test content for K-PREP. These individuals enter into an agreement with the K-PREP testing contractor outlining the tasks, proposed compensation, and guidelines for submitting completed work.

Kentucky's testing contractor provides extensive training for writers prior to item or task development. For K-PREP, 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. In addition, Kentucky's testing contractor discusses policies of content security and ownership. Training provides the foundation of best practices for item development.

Item Authoring

Once items are submitted by item writers, the testing contractor executes a process of review and editing before the items are included into item banking applications. During this phase of item development, subject matter experts from the testing contractor review item metadata (e.g., standard/benchmark/objective, answer key, cognitive level, etc.) for accuracy, making revisions as needed. Also, items are reviewed for appropriate and accurate content as well as 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. The process of reviewing and editing the items submitted by item reviews allows the testing contractor to publish items suitable for use in large-scale testing.

Quality Control

Throughout the item development process, quality control is instituted in a variety ways. From the initial review of submitted items, multiple staff persons from the testing contractor 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, researched by the testing contractor. In the case of inaccurate information 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. Also, scoring rubrics, for performance items, are reviewed for what could lead to errors or other issues in hand scoring. Furthermore, all revisions to items and other test content are made through the consultation of staff from the testing contractor for agreement, rather than through a single individual.

Content Advisory Committees

Kentucky educators and other stakeholders take part in the development of K-PREP test content through participation in item review committees. The content advisory committee reviews newly-developed items for content, alignment to the standards, and appropriateness at the intended grade level. The educators work in groups, facilitated by the testing contractor, to recommend that items are accepted for testing, rejected for testing, or conditionally accepted (i.e., acceptance with minor modifications to the items).

Bias and Sensitivity Review

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

Table 2.1 provides a demographic summary of participation in K-PREP Bias and Content review meetings. These meetings were held during the phase of K-PREP item development which occurred during the first few years of the K-PREP assessment program (i.e., 2011-2013). The table below provides a summary of the participants including gender and ethnicity distribution, highest level of education attained, professional position at the time of participation, experience with special population (e.g., ESL), and average number of years in education. For summary purposes, the professional positions of the participants were classified into three groups: teacher, non-teacher educator, and general public.

Teachers were those individuals that were responsible for classroom instruction at the time of their participation in the bias and content meetings. Non-teacher educators were those individuals with a background in education, but were not K-12 classroom teachers. These individuals include curriculum specialists, administrators, and university instructors. Finally, the general public category was used for individual that were not directly involved with education at the time of participation in the bias and content meetings. However, these individuals may have been previously involved in education (e.g., retired teachers).

The table summarizes participation across six bias and content review meetings. The distribution of individuals with special populations may vary across meetings depending on the purpose of the meeting. For example, the percentage of individuals having experience with students of special populations for the 2013 Panel 1 is the result of that meeting having a focus of reviewing for bias in K-PREP assessment content.

Table 2.1 K-PREP Bias and Content Review Meeting Participation Summary

Category	Group	2011 Panel 1	2011 Panel 2	2012 Panel 1	2012 Panel 2	2013 Panel 1	2013 Panel 2
Gender	Female	77%	82%	65%	81%	78%	88%
	Male	23%	18%	35%	19%	22%	12%
Ethnicity	African American	7%	8%	29%	6%	11%	12%
	Asian	7%	2%	6%	3%	--	6%
	White	86%	90%	59%	91%	89%	82%
	Missing	--	--	6%	--	--	--
Education Attainment	Bachelors	5%	4%	--	--	--	--
	Masters	53%	42%	35%	45%	67%	35%
	Doctorate	7%	12%	18%	9%	--	12%
	Rank I/Education Specialist	35%	42%	47%	45%	33%	53%
Position Type	Teacher	70%	58%	47%	67%	67%	59%
	Non-Teacher Educator	21%	32%	24%	30%	22%	29%
	General Public	9%	10%	29%	3%	11%	12%
Special Populations	No	93%	94%	71%	94%	44%	94%
	Yes	7%	6%	29%	6%	56%	6%
Years Teaching	Number of Years	14.6	18.3	22.9	17.7	17.4	19.4
	Number of Years in KY	13.6	17.1	20.8	16.9	15.5	17.9

Item Editing

After the various reviews are conducted, the testing contractor 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 – administered to students within a standard testing environment for the purposes of collecting item performance data.

Scoring Guides

For constructed response items—short answer and extended response items, on K-PREP—scoring guides are required to describe criteria that differentiate item responses by the achievable score points. For K-PREP, 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 11, "Performance Scoring").

Forms Development

Developing test forms is a process by which assessment specialists select and sequence items that assess subject area content as specified by 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.

Test Design and Blueprints

The *test design* can be thought of as 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, etc.), and the number of sections a test may be divided into, if preferred. 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. For K-PREP, norm-referenced test material is included which adds design considerations to the overall assessment forms. In particular, decisions were made on where this additional material would be located within the test form as well as how many items would be included. Also, large-scale assessments often include field-test items; the placement of these items within the test form – in one section or spread throughout—becomes an additional design factor.

Test blueprints, on the other hand, mainly provide specifications on content coverage —the number of items required per domain/reporting category. This includes how item types – e.g., multiple choice and constructed response items—are chosen across domains/reporting categories 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. through Table 2. provides the test blueprints used for constructing the

K-PREP tests in Reading, Mathematics, Science, and Social Studies. Note: The Science test blueprint is included for historical reference.

Table 2.2 K-PREP Reading Test Blueprint

Grade(s)	Domain	Item Types			Domain Coverage	Passage Genre Item Coverage	
		MC	*SA*	*ER*		*Literary*	*Informative*
3	Key Ideas	X	X	NA	25%	50%	50%
	Craft and Structure	X	X (rotate)	NA	25%	50%	50%
	Integration of Ideas	X	X (rotate)	NA	25%	50%	50%
	Vocabulary and Acquisition	X		NA	25%	50%	50%
4, 5	Key Ideas	X	X		25%	50%	50%
	Craft and Structure	X	X		25%	50%	50%
	Integration of Ideas	X		X	25%	50%	50%
	Vocabulary and Acquisition	X			25%	50%	50%
6-8	Key Ideas	X	X		25%	45%	55%
	Craft and Structure	X	X		25%	45%	55%
	Integration of Ideas	X		X	25%	45%	55%
	Vocabulary and Acquisition	X			25%	45%	55%

Table 2.3 K-PREP Mathematics Test Blueprint

Grade(s)	Domain	Item Types			Domain Coverage
		MC	*SA*	*ER*	
3	Operations and Algebraic Thinking	X	3 SA/form - rotate	NA	25%
	Number and Operations in Base Ten	X		NA	20%
	Number and Operations – Fractions	X		NA	25%
	Measurement and Data, Geometry	X		NA	30%
4, 5	Operations and Algebraic Thinking	X	3 SA/form - rotate	1 ER /form - rotates	20%
	Number and Operations in Base Ten	X			25%
	Number and Operations – Fractions	X			25%
	Measurement and Data, Geometry	X			30%
6, 7	Ratios and Proportional Relationships	X	3 SA/form - rotate	1 ER /form - rotates	20%
	The Number System	X			20%
	Expressions and Equations	X			20%
	Geometry	X			20%
	Statistics and Probability	X			20%
8	The Number System and Expressions & Equations	X	3 SA/form - rotate	1 ER /form - rotates	30%
	Functions	X			20%
	Geometry	X			30%
	Statistics and Probability	X			20%

Table 2.4 K-PREP Science Test Blueprint (2012-2015)

Grade(s)	Domain	Item Types			Domain Coverage
		MC	*SA*	*ER*	
4	Physical Science	X	N/A	3 ER /form - rotates	25%
	Earth/Space Science	X			25%
	Life Science	X			30%
	Unifying Ideas	X			20%
7	Physical Science	X	N/A	3 ER /form - rotates	25%
	Earth/Space Science	X			25%
	Life Science	X			20%
	Unifying Ideas	X			30%

Table 2.5 K-PREP Social Studies Test Blueprint

Grade(s)	Domain	Item Types			Domain Coverage
		MC	*SA*	*ER*	
5	Government and Civics	X	N/A	3 ER /form - rotates	20%
	Cultures and Societies	X			15%
	Economics	X			15%
	Geography	X			20%
	Historical Perspective	X			30%
8	Government and Civics	X	N/A	3 ER /form - rotates	25%
	Cultures and Societies	X			15%
	Economics	X			15%
	Geography	X			15%
	Historical Perspective	X			30%

For On-Demand Writing, three essays are administered within each grade, but students are required to respond to only two of the essays. For each grade, there is one passage-based essay and two stand-alone essays. All students must respond to the passage-based essay and choose one of the stand-alone essays. The mode type of the essays varies by and within grade and will vary across years of the Writing assessment. For 2017-2018, one form was administered in each grade. Table 2. through Tables 2.6 through 2.12 show the test blueprint used for the 2012 through 2018 On-Demand Writing assessments.

Table 2.6 2012 K-PREP On-Demand Writing Test Blueprint

Grade	Prompt Mode		
	Stand-Alone A	*Stand-Alone B*	*Passage-Based*
5	Narrative	Opinion	Informative/Explanatory
6	Narrative	Argumentative	Informative/Explanatory
8	Narrative	Informative/Explanatory	Argumentative
10	Informative/Explanatory	Informative/Explanatory	Argumentative
11	Argumentative	Argumentative	Informative/Explanatory

Table 2.7 2013 K-PREP On-Demand Writing Test Blueprint

Grade	Prompt Mode		
	Stand-Alone A	*Stand-Alone B*	*Passage-Based*
5	Narrative	Opinion	Informative/Explanatory
6	Narrative	Informative/Explanatory	Argumentative
8	Narrative	Argumentative	Informative/Explanatory
10	Argumentative	Argumentative	Informative/Explanatory
11	Argumentative	Argumentative	Informative/Explanatory

Table 2.8 2014 K-PREP On-Demand Writing Test Blueprint

Grade	Prompt Mode		
	Stand-Alone A	*Stand-Alone B*	*Passage-Based*
5	Informative/Explanatory	Narrative	Opinion
6	Informative/Explanatory	Narrative	Argumentative
8	Informative/Explanatory	Narrative	Argumentative
10	Argumentative	Argumentative	Informative/Explanatory
11	Informative/Explanatory	Informative/Explanatory	Argumentative

Table 2.9 2015 K-PREP On-Demand Writing Test Blueprint

Grade	Prompt Mode		
	Stand-Alone A	*Stand-Alone B*	*Passage-Based*
5	Informative/Explanatory	Narrative	Opinion
6	Argumentative	Narrative	Informative/Explanatory
8	Narrative	Argumentative	Informative/Explanatory
10	Informative/Explanatory	Informative/Explanatory	Argumentative
11	Informative/Explanatory	Informative/Explanatory	Argumentative

Table 2.10 2016 K-PREP On-Demand Writing Test Blueprint

Grade	Form	Prompt Mode		
		Stand-Alone A	*Stand-Alone B*	*Passage-Based*
5	1	Narrative	Opinion	Informative/Explanatory
	2	Informative/Explanatory	Narrative	Opinion
	3	Informative/Explanatory	Narrative	Opinion
6	1	Argumentative	Narrative	Informative/Explanatory
	2	Argumentative	Narrative	Informative/Explanatory
	3	Informative/Explanatory	Narrative	Argumentative
8	1	Narrative	Argumentative	Informative/Explanatory
	2	Narrative	Argumentative	Informative/Explanatory
	3	Narrative	Informative/Explanatory	Argumentative
10	1	Argumentative	Argumentative	Informative/Explanatory
	2	Informative/Explanatory	Informative/Explanatory	Argumentative
	3	Informative/Explanatory	Informative/Explanatory	Argumentative
11	1	Argumentative	Argumentative	Informative/Explanatory
	2	Informative/Explanatory	Informative/Explanatory	Argumentative
	3	Informative/Explanatory	Informative/Explanatory	Argumentative

Table 2.11 2017 K-PREP On-Demand Writing Test Blueprint

Grade	Prompt Mode		
	Stand-Alone A	*Stand-Alone B*	*Passage-Based*
5	Informative/Explanatory	Narrative	Opinion
8	Narrative	Informative/Explanatory	Argumentative
11	Informative/Explanatory	Informative/Explanatory	Argumentative

Table 2.12 2018 K-PREP On-Demand Writing Test Blueprint

Grade	Prompt Mode		
	Stand-Alone A	*Stand-Alone B*	*Passage-Based*
5	Narrative	Opinion	Informative/Explanatory
8	Narrative	Argumentative	Informative/Explanatory
11	Argumentative	Argumentative	Informative/Explanatory

Form Content Alignment

For new forms development, the testing contractor utilizes two content specialists per K-PREP test form developed. The first content specialist is responsible for constructing a test form meeting both content and statistical requirements. 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, as well as the accuracy of specified item characteristics (e.g., depth of knowledge 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, subsequently, client review.

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

Statistical Guidelines

In addition to content considerations for constructing test forms, statistical considerations must be considered as well. Item statistics are discussed more in 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 examinees, including representing a variety of difficulty. Specific guidelines include:

- 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. As previously mentioned, items are reviewed for bias during the item development phase, prior to obtaining statistical data. Therefore, it is recommended that statistics not become the sole deciding factor in item use given previous scrutiny during item development.

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; overexposed test content could lead to questions of test validity. Item development activities occur during each year of the assessment, or as stipulated in work scopes. These items are developed and reviewed through activities discussed at the beginning of this chapter. A step in the item development process that has not been mentioned is when the items are "field-tested" or administered to examinees to obtain low-stakes performance data.

Field-test items are items that are administered to examines to obtain performance data, but are not included in students' test scores. These items are administered to obtain data that support their future use as items that contribute to students' test scores. The number of field test forms is determined based on item bank needs and affects the number of responses obtained on field test items. For multiple-choice items, the minimum number of responses per field test item can be a few thousand responses. However, for constructed response items—short answer and extended response—only 1,500 responses are selected and scored for item analysis. The selection of responses is random such that all achievable scores are represented for analysis.

After field-testing, student performance is analyzed and decisions are made regarding the future use of these items. In some cases, the statistics of an item will lead to item reviews that may deem the item inappropriate for future use. For K-PREP, items were field-tested in Reading, Mathematics, and Writing. The next two sections discuss the approaches of field-testing items within these subjects.

When field-test items are included on the test forms, the location of the field-test items is not known to the examinees, thus allowing for maximum effort by the examinees. All item types—multiple choice, short answer, and extended response—are field-tested as needed for maintaining a suitable pool of items for subsequent test forms. Performance data from the field-item items are used during test construction for selecting appropriate test items.

On-Demand Writing

Field-testing for the On-Demand Writing assessment occurred through a *stand-alone field-test* administration. The essay prompts developed for the On-Demand Writing program were administered to Kentucky students in October 2011. Given this unique test administration, a sampling plan was proposed to utilize the minimum population necessary to obtain adequate performance data on each prompt. Unlike Reading and Mathematics, students were aware that the prompts were being field-tested and that their scores would not count toward the academic standing. However, the prompts were administered under live testing conditions, as specified through test administration instructions. Performance data gathered from this test administration were used to select the writing prompts that would be used for the operational test administrations.

Test Booklet Design

For K-PREP, each grade has one test booklet that contains all content areas assessed at that grade (except for Science). For example, third grade test booklet contains Reading and Mathematics only, but the fifth grade test booklet contains Reading, Mathematics, Socials Studies, and Writing. Table 2.13 shows the content areas and order of appearance in the test booklet by grade. In 2018, the Science assessment was administered in a separate test booklet.

Table 2.13 K-PREP Test Booklet by Grade

Grade						
3	4	5	6	7	8	11
R	R	R	R	R	R	ODW
M	M	M	M	M	M	
		SS	ODW		SS	
		ODW			ODW	

R = Reading, M = Mathematics, SS = Social Studies,
ODW = On-Demand Writing

Braille and Large Print Test Materials

Federal and state laws require accessibility of test material for all students. Test material must be developed to accommodate the various needs of students within a testing population. Visually-impaired students participate in the K-PREP assessment program via Braille or large-print versions of the test material. Test forms for these students are modified reproductions of the test form constructed for the general population. For Braille test forms, though, 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 examinees' test scores who use the Braille form.

For K-PREP, items that were not appropriate for Braille were removed from inclusion in the Braille examinees' test scores, thus reducing the maximum number of test points for Braille examinees. As discussed in chapter 7, "Scaling", 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 K-PREP, 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, which stress the importance of adhering to these guidelines. For K-PREP, the *District and Building Assessment Coordinators' Manual (DAC/BAC Manual)* is a manual developed in collaboration between KDE and the testing contractor that outlines administration procedures for before, during, and after the test administration. This chapter will highlight some of the topics presented in the *DAC/BAC Manual* regarding overall test administration procedures including testing dates, student eligibility, and testing accommodations. Also, this chapter will discuss other manuals that are published to guide the administration of K-PREP.

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 K-PREP is based on a district's last day of school, although a general test administration window is specified. Each district is required to administer K-PREP for five consecutive days 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.

Test Make-up Procedures

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

Eligibility Requirements and Exemptions

All students enrolled in grades 3 through 8, and 11 are required to take K-PREP, unless they are participating in the Alternate K-PREP. Participation in K-PREP test administration includes:

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

Students who do not participate in K-PREP include:

- Those participating in Alternate K-PREP
- Those expelled and not receiving academic services
- Foreign exchange students
- Those medically unable to take the assessment
- Those moved out of the Kentucky public school system during testing window
- Those qualifying for an "extraordinary circumstance" exemption (see below).

[2] English learners in their first year must participate in K-PREP Mathematics where tested at their grade.

Students may be exempt from K-PREP based on factors not mentioned above. A medical exemption, for example, can be filed for extenuating medical circumstances. An "extraordinary circumstance" exemption, however, can be filed in the extreme cases of a student not being able to participate in the K-PREP test administration (e.g., parental kidnapping or belonging in protective custody). Appendix A of the *Yearbook* contains a table of participation rates for each content area of K-PREP.

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 (e.g., Individualized Education Program or 504 Plan).

Accommodations and their acceptable use are clearly defined in the manuals published for K-PREP test administration. Below is a list of the accommodations used on K-PREP.

- Use of assistive technology
- Manipulatives
- Readers
- Scribes
- Paraphrasing
- Extended time
- Reinforcement and behavioral modification strategies
- Prompting and cueing
- Interpreters for students with deafness or hearing impairment (signing)
- Simplified language and oral native language support for EL.

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 the testing contractor to develop and provide the training and documentation necessary for K-PREP to be administered under standardized conditions throughout all testing environments.

District Assessment Coordinators

Training for K-PREP test administration is provided to District Assessment Coordinators (DAC) through OAA, Office of Support and Research. 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 K-PREP test administration, including providing test materials and training to the BACs. The DACs also serve as the point of contact for the testing contractor in the case of issues with test materials (e.g., damaged boxes during shipping, additional materials ordering, etc.).

District and Building Assessment Coordinators' Manual

As previously mentioned, the *District and Building Assessment Coordinators' Manual (DAC/BAC Manual)* provides instructions and comments regarding the administration of K-PREP. Included in this manual are instructions for completing the various pre- and post-administration forms as well as instructions for maintaining test security. The assessment coordinators are instructed to read the *DAC/BAC Manual* in preparation for K-PREP test administration.

Test Administrators' Manual

The *Test Administrators' Manual (TAM)* provides much of the same information as the *DAC/BAC Manual*, but also includes explicit directions and scripts to be read aloud to students by test administrators. The *TAM* provides test administrators guidelines on preparing testing environments and the assembly of test materials for returning to the BACs. Given its content and purpose, the *TAM* further promotes the standardization of K-PREP test administration. The assessment coordinators are instructed to read the *TAM* in preparation for K-PREP test administration.

Interpretive Guide

Student performance on K-PREP can be presented in numerous ways. However, it is important to consider how test results should be interpreted and used when compiling data into reports for distribution (see chapter 10, "Validity"). Test results from K-PREP are summarized in various reports from the individual student to the district level. The *K-PREP Interpretive Guide* provides a synopsis of the assessment program and an explanation of some of the score reports that are provided to the schools and districts. The purpose of this guide is to provide guidelines on understanding the reports. A separate, but related document, the *K-PREP Parent Guide* provides a brief description on the performance levels and scale score system used for classifying Kentucky students on achievement.

Test Security

The high-stakes nature of the K-PREP assessment program necessitates the need for test security measures to protect the integrity of the program. Policies for K-PREP test security are outlined in both the *DAC/BAC Manual* and *TAM* and all individuals participating in the administration of K-PREP must adhere to these policies. Adhering to test security policies include reporting any suspicions of security breaches immediately to the appropriate authority, as outlined in the manuals. KDE investigates all allegations of test security breaches.

Receipt and shipping of materials are handled by DACs, using tracking sheets provided by the testing contractor. The *DAC/BAC Manual* provides detailed specifications on inventorying test materials upon arrival and prior to return shipping to the testing contractor. 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 K-PREP 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 administration of K-PREP.

4. Reports

Multiple reports are used to document student performance on the K-PREP assessments. These reports present different levels of summary information about K-PREP and target different audiences. This chapter discusses the various score reports used for K-PREP, including specific pieces of information as well as general cautions on using the reports. Sample reports are provided in Appendix B of the *Yearbook*.

Appropriate Uses for Scores and Reports

The test forms constructed for K-PREP cover a sampling of curriculum content as specified through test blueprints; the tests do not assess all of the possible of content on one test form. Also, the content is assessed through a limited range of item types. Furthermore, the K-PREP 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 be only be interpreted and used in the context from which they are obtained. In other words, K-PREP 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. In addition, academic placement decisions and promotions should not be based solely on K-PREP test scores, but should include other indicators of achievement.

Individual Student Report

The Individual Student Report (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 and will be discussed later in this chapter. The ISR provides the "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.

Kentucky Performance Report

Test scores are also summarized in 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, delineating areas, at a group level, where progress in achievement may be necessary.

Description of Scores

Raw Score

Raw scores are the sum of points from each item within that test. The K-PREP assessments, except for Writing, include a mix of item types that differ in points: multiple-choice items are one point each, short answer items are two points each, and extended response items are four points each. Raw scores can be computed at the domain level (see chapter 2, "Test Development") in addition to the overall test. For Writing, the raw score is the sum of points earned on each writing task.

Scale Score

Scaled scores are derived scores from a statistical transformation of the raw scores. These scores represent a metric that is consistent across test forms and allow for comparisons across test administrations within subject and grade. As discussed in more detail in chapter 7, "Scaling", scaled scores are used to identify the proximity of test performance to established criteria (e.g., passing the test). Scaled scores can also be computed at the domain level to indicate achievement on groups of items – Geometry on a math test, for example. For K-PREP the range of scaled scores is set

100-300 for each test. The range of scaled scores for the domains of each test is also set to 100-300.

Student Performance Level

Student achievement on K-PREP is defined by *performance levels*, within a classification system of achievement from low proficiency to high proficiency. In Kentucky, there are 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 category. Performance level summaries are included on the K-PREP score reports at all levels of reporting—student, school, district, and state. The performance level descriptor, however, is only included on the student report (ISR) since it provides a description of individual student achievement. Chapter 5, "Performance Standards" discusses the performance levels and descriptors and chapter 7, "Scaling", discusses the alignment of scaled scores to the performance levels.

Lexiles and Quantiles

Lexiles are measures used to describe a person's reading proficiency; quantiles are measures used to describe a person's mathematical achievement. These measures also describe the difficulty of content-specific material (e.g., books for reading, or mathematical concepts) so that a person's measure can be used to locate content material at or near the same level of difficulty. The Lexile and Quantile measures are captured in the ISR with instructions on how to use the measures. Chapter 7, "Scaling" provides more information on these measures.

Description of Reports

Student Report

The individual student report (ISR) provides test score information at the student level for each subject test assessed. Scaled scores are reported along with the designated performance level—Novice, Apprentice, Proficient, and Distinguished. As previously mentioned, the performance levels are accompanied with the appropriate performance level descriptor 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 scale score is reported with the corresponding performance level and performance level descriptor. 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. The Lexile and Quantile measures are provided with instructions on how to use them for fostering continued achievement.

School Listing Report

The school listing report provides a list of all students within a particular school along with their test scores: scaled score, performance level, Lexile, and Quantile. This report is created by grade and varies due to the different subject areas assessed within each grade. The school listing report also identifies those students that used test accommodations.

Kentucky Performance Report

The School, District, and State Summary reports provide test score summary information at these three levels of score reporting. These reports provide

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 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—Novice, Apprentice, Proficient, and Distinguished—along with the percentages at the district and state levels. The mean scores by domain ("reporting category") are also presented for the school, in addition to the mean scores 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.

Cautions for Score Interpretations and Use

K-PREP test results can be interpreted in many different ways and used to make inferences about a student, educational program, school, or district. As mentioned earlier in this chapter, these results must be used appropriately to prevent inaccurate interpretations.

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 particular 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 9, "Reliability", provides information on evidence gathered that indicates measurement error on the K-PREP assessments is within an acceptable range.

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, as previously mentioned, test scores are expected to change with multiple testing attempts. As a result, those 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.

Limitations When Comparing Scale Scores at Reporting Group Levels

Test scores of demographic or program groups can be compared within a subject 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 for a particular, but it does not provide all of the information regarding the score distribution. Two groups with similar mean scaled

scores can have different score distributions. Therefore, when viewing group mean test scores, conclusions about the overall distributions cannot be made.

Inappropriateness of Comparing Scale Scores Between Content Tests

Test scores between content tests are not on the same scale and, therefore, should not be compared. As discussed in chapter 8, "Equating", test scores within a particular content test and grade level are placed on the same scale such that scores can be compared across test administrations.[3] The constructs (traits) measured across content tests vary to the extent that the scores cannot be used interchangeably for comparisons.

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 15.4 in the *Standards for Educational and Psychological Testing*, "In program evaluation or policy studies, investigators should complement test results with information from other sources to generate defensible conclusions based on the interpretation of the test results." The Kentucky assessments do 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] The equating of scores began with the 2013 test administration.

5. Performance Standards

In adopting the Kentucky Academic Standards, the Commonwealth of Kentucky began a process of aligning its state educational accountability system toward the goal of measuring students' readiness for post-secondary success. In order to use K-PREP to this end, performance standards were derived that indicate the mastery level needed to be considered "on track" for college and career readiness at pre-secondary levels. This chapter provides a general discussion of determining performance standards for K-PREP Reading, Mathematics, and On-Demand Writing assessments. A separate, and detailed, report of the process is available for interested readers. The final section of this chapter covers the standards determined for Science (2012-2014) and Social Studies.

Performance Level Descriptions and College/Career Readiness

In practice, setting performance standards begins with a set of definitions outlining student achievement requirements at different performance levels. These definitions are often policy- and curriculum-driven, based on grade-specific achievement expectations considered most important by state education agencies. *Performance level descriptors* are the definitions that describe the knowledge and skills necessary to be classified into each performance level defined within an assessment program. In Kentucky, the performance levels of achievement are *Novice*, *Apprentice*, *Proficient*, and *Distinguished*. Given the goal of college and career readiness, the performance level descriptors should include knowledge and skills considered most important for being college/career ready. Extending achievement expectations at the primary grade levels to the idea of college and career readiness, though, is challenging since this level of expectation is not readily accessible for those grades. The K-PREP Reading and Mathematics assessments were aligned to the notion of college and career readiness through a multi-step process of statistical analyses and human judgment. The next section provides a general overview of the steps taken to determine performance standards for Reading and Mathematics.

K-PREP and College/Career Readiness

The expectations of college and career readiness (CCR) are rooted in Kentucky's end-of-course (EOC) assessment program, which uses a modified version of ACT's Quality Core EOC assessments. CCR benchmarks for the EOC assessments were derived from investigations performed by Kentucky's Council on Postsecondary (CPE). These benchmarks established common expectations across high schools, Community and 4-year colleges and were used to determine scores that define students by performance level for each EOC assessment. Applying these scores to Kentucky's ACT Reading and Mathematics test results, HumRRO used the performance level distributions as reference to perform an equipercentile statistical approach to derive cut scores for the K-PREP Reading and Mathematics (grades 3 through 8) assessments. This approach assumed the same proportion of students in each performance level as the ACT referent test, maintaining a degree of correspondence to the EOC exams. More information can be found at Policy Capture for Setting Cut Scores (HumRRO).

The derived cut points were presented to Kentucky educators tasked with creating performance level descriptors using the cut points and test content. Test items were divided into levels—representing the four performance levels previously mentioned—based on the cut points and educators used the groups of items to create

performance level descriptors outlining the knowledge and skills represented by each group. During this process, items may have been viewed as being "misplaced" within a group; for example, an item in the "Apprentice" category may require lower proficiency and, therefore, fit more appropriately with items in the "Novice" category. The educators were provided with guidelines on how items could be shifted across adjacent performance level groups for better fit, but all recommended changes required approval by KDE.

The outcome of this process was a set of performance level descriptors for each grade of the Reading and Mathematics assessments. Additionally, the educators endorsed the cut points through their discussion and creation of the performance level descriptors, including making any recommended adjustments. Once approved by KDE, the performance level descriptors and cut points are used to categorize Kentucky students within the performance levels—Novice, Apprentice, Proficient, and Distinguished. Table 5.1 shows the final theta cut points and impact data (i.e., the percentage of students in each performance level from the 2012 assessments) produced by this approach. For reporting, however, scaled values of the cut points are used to determine the performance levels (see Chapter 7).

Table 5.1 Reading and Mathematics Final Cut Points and Impact Data

Subject	Grade	Theta Cuts			Raw Score Cut Points			Final Impact Data			
		N-A	*A-P*	*P-D*	*N-A*	*A-P*	*P-D*	*N*	*A*	*P*	*D*
Reading	3	-0.0277	0.6911	1.6645	19	25	32	25%	25.6%	32.2%	17.2%
	4	-0.0329	0.7559	1.7576	21	28	35	25%	27.8%	31%	16.2%
	5	-0.0429	0.6559	1.6410	21	27	34	29.4%	23%	31.2%	16.5%
	6	0.1154	0.7865	1.7981	25	32	40	31.3%	22.7%	29.2%	16.9%
	7	-0.0514	0.6286	1.5600	24	31	39	27.1%	25%	31%	16.8%
	8	-0.0362	0.6237	1.5378	24	31	39	28.9%	24.3%	30.1%	16.7%
Mathematics	3	-0.1051	0.9970	2.4321	24	34	43	22.6%	34.6%	34.4%	8.4%
	4	-0.4514	0.5026	1.6434	21	31	42	21.7%	38.7%	29.3%	10.4%
	5	-0.6058	0.4755	1.5902	19	30	40	19.9%	41.1%	27.6%	11.4%
	6	-0.6396	0.4745	1.7376	19	31	43	20.4%	38%	32.1%	9.6%
	7	-0.8555	0.2222	1.5058	16	28	42	22.7%	38.6%	28.7%	9.9%
	8	-0.6391	0.4255	1.7158	18	30	43	20.9%	37.5%	32.2%	9.4%

On-Demand Writing

For On-Demand Writing, KDE chose to use a different process for setting performance standards than was used for Reading and Mathematics. The performance standards for Writing were based on procedures from the Body of Work methodology (Kingston, Kahl, Sweeney, & Bay, 2001) which included a multi-step process of reviewing and rating student work to derive cut points differentiating student writing proficiency in the four performance levels. Educators used student work from the 2012 test and a collection of ancillary material—performance level descriptors and scoring rubric—to form judgments of what level of writing proficiency is necessary to be classified into each performance level. Different from Reading and Mathematics, performance level descriptors for Writing were available for use during this process; the performance level descriptors were crucial in the educators' judgments of writing proficiency.

This process utilized two rounds of judgment in which the educators rated each selected collection of student work to the performance level descriptors – assigning a performance level rating to each collection of work. After the ratings, these judgments were transformed, statistically, into cut points differentiating student

performance into *Novice*, *Apprentice*, *Proficient*, and *Distinguished* categories. The educators were then provided with both the derived cut points, from their ratings, and the actual test scores given by trained scorers. Using this information, the educators compared the cut points with the test scores and discussed if the cut points matched their expectations of student achievement. For example, if the derived cut point for *Proficient* was 10, the educators reviewed the student work that received a test score of 10 and considered if that student work matched the expectations described in the *Proficient* performance level descriptor.

Having two rounds of performance level ratings allowed the educators to share perspectives on their individual ratings and learn perspectives of student achievement expectations; educators may think differently about the student work during the second judgment round, based on what they learned from their peers after the first judgment round. After the second judgment round, though, the educators were provided impact data—the percentage of students in each performance level—based on the derived cut points from the round's judgments. The educators used this data as a "reality check" of their own expectations of student writing.

For the final task of this performance standards process, the educators provided cut score recommendations, having considered the work and feedback data that they reviewed and discussed throughout the process. Reviewing student work was not a planned part of this task, but educators could refer to student work as they considered their recommendations. Tables 5.2 and 5.3 provide the final cut score recommendations and impact data from this process. Note: Cut scores for grades 6 and 10 are presented for historical reference. Chapter 13 describes the scale score metric that will be used for students' ODW scores beginning in 2015.

Table 5.2 ODW Final Performance Level Cut Points[4]

Grade	Performance Level Cut Points		
	Apprentice	*Proficient*	*Distinguished*
5	7	10	14
6	6	9	13
8	7	11	14
10	7	11	14
11	7	10	14

[4] The score range is 0 to 16.

Table 5.3 ODW Final Round Impact Data

Grade	Performance Levels			
	Novice	*Apprentice*	*Proficient*	*Distinguished*
5	19%	49%	30%	2%
6	18%	43%	35%	4%
8	11%	46%	34%	9%
10	12%	46%	36%	7%
11	19%	35%	40%	6%

Science and Social Studies

The K-PREP Science (2012-2015) and Social Studies assessments remained similar in curriculum to the previous assessment program (KCCT). However, some modifications to the test structure (blueprint) in addition to a change in measurement framework lead to a modification of the cut points from KCCT. Standard setting procedures outlined in the previous sections of this chapter were not necessary for Science and Social Studies; instead, the performance level distributions from the 2011 KCCT administration were used to determine cut point for K-PREP.

Table 5.4 shows the percentage of students in each performance level from the 2011 test administration. From scaling procedures—discussed in the next chapter—cut points were found that provided 2012 performance level distributions that were approximately the same as in 2011. Table 5.5 provides the cut points derived using this methodology and the final performance level distributions. Note: Science is included for historical reference.

Table 5.4 2011 Science and Social Studies Performance Level Distribution (KCCT)

Subject	Grade	Performance Level Percentages			
		Novice	*Apprentice*	*Proficient*	*Distinguished*
Science	4	6%	24%	42%	29%
	7	10%	26%	44%	20%
Social Studies	5	11%	29%	44%	16%
	8	10%	30%	41%	19%

**Due to rounding, total percentage may not equal 100.*

Table 5.5 2012 Science and Social Studies Cut Points and Impact Data (K-PREP)

Subject	Grade	Theta Cuts			Performance Level Distribution			
		N/A	A/P	P/D	*N*	*A*	*P*	*D*
Science	4	-0.7197	0.3172	1.4062	6.1%	24.7%	40.5%	28.7%
	7	-0.7215	0.1689	1.4347	10.5%	27.1%	44.5%	17.9%
Social Studies	5	-0.6026	0.4205	1.8593	10.3%	29.6%	45.2%	14.9%
	8	-0.7279	0.4160	1.8512	10.1%	30.7%	40.5%	18.8%

In 2018, new Science assessments were administered in grades 4 and 7. New performance standards, cut scores and performance level descriptors, were set on these assessments following the 2018 test administration. Educators convened to review the content standards and assessments and engaged in a multi-round judgment process that resulted in a set of cut scores that defined student performance within Kentucky's performance levels: Novice, Apprentice, Proficient,

and Distinguished. A separate technical report describes that standard setting process, the recommended cut scores, and performance level descriptors.

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 chapter 2, "Form Development", 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 K-PREP assessments.

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 open-ended (or 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 particular 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.

Item difficulties from the K-PREP assessments are presented in Appendix C the *Yearbook*. The items summarized in these tables are the *operational* items – test items scored and used for determining students' K-PREP achievement. 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. As discussed previously in this report, the K-PREP assessments include a blend of criterion-referenced and norm-referenced content. Some of the norm-referenced content is used with the criterion-referenced content to determine K-PREP test scores. The *Yearbook* includes the multiple-choice item difficulties by *p*-value ranges, including the average *p*-value for all items, for each grade and content area. The *Yearbook* also contains summaries of item difficulty for the constructed response—short answer and extended response—items.

Item-Test Score Correlations

Judging items' appropriateness for testing, however, 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 those students who should have answered the item correctly, based on their total test score, did answer the correctly and those 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 to +1, 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* contains summaries of the item-test score correlations for the multiple-choice constructed response items, including the median correlation across all items, for each grade and content area.

In addition to the correlation between item score and total test score, each answer option of multiple-choice items 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 of the incorrect answer options ("distracters") will be lower than that of the correct answer. In fact, the correlation for the distracters should be less than 0 since students who answer an item incorrectly should have lower test scores than those who answered the item correctly. However, a distracter correlation may be positive (slightly above 0), indicating that even students with higher test scores chose that wrong answer. Positive correlations for item distracters may indicate something systematically causing students to choose the incorrect answer option. In this case, the item's content and answer option should be reviewed.

Differential Item Functioning

During item development, items are reviewed for potential bias against any student subgroup (e.g., gender, ethnicity, disability, etc.). Items that are identified as displaying potential bias are either revised or removed from consideration for future use. Once items have been field-tested, though, 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.

Differential item functioning (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" while 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, while Males are the reference group; in ethnicity DIF analyses, African-Americans are a focal group, while Whites are the reference group. DIF analyses on ethnicity can be extended to other ethnic groups to represent the focal group—and comparing them each to Whites. Since students are matched on proficiency across focal and reference groups, statistical differences found between the groups are not confounded by student proficiency.

There are multiple statistical procedures for analyzing DIF, one of which is based on the Mantel-Haenszel chi-square statistic (M-H χ^2) for multiple-choice items (Holland and 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-by-2-by-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 2x2 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 that are designated with 'B' or 'C' DIF classifications are recommended for review before continued use on assessments. However, caution must be exercised when analyzing DIF to prevent over-interpretation of the statistics.

The *standardized mean difference* (SMD: Zwick, Donoghue, and Grima, 1993) procedure is also used for detecting DIF; for K-PREP this statistic is used on constructed response items. SMD is a summary statistic 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. As previously mentioned, caution must be exercised when analyzing DIF to prevent over-interpretation of the statistics.

Appendix E of the *Yearbook* provides the number of operational and field-test 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 or not the item content lends itself to differential item functioning. 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.

Item Response Theory

Item Response Theory (IRT) is a measurement framework that analyzes test item properties and item responses simultaneously. IRT has become the focal point in large-scale assessment, surpassing *classical test theory*, its predecessor. Measurement models under IRT specify the probability of a correct response to an item dependent upon proficiency and item characteristics. While discussed as an overview in this report, readers interested in IRT and its models should seek the multitude of books on this topic. The relevance if mentioning IRT here is that one fundamental aspect of the framework is the difficulty of test items.

The simplest IRT model is the *one-parameter logistic* (1PL; Rasch, 1980) measurement model, represented as:

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

where $P_i(\theta)$ is the probability that a person 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 provides a graphical display of the 1PL model for an item.

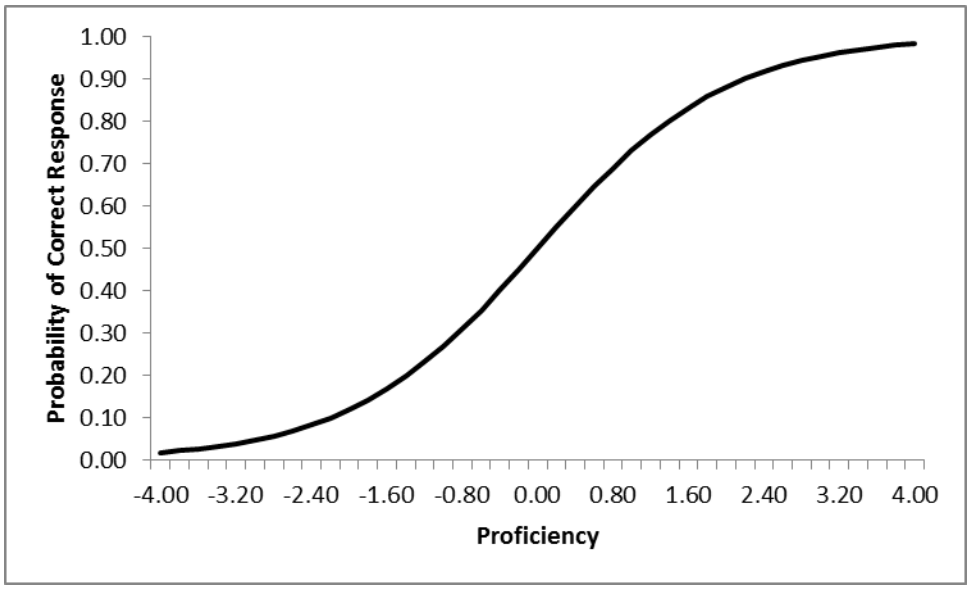


Figure 6.1 Graph of 1PL Model

However, this model applies to multiple-choice items only. Given that K-PREP 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 at the beginning of this chapter. 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 person 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 and PCM is that the PCM has multiple difficulties associated with an item as opposed to the single item difficulty in the 1PL. However, the difficulties in PCM represent the difficulty in transitions from one score category to the next. For an item with three score categories—0 to 2 points, for example—there would be two transitions ("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 the partial credit model for a three-category 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.

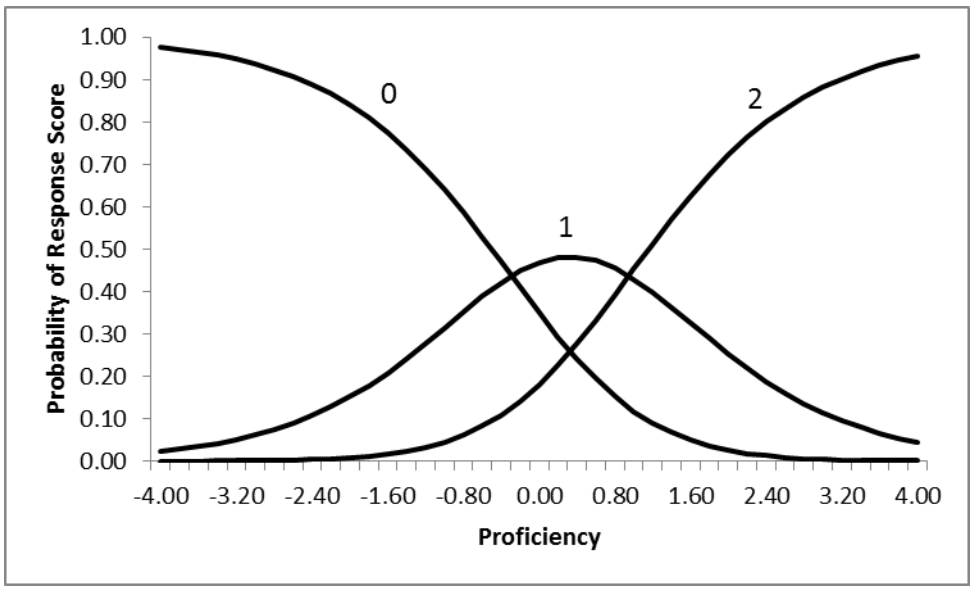


Figure 6.2 Graph of Partial Credit Model for Three-point Item

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 K-PREP, it is how well the 1PL and partial credit models 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 K-PREP is investigated through *mean-square* fit statistics which provide evidence on how well the pattern of observed responses are predicted by

measurement models, 1PL and partial credit model. *Outfit* mean-square statistics are influenced by unexpected response patterns to items far from a person's proficiency measure. *Infit* mean-square statistics are influenced by unexpected response patterns to items near a person's proficiency measure. Linacre (2011a) provides a classification of fit mean-square estimates useful for interpretation (see 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. From this figure, 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.

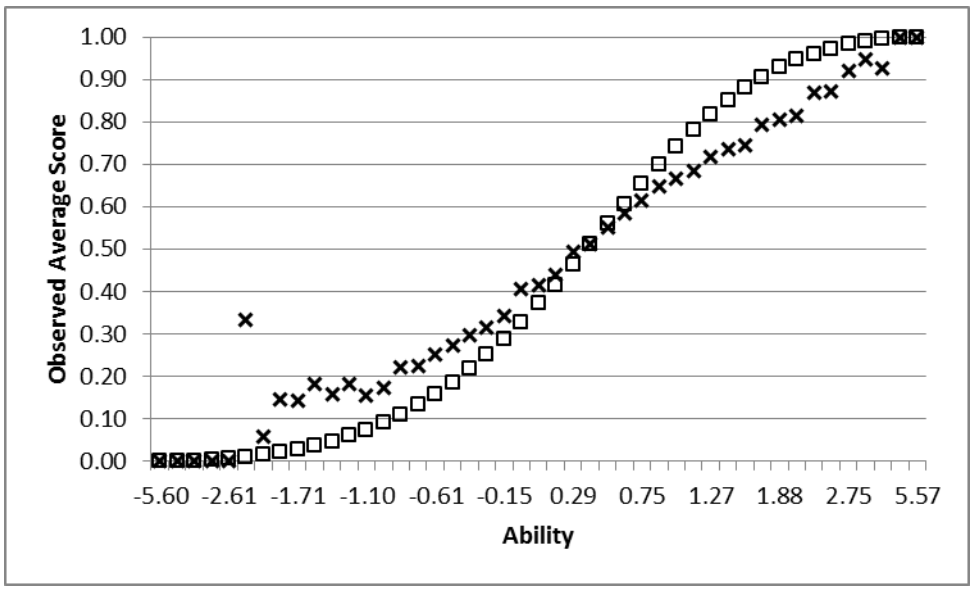


Figure 6.3 Observed and Expected Performance on Item of Average Difficulty

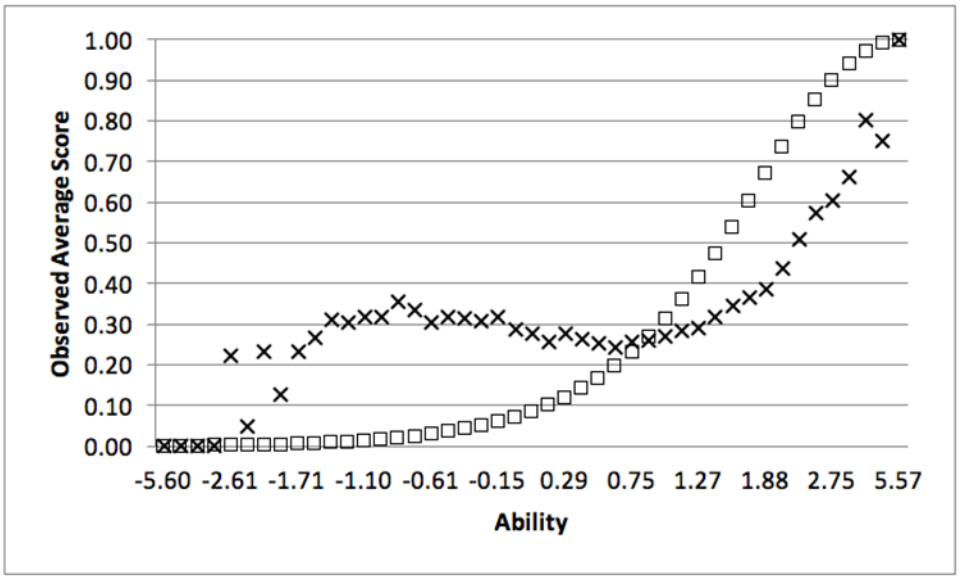


Figure 6.4 Observed and Expected Performance on Difficult Item

The IRT parameter estimates—item difficulty and item fit—are summarized in Appendix F of the *Yearbook*.

On-Demand Writing Item Analysis

Essay prompts were field-tested in 2011 for the On-Demand Writing assessment program to gather student performance data on a variety of writing tasks. These tasks included passage-based and stand-alone stimuli and covered several modes of writing: argumentative, narrative, opinion, and informatory/exploratory. Twenty-four prompts were administered per grade and a sampling plan was designed to select a testing sample that reflected the student population of Kentucky. After the prompts were administered, student performance was analyzed in multiple ways.

- *Mean total scores*: Overall mean total scores and mean total scores by student subgroups (e.g., gender, ethnicity, and Limited English Proficiency).
- *Score point frequencies*: Overall percentages of each total score point and frequency counts of invalid scores (e.g., blank, off-topic, etc.).
- Scorer agreement: Each student response was double-scored, allowing for indices of 'perfect', 'adjacent', and 'non-adjacent' agreement to be computed.[5]

These computations provided a context for determining which prompts should be used for live testing, and subsequently for providing appropriate information about student writing in Kentucky.

Beginning in 2015, IRT was implemented in order to establish a stable reporting scale and equate scores over subsequent test forms. As a result, student scores on the writing prompts were calibrated according to the Rasch Partial Credit model and the prompt IRT estimates were used to generate overall scores. Chapter 13 describes the implementation of IRT into the analysis of On-Demand Writing.

[5] Adjacent scores occur when a student responses receives two scores that differ by one point; non-adjacent occurs when the difference is more than one point.

7.Scaling[6]

Rationale

Total test scores for examinees 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 an examinee's performance on a test. Also, examinees can be rank ordered based on their test performance using the raw scores and group statistics can be computed (i.e., average, standard deviation, etc.) and interpreted. However, raw scores can be limiting for comparisons 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, to name a couple of reasons. 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 exactly the same level of difficulty as other test forms of similar content and therefore 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 *scaling* process of raw score transformations; the next chapter, "Equating", discusses more aspects of adjusting difficulty difference between test forms.

Measurement Models

The Rasch and Partial Credit models were introduced in chapter 6, "Item Analyses", 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 examinee based on their response patterns and the simultaneous estimation of the item parameters. As mentioned in the previous chapter, 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, then there will be 41 person proficiency estimates, one for each raw score including zero. The proficiency estimates will also increase from the lowest to highest value in relation to the ascending order of the raw scores.

It should be noted that 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 an examinee 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.

[6] For 2017-2018, scaling was conducted for Reading and Mathematics.

As will be demonstrated later in this chapter, the proficiency estimates are used to transform examinees' test scores into a metric that can be used to compare performance across test forms.

Process

This section outlines the process by which the K-PREP assessments were scaled according to the IRT models previously discussed. While the following description is an overview of the process, some level of detail is required so that the reader can gain an understanding of how reportable scores are derived from examinee test responses.

Overview

Pearson performed item calibrations to obtain the Rasch item parameters and proficiency estimates for the K-PREP assessments. HumRRO performed an independent execution of the analyses as a third-party verifier of the process and results. Pearson created analysis specifications ("Calibration and Equating Specifications") that outlined, in detail, the process and methodology for scaling the K-PREP assessments. These specifications included 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 further.

The scaling process utilized approximately the entire testing population of K-PREP; exclusion rules were applied to remove examinees that 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 any other testing accommodation made available for K-PREP. All students participate in K-PREP using the same test form of operational items, regardless of testing accommodation. In the case of Braille examinees, however, some test items are considered not appropriate for Braille reproduction and, therefore, are removed from administration and scoring for these examinees. As a result, separate analyses may be conducted for Braille examinees due to the difference in maximum test score.

Prior to scaling, examinee data is inspected primarily to identify any items that potentially may have been scored incorrectly. In other words, 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 is dependent upon accurately scored examinee data and all items must be considered to have been properly scored prior to analysis.

Examinee response data is analyzed through Winsteps Version 3.73 (Linacre, 2011), a Rasch modeling statistical software. Each K-PREP assessment is analyzed separately through this software; the operational items for each subject/grade test is analyzed first, followed by the field-test items (discussed in the next chapter). As previously mentioned, the output from this process includes item parameters ("difficulty") and proficiency estimates both on a logit scale. Discussed in more detail later in this chapter, the proficiency estimates are used to derive *scaled scores* for performance comparisons across test forms.

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 necessary files—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. *(As part of its internal processes Pearson utilized two independent replications of the analysis.)* In addition to feedback throughout the analysis, Pearson, HumRRO, and KDE participated in a conference call each day during the analysis to share general impressions and discuss any concerns with the current results. To utilize the daily conference call effectively, Pearson proposed a schedule of analysis such that Pearson and HumRRO would perform the same analyses concurrently and 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.

Scaled Scores

Transformation of Raw Scores

This chapter has been devoted to setting the foundation for scaled scores – scores derived from raw scores to a metric usable for communicating and interpreting examinee performance. Scaled scores can be derived through either linear or nonlinear transformations of the raw scores. For K-PREP, the scaled scores are derived through linear transformations 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 examinee response data. In 2015, scaled scores for the Writing test were computed for the first time in K-PREP. Chapter 13, *On Demand Writing*, describes the process by which scaled scores were determined for this test. The remaining portions of the current chapter, however, describe scaled scores for Reading, Mathematics, and Social Studies.

The scaled score metric for K-PREP was chosen to range from 100 to 300, for each subject, with 210 representing the minimum scaled score for passing *(Proficient)*. To achieve this score metric, the following linear transformation was proposed:

$$SS = m(\theta - \theta_p) + b,$$

where the slope (m) was set to 16.67, the intercept (b)[7] was set to 210, and θ is the person proficiency estimate defined as before. This transformation, however, includes θ_p, the person proficiency estimate identified as the minimum value for

[7] In this context, b should not be confused with b_i used as item difficulty in the IRT models.

Proficient. This term was included in the transformation so that the proposed minimum scaled score for passing (210) would always exist. Therefore, the value of 210 has the same meaning regardless of which form is taken. The values used for this term are provided below in Table 7.1. The derived scaled scores are discussed more in the remaining sections of this chapter.

For Reading and Mathematics, the values for θ_p were determined during standard setting meetings (see chapter 5, "Performance Standards"); for Science, these determined during standard settings in 2018, as documented in a separate report; and for Social Studies, these values were determined by using 2011 KCCT performance data and finding similar performance patterns in the 2012 K-PREP test data. The determination of criterion values indicating performance standards is discussed in Chapter 5, "Performance Standards."

Table 7.1 Proficient Cut Points for Derived Scaled Scores

Subject	Grade	θ_p
Reading	3	0.6911
	4	0.7559
	5	0.6559
	6	0.7865
	7	0.6286
	8	0.6237
Mathematics	3	0.9970
	4	0.5026
	5	0.4755
	6	0.4745
	7	0.2222
	8	0.4255
Science	4	0.2775
	7	-0.0138
Social Studies	5	0.4205
	8	0.4160

Scaled scores for each reporting category (domains outlined in Chapter 2, "Test Development") ("subdomain") of each content area were computed to help illustrate students' specific strengths and weaknesses. In 2012, to prevent confusion with the total test scale scores, a different scale was used for computing scale scores of the reporting categories. The score transformation to achieve this scale was

$$SS = 1.667 * (\theta - \theta_p) + 21,$$

where θ is the IRT person proficiency estimate, from total test performance, and θ_p is the minimum person proficiency estimate for *Proficient*, as determined through standard setting. The slope and intercept—1.667 and 21, respectively—were adapted from the scale score transformation of the total test scores.

In addition to individual student performance information, the scale scores for the reporting categories were 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 reporting category. For example, school, district, and state scale score averages could be compared for the *Key Ideas* subdomain in Reading and for the *Geography* subdomain in Social Studies. Summary reports contained these comparisons for all reporting categories in Reading, Mathematics, and Social Studies.

This transformation of subdomain scores, however, presented a challenge for interpretations because the scale became too small to detect differences in the mean score values. In other words, the mean values were so close between school, district, and state summary levels, that users could not discern meaningful differences. In addition, users experienced difficulty in determining strengths and weaknesses at a summary level. This challenge from the 2012 score reports led KDE and the testing contractor to modify the scale score transformation for domain reporting categories as well as the intended interpretations, for future score reporting.

Beginning in 2013, the reporting category scale scores were derived using the same transformation as the scale scores for the total test,

$$SS = 16.67 * (\theta - \theta_p) + 210,$$

to achieve a scale with 100 and 300 as the minimum and maximum scale score values, respectively. The scale scores are aggregated by schools, districts, and state, but instead of implying comparisons between these aggregate levels, a criterion was imposed for determining strengths and weaknesses. Specifically, the average scale scores by reporting category for each aggregate level are compared to the *Proficient* standard (or '210' scale score). Average scale scores at or above 210 indicate that students at that particular aggregate level are "on-track" for mastering the concepts within a particular subdomain; scores below '210' indicate that improvement is needed for a particular subdomain. The goal of this modification in score reporting is to help school and district administrators identify areas for improvement in student achievement.

Although the transformation of raw scores to scale scores was modified after 2012, the interpretability of the scale scores remains the same. The meaning of the scale scores is only affected by the threshold(s) and threshold definitions attached to the overall scale.

Considerations and Limitations

There are limitations on using scaled scores for interpreting examinee performance. First, the scaled scores are not on a *vertical scale*, which limits interpretations on performance differences on a subject 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 5 points can be the result of a small difference in proficiency estimate. Also, differences in scaled scores within a test vary along scale. For example, in table 7.2, scaled scores near the middle of the scale—for raw scores 23 through 27—will have a smaller difference than the lowest or highest scaled scores—for raw scores 38 through 40, for example.

Table 7.2 Raw Score to Scale Score Conversion

Raw Score	Scale Score
0	100
1	132
2	144
3	152
4	157
.	.
.	.
23	206
24	208
25	210
26	212
27	214
.	.
.	.
37	246
38	253
39	265
40	289

The scaled score system was created to indicate the proximity of examinee performance in line with the state performance standards (see Chapter 5). The scaled scores align to definitions of achievement—performance levels (see Table 7.3). As mentioned in Chapter 5, the scale scores presented in Table 7.3 are used to differentiate student performance levels for reporting. The 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. The scaled scores for the reporting categories, however, are further restricted in use and interpretation. These scaled scores are not aligned to the performance levels, but provide supplementary information on within-subject achievement.

Table 7.3 Scaled Scores by Performance Level

Subject	Grade	Novice	Apprentice	Proficient	Distinguished
Reading	3	100-197	198-209	210-225	226-300
	4	100-196	197-209	210-226	227-300
	5	100-197	198-209	210-225	226-300
	6	100-198	199-209	210-226	227-300
	7	100-198	199-209	210-225	226-300
	8	100-198	199-209	210-224	225-300
Mathematics	3	100-191	192-209	210-233	234-300
	4	100-193	194-209	210-228	229-300
	5	100-191	192-209	210-228	229-300
	6	100-190	191-209	210-230	231-300
	7	100-191	192-209	210-230	231-300
	8	100-191	192-209	210-231	232-300
Science	4	100-190	191-209	210-225	226-300
	7	100-191	192-209	210-228	229-300
Social Studies	5	100-192	193-209	210-233	234-300
	8	100-190	191-209	210-233	234-300

Results

Appendix G of the *Yearbook* contains the tables of derived scaled scores for each K-PREP assessment. Each table contains the raw scores, proficiency estimates ("theta"), scaled scores, and conditional standard error of measurement. The conditional standard error of measurement represents the standard deviation of observed scores of students with the same true score and as discussed more in Chapter 9, "Reliability."

Descriptive statistics—mean, standard deviation, minimum, maximum—for the scale scores for each K-PREP assessment are provided in Appendix K of the *Yearbook*. The descriptive statistics are provided for the overall testing population, as well as by subgroups— gender, ethnicity, free/reduced lunch status, and accommodations. Scaled score frequency distributions for each K-PREP assessment are provided in Appendix I of the *Yearbook*. Appendix L of the *Yearbook* contains tables of performance level distributions for each K-PREP assessment.

Lexiles and Quantiles

For K-PREP Reading and Mathematics, examinee performance is aligned to external indicators of reading and math fluency. *Lexiles*® are measures that indicate a person's reading proficiency or the reading difficulty level of a book or other piece of text. Regardless of the object—person or text—the person Lexile measure can be directly compared to the Lexile measure of text. Knowing both a person's and a book's Lexile measure, for example, one can predict how well that person will understand that book. *Quantiles*®, on the other hand, indicate how well one understands the mathematical concepts at his/her grade level. Similar to Lexiles, Quantiles are applied to both person's mathematical proficiency and the difficulty of mathematical concepts. In Lexile and Quantile frameworks, the higher measure a person receives, the higher proficiency that person exhibits.

MetaMetrics® provided scaling transformations to derive student Lexile and Quantile measures based on K-PREP test performance. Although the results of those transformations are not presented in this report, it is important to mention this unique scaling application of K-PREP performance.

8. Equating[8]

Rationale

In large scale assessment programs, multiple test forms are created that reflect similar content and difficulty. These forms can be used for different testing administrations (i.e., years) or within the same testing administration but on different subsets of the testing population. Regardless of when the forms are used, they are constructed such that performance across forms can be directly compared. However, no two test forms will have the exact same level of difficulty, which confounds the comparison of performance across 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.

Process

Equating test forms can be accomplished in many different 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 through the use of a set of common items included on both forms. The testing populations are considered nonequivalent as they do not consist of the same examinees taking both forms. The equating result is a scale transformation that accounts for differences in difficulty across two (or more) test forms. The end result is that scores from both test forms exist on a single scale. The rest of this section describes the equating process for the K-PREP assessments, as conducted by the testing contractor.

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 K-PREP, the linking items are *internal* in that they are treated as operational items contributing to students' test score. 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 (mini version) of the overall test. For the 2018 tests, linking items were chosen to best represent the range of item difficulty while adhering the content distribution of the blueprint.

Since K-PREP assessments include constructed response item types on each form, the linking sets included those item types. For Reading and Mathematics, only the short answer constructed response item type was included in the linking set even though extended response item types were administered as well. In the Reading and Mathematics test designs, each test form contains at least two short answer items, while there is only one operational extended response item administered. In this design, utilizing the extended response item in the linking set may present an exposure concern as well as put limits on the selection of other operational items during the forms development process. For Social Studies, there are multiple

[8] For 2017-2018, equating was conducted for Reading and Mathematics.

extended response items administered, but no short answer items. In this case, the extended response item type is represented in the linking sets. Table 8.1 provides the distribution of the linking items by item type selected for the Reading and Mathematics tests. In some cases, only multiple-choice items were selected due to availability within the item pools. Tables 8.2 and 8.3 provide the linking items by test blueprint distribution.

Table 8.1 **2018 Linking Items by Item Type**[9]

Subject	**Grade**	**Item Type**		
		Multiple-choice	*Short Answer*	*Extended Response*
Reading	3	10	1	--
	4	10	1	
	5	13	--	
	6	15	--	
	7	13	1	
	8	14	--	
Mathematics	3	15	--	--
	4	14	1	
	5	13	1	
	6	14	1	
	7	15	--	
	8	15	--	

[9] For 2018, equating, and the use of linking items, was conducted for the Reading and Mathematics assessments.

Table 8.2 **2018 Reading Linking Items by Test Blueprint**

Grade(s)	Domain	Linking Items (Points)			Domain Coverage
		MC	*SA*	*ER*	
3	Key Ideas	2	0	NA	17%
	Craft and Structure	2	2		33%
	Integration of Ideas	3	0		25%
	Vocabulary and Acquisition	3	0		25%
4	Key Ideas	1	2	NA	25%
	Craft and Structure	2	0		17%
	Integration of Ideas	4	0		33%
	Vocabulary and Acquisition	3	0		25%
5	Key Ideas	3	0	NA	23%
	Craft and Structure	3	0		23%
	Integration of Ideas	4	0		31%
	Vocabulary and Acquisition	3	0		23%
6	Key Ideas	3	0	NA	20%
	Craft and Structure	4	0		27%
	Integration of Ideas	5	0		33%
	Vocabulary and Acquisition	3	0		20%
7	Key Ideas	4	0	NA	27%
	Craft and Structure	2	2		27%
	Integration of Ideas	4	0		27%
	Vocabulary and Acquisition	3	0		20%
8	Key Ideas	3	0	NA	21%
	Craft and Structure	3	0		21%
	Integration of Ideas	4	0		29%
	Vocabulary and Acquisition	4	0		29%

Table 8.3 **2018 Mathematics Linking Items by Test Blueprint**

Grade(s)	Domain	Linking Items (Points)			Domain Coverage
		MC	*SA*	*ER*	
3	Operations and Algebraic Thinking	3	0	NA	20%
	Number and Operations in Base Ten	4	0		27%
	Number and Operations – Fractions	4	0		27%
	Measurement and Data, Geometry	4	0		27%
4	Operations and Algebraic Thinking	4	0	NA	25%
	Number and Operations in Base Ten	3	0		19%
	Number and Operations – Fractions	4	0		25%
	Measurement and Data, Geometry	3	2		31%
5	Operations and Algebraic Thinking	3	0	NA	20%
	Number and Operations in Base Ten	2	2		27%
	Number and Operations – Fractions	3	0		20%
	Measurement and Data, Geometry	5	0		33%
6	Ratios and Proportional Relationships	3	0	NA	19%
	The Number System	3	0		19%
	Expressions and Equations	3	0		19%
	Geometry	3	0		19%
	Statistics and Probability	2	2		25%
7	Ratios and Proportional Relationships	3	0	NA	20%
	The Number System	3	0		20%
	Expressions and Equations	3	0		20%
	Geometry	3	0		20%
	Statistics and Probability	3	0		20%
8	The Number System and Expressions & Equations	4	0	NA	27%
	Functions	4	0		27%
	Geometry	4	0		27%
	Statistics and Probability	3	0		20%

Analysis

The equating analysis was performed by the testing contractor and an independent contractor of KDE, using analysis specifications created and maintained by the testing contractor. Four process checkpoints were implemented for verification across the independent replications:

- Initial calibration item parameters
- Robust Z statistics for linking item analysis
- Final (equated) item parameters
- Raw-score-to-scale-score ("RS-SS") conversion tables

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

1. Calibrate the items through Winsteps (Linacre, 2011b) software using student item response data.
2. Perform item stability analysis of linking items using Robust z statistical methodology (Huynh, 2000; Huynh & Rawls, 2009; Huynh & Meyer, 2010)—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. Produce score conversion tables, including scale 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, then those items are not appropriate for equating one test form onto the other. Each linking set is tested through this procedure. Although items may be considered unstable for equating, internal linking items remain as scored items for students' test score. The majority of linking items, across all grades, are considered stable for equating.

Table 8.4 Unstable Linking Items

Subject	Grade(s)	Number of Linking Items Dropped		
		MC	SA	ER
Reading	3	0	0	--
	4	0	0	--
	5	0	--	--
	6	1	--	--
	7	0	1	--
	8	0	--	--
Mathematics	3	0	--	--
	4	1	0	--
	5	0	1	--
	6	0	1	--
	7	0	--	--
	8	1	--	--

After dropping the linking items that are considered unstable for equating, the remaining linking items are used to produce equated parameter estimates of non-linking items. These item parameter estimates are produced through item calibration with Winsteps, similar to the initial step of the analysis, but with the linking items used as an anchor scale.

Field-test Item Calibration

When necessary, new items are included on test forms to gather student performance data while not contributing to examinees' test scores. These *field-test items* are administered so that they can be used toward examinees' test scores on a future test form. During the item analyses, the field-test items are placed on the same measurement scale as the operational items via Winsteps, using the operational items as the base scale. This process requires two steps: 1) calibrate the operational items via Winsteps, and 2) calibrate the field-test items via Winsteps, but specify the operational items—their item parameters—as the base. Through this process, the field-test items are added to the calibrated item pool and will be used for future form development analyses through IRT.

9. Reliability

Reliability is the consistency of the results obtained from a measurement. When a score is reported for a student, there is an expectation 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 (that is, 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 (that is, showing evidence of valid use of the results). However, a reliable test score is not necessarily a valid one; and a reliable test score is not valid for every purpose. A measure can be consistent and support certain score interpretations but still not support all the inferences a user of the test wishes to make. The concept of test validity is discussed in chapter 10, "Validity."

Definition of Reliability

The basis for developing a mathematical definition of reliability can be found by examining the fundamental principle at the heart of classical test theory: All measures consist of an accurate or "true" part and an inaccurate or "error" component. This is commonly expressed as,

$$Observed\ Score = True\ Score + Error.$$

Errors occur as a natural part of the measurement process and can never be eliminated entirely. For example, uncontrollable factors such as differences in the physical world and changes in examinee disposition may work to increase error and decrease reliability. In classical test theory, error is typically assumed to be the result of random, unsystematic influences. If there are systematic influences contributing to the error term, then derived reliability indices are likely to be compromised. For example, if a test is administered under very poor lighting conditions, the results of the test are likely to be biased against the entire group of students taking the test under the adverse conditions.

From the equation above, it is apparent that scores from a reliable test generally have little error and vary primarily because of true score differences. One way to consider reliability is to define reliability as the proportion of true score variance relative to observed score variance:

$$Reliability = \frac{\sigma_T^2}{\sigma_O^2} = \frac{\sigma_T^2}{\sigma_T^2 + \sigma_E^2} = 1 - \frac{\sigma_E^2}{\sigma_O^2},$$

where σ_T^2 is the true score variance, σ_O^2 is the observed score variance, and σ_E^2 is the error variance. When there is no error, the reliability is the true score variance divided by true score variance, which is unity. However, as more error influences the measure, the error component in the denominator of the ratio increases and the reliability decreases.

Using classical test theory, an alternative formulation can be derived. Reliability (the ratio of true variance to observed variance) can be shown to equal the correlation coefficient between observed scores on two *parallel* tests. The term parallel has a specific meaning: The two tests meet the standard classical test theory assumptions, as well as yielding equivalent true scores and error variances. The proportion of true

variance formulation and the parallel test correlation formulation can be used to derive sample reliability estimates.

Estimating Reliability

There are a number of different approaches available to estimate reliability of test scores. Discussed below are test-retest, alternate forms, and internal consistency methods.

Test-Retest Reliability Estimation

Reliability can be estimated by calculating the correlation coefficient between scores from a test given on one occasion with scores from the same test given on another occasion to the same students. Essentially, the test is acting as its own parallel form. Using the test-retest reliability method has potential pitfalls. A long interval between testing sessions likely will result in student growth in knowledge of the subject matter, while a short interval increases the chance students will remember and repeat answers from the first session. In addition, the test-retest approach requires students to take the same test twice. In Kentucky, students do not take the same test twice under any circumstances; therefore, test-retest reliability estimation is not used on the Kentucky assessment.

Alternate Forms Reliability Estimation

Alternate forms reliability is similar to test-retest, except that instead of repeating the same test, two presumably equivalent forms of the test are administered to the same students. The accuracy of the alternate forms coefficient greatly depends upon the degree to which the two forms are equivalent. Ideally, the forms would be parallel in the sense given previously. For Kentucky assessment, alternate forms reliability estimation is not possible because no student takes more than one form of the test during any test administration.

Internal Consistency Reliability Estimation

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 requiring multiple tests. Probably the most frequently used internal consistency reliability estimate is the coefficient alpha (Cronbach, 1951). Coefficient alpha is based on the assumption that inter-item covariance constitutes true-score variance and the fact that the average true score variance of items is greater than or equal to the average inter-item covariance. The formula for the coefficient alpha is

$$\alpha = \left(\frac{N}{N-1}\right)\left(1 - \frac{\sum_{i=1}^{N} s_{Y_i}^2}{s_X^2}\right),$$

where N is the number of items on the test, $s_{Y_i}^2$ is the sample variance of the i^{th} item (or component) and s_X^2 is the observed score sample variance for the test.

Coefficient alpha estimates for each overall test and by item type—multiple-choice and constructed response—are provided for each grade and subject in Appendix M of the *Yearbook*. These reliability estimates are provided the overall testing population

as well as by gender, ethnicity, and other student breakout groups. In addition, coefficient alpha estimates are provided, for each major subscale (*see Domain Reliability Estimation*).

Domain Reliability Estimation

The Kentucky assessment consists of item clusters that divide content areas into domains *(refer to chapter 2)*. Scores are provided for the domains, in addition to the total score for the content areas. Reliability at the domain level, though, will generally be lower than total score reliability because reliability is influenced by the number of items (as well as their covariance). In some cases, the number of score points associated with a domain score is small (ten or fewer). Results involving domain scores must be interpreted carefully, as in some cases these measures have low reliability due to the limited number of points attached to the score.

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 (unreliability). The SEM is an estimate of how much error there is likely to be in an individual's observed score, or alternately, how much score variation would be expected if the individual 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 (standard deviation of the raw scores) and $\rho_{xx'}$ is a reliability estimate for the set of test scores.

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 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 (that is, 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 2 out of 3 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. By increasing the range of the confidence interval, one improves the likelihood the confidence interval includes the true score; 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.

The SEM is reported for Kentucky assessment in the *Yearbook* in the reliability tables (Appendix M). The SEM is reported for total scores and domain scores for the overall testing population, gender, ethnicity, and other student breakout groups.

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. To formalize this notion, classical test theory postulates that every student has a true score. This is the score the student would receive on the test if no error were present. The SEM for a particular true score is defined as the standard deviation of the observed scores of students with that true score. 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, then 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. Therefore, the CSEM cannot be calculated simply by grouping students by their true score and computing the conditional standard deviation. However, item response theory (IRT) allows for 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

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

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.

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

Scoring Reliability for Open-Ended Items

Reader Agreement

Kentucky's testing contractor uses several procedures to monitor scoring reliability. One measure of scoring reliability is the between-reader agreement observed in the required second reading of 1) all On-Demand Writing test responses and 2) a percentage of students' short-answer and extended-response item responses for Reading, Mathematics, Science, and Social Studies. These data are monitored on a

daily basis by Kentucky's testing contractor during the scoring process. Reader agreement data show the percent perfect agreement of each reader against all other readers.

Reader agreement data 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 Kentucky's testing contractor 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 Kentucky's testing contractor's scoring directors, and appropriate actions are initiated as needed, including the retraining or termination of readers.

Appendix N in the *Yearbook* provides scoring metrics—reliability, validity, and score distributions—for constructed response items across content areas. As mentioned above, 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 11, "Performance Scoring."

Score Resolutions

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

Reliability of Performance Level Categorization

Every test administration will result in some error in classifying examinees. 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 achievement levels. For example, some students may have a true achievement 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 achievement level. As discussed in the section on the SEM, a student's true score is most likely to fall into a standard error band around his or her observed score. Thus, the classification of students into different achievement levels can be imperfect; especially for the borderline students whose true scores lie close to achievement level cut scores.

For the Kentucky assessment, the levels of achievement are *Novice*, *Apprentice, Proficient,* and *Distinguished*. A description and analysis of classification *accuracy* and *consistency* indices is provided below.

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— "true scores". Since true scores are not available, an estimate of the true score distribution must be determined in order 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. In Kentucky, however, this is impractical. 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 K-PREP assessments.

Calculating Accuracy

To calculate accuracy, a 4 x 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 score fall into performance level x and whose observed scores fall into performance level y. Table 9.1 is an example accuracy table where the columns represent test-based student achievement and the rows represent true achievement level decisions. In this example, the total accuracy is approximately 75%, the sum of the diagonal (shaded) cells.

Table 9.1 Example Accuracy Classification Table

True Score	Observed Score				Total
	Novice	Apprentice	Proficient	Distinguished	
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* versus *Proficient* or *Distinguished* because Kentucky uses *Proficient* and above as proficiency for Adequate Yearly Progress (AYP) decision purposes as well as for an index 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 9.2 indicated classification accuracy around the Proficient cut point of approximately 90%. The percentage of examinees incorrectly classified as *Apprentice* or lower, when their true score indicates *Proficient* or above, is approximately 3%.

Table 9.2 Example Accuracy Classification Table for Proficient Cutpoint

True Score	Observed Score				Total
	Novice	Apprentice	Proficient	Distinguished	
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

Calculating Consistency

Consistency can be calculated in the same manner, via 4 x 4 contingency table, albeit with data indicating an estimate of the joint distribution of classifications on (hypothetically) two independent, parallel test forms. Table 9.3 shows sample statistics of consistency classification. Based on this sample data, the overall consistency is approximately 67%. The consistency at *Proficient* is 87%. The agreement rates are lower than those for accuracy because both classifications contain measurement error; whereas, in the accuracy table, true score classification is assumed to be without error.

Table 9.3 Example Consistency Classification Table

First Form	Second Form				Total
	Novice	Apprentice	Proficient	Distinguished	
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

Calculating Kappa

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

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

where P is the proportion of consistent classifications and P_c is the proportion of consistent classification by chance. Using Table 9.3, 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 9.3 is 0.548.

Appendix P of the *Yearbook* contains tables of classification accuracy and consistency indices – including kappa coefficients—overall performance level classification and at the Proficient cut point for each grade and subject.

10. Validity

Validation is the process of collecting evidence to support inferences from assessment results. A prime consideration in validating a test is determining if the test measures what it purports to measure. During the process of evaluating if the test measures the construct of interest, a number of 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, the test content may not span the entire range of the construct to be measured, etc. Any of these threats to validity could compromise the interpretation of test scores.

Beyond verifying 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 chapter 4, "Reports" (in the section "Cautions for Score Interpretation and Use") and chapter 7, "Scaling" (in the section "Scaled Scores: Limitations of Interpretations").

Demonstrating that a test measures what it is intended to measure and 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).

Argument-Based Approach to Validity

The fifth edition *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association and the National Council on Measurement in Education, 1999) recommends establishing the validity of a test through the use of a *validity argument*. This term is defined in the *Standards* as "An explicit scientific justification of the degree to which accumulated evidence and theory support the proposed interpretation(s) of test scores."

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. An *interpretative argument* specifies all of the inferences and assumptions made in the process of assigning scores to individuals 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*. The validity argument 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 Kentucky assessment, the stages can be broken out as

scoring, *generalization*, *extrapolation* and *implication*. Descriptions of each stage are given below along with examples of the validity arguments within each stage.

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 Kentucky assessment uses IRT models, so it is also important to verify the assumptions underlying these models.

Generalization

The second stage of the interpretative argument involves the inferences about the *universe score* made from the observed score. Any test contains only a sample of all 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 questions could be administered. Two major requirements for validity at the generalization stage are: (1) the sample of items administered on the test is representative of the universe of possible items and (2) 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 content validity is upheld and test specifications are met. For the second requirement, estimates of test reliability and the standard error of measurement are key components to demonstrating that random measurement error is controlled.

Extrapolation

The third stage of the interpretative argument involves inferences from the universe score to the *target score*. Although the universe of possible test questions is likely to be quite large, inferences from test scores are typically made to an even larger domain. In the case of the Kentucky assessment, for example, not every standard and benchmark 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 Kentucky test is 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.

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 as well as 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.

Validity Argument Evidence for the Kentucky Assessment

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, the majority of this manual can be considered validity evidence for the Kentucky 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.

Scoring

Scoring validity evidence can be divided into two sections. These sections are the evidence for the scoring of performance items and the evidence for the fit of items to the measurement model.

Scoring of Performance Items

The scoring of constructed-response items and written compositions on the Kentucky assessment is a complex process that requires its own chapter to describe fully. Chapter 11, "Performance Scoring," gives complete information on the careful attention paid to the scoring of performance items. 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 *Yearbook* tables reporting inter-rater agreement and inter-rater reliabilities (Appendix N). The results in those tables show both of these measures are generally high for the Kentucky assessments.

Model Fit and Scaling

IRT models provide a basis for the Kentucky assessment. IRT models 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 carefully 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 Kentucky assessment assume the domain being measured by the test is relatively unidimensional. To test this assumption, a principal components analysis is performed. The scree plots for the principal component analyses for each subject and grade are provided in Appendix Q of the *Yearbook*. A scree plot implying a unidimensional factor structure shows that the slope begins to flatten at the second dimension. In other words, the first factor shows the highest loading in the factor structure, followed by less relevant factors. This type of result in a scree plot is evidence the Kentucky assessment measures a single dimension.

To go along with the principal component analyses, confirmatory factor analyses were conducted to test the model of one factor construct within the K-PREP 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 and Bentler, 1999). However, this statistic is sensitive to sample size and often rejects the model when large samples are used (Bentler and Bonnet, 1980). Alternatives to the chi-square, the goodness-of-fit statistic (GFI: Jöresky and Sörbom, 1993) and adjusted goodness-of-fit (AGFI: Tabachnick and Fidell, 2007), are also sensitive to sample size which has led to researchers reporting them along with other fit indices (Hooper, Coughlan, and Mullen, 2008).

The root mean square error of approximation (RMSEA), a comparative fit index, tells how well the model would fit the population covariance matrix (Byrne, 1998). This fit index favors parsimony since it is sensitive to the number of estimated parameters in the model. There have been a few suggestions of index threshold cut-offs of good fit. The most stringent criterion is 0.06 as suggested in Hu and Bentler (1999). In addition, a confidence interval can be constructed for RMSEA, with a lower limit close to 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 (1999) suggests well-fitting models having an SRMR less than 0.05. Hooper, Coughlan, and Mullen (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. The estimates of these indices are presented in Appendix R of the *Yearbook*. These estimates provide support of the one-factor construct for the K-PREP assessments.

In addition, correlations among the total test score and subdomains are provided in Appendix S of the *Yearbook*. These correlations quantify the relationships among subdomains and the overall test score. These correlations demonstrate that the subdomains comprising the overall test are moderate to highly related (as demonstrated through high correlations) to the overall test while also distinct in the factors they are measuring. Put another way, high correlations are indicative that the assessment is measuring one underlying construct.

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, if an item has a high item total correlation (i.e., 0.30 or above), it 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; 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 item-total correlations in the tables of item statistics.

Justification for the scaling procedures used for the Kentucky assessment is found in chapter 7, "Scaling."

Generalization

There are two major requirements for validity that allow generalization from observed scale 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. Content validity is documented through evidence that the test measures the state standards and benchmarks, to the extent possible. The second requirement for validity at the generalization stage is that 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 Kentucky assessments for different student populations. These sources of evidence are reported in the sections that follow.

Evidence of Content Validity

The tests of the Kentucky Assessment system 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 annually to review new 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 and was 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, that is, 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, "Test Development", Kentucky's testing contractor 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.

Evidence of Control of Measurement Error

Reliability and the SEM are discussed in chapter 9, "Reliability." Appendix G of the *Yearbook* has tables reporting the conditional SEM for each scale score point and Appendix M provides the coefficient alpha reliabilities for raw scores (coefficient alpha is reported for all students and for gender and ethnic groups). Further evidence is supplied to demonstrate that the IRT model fits the data well. Item-fit

statistics and tests of unidimensionality apply here, as they did in the section describing evidence argument for scoring. As previously indicated, results of these analyses can be found in Appendices Q, R, and S of the *Yearbook*.

Validity Evidence for Different Student Populations

It can be argued from a content perspective that the Kentucky assessment is not more or less valid for use with one subpopulation of students relative to another. The Kentucky assessment measures 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, "Test Development," 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 Kentucky's testing contractor and KDE after field-test data are collected.

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. These lines of evidence are detailed below.

Analytic 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 Kentucky assessment is 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, the majority of 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 Kentucky assessment. 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.

Implication

There are inferences made at different levels based on the Kentucky assessment. 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 Kentucky assessment reports 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 in detail evidence showing that the Kentucky assessment is a valid measure 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. Appendix T of the *Yearbook* includes non-response rates for the short answer and extended response items of the Kentucky assessment.

One of the most important inferences to be made concerns the student's proficiency level, especially for accountability tests like the Kentucky assessment. 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 assessment, separate chapters are devoted to them in this manual. Chapter 5 discusses the details of setting 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 statewide), the implication validity of school accountability assessments like the Kentucky assessment 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. As mentioned before, 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.

Summary of Validity Evidence

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

11. Performance Scoring

K-PREP assessments require students to construct their own response to some of the test questions. For example, examinees may be required to provide a short, written response to demonstrate the application of a mathematical formula or a scientific concept. As mentioned earlier in this report, K-PREP tests have 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; extended response items are designed such students may respond completely in no more than one page. For the On-Demand Writing test, students are required to write an essay based on a given prompt. Students are provided multiple sheets, within the test response booklet, to respond to the essays.

All *constructed-response* items are scored against a rubric by human scorers. For Writing, one rubric is applied to all essay responses across grades. However, there are specific conditions of writing mastery included for particular grades and/or modes (e.g., counterarguments for grades 8 and 11). For the remaining content tests, however, the short answer and extended response items are scored with rubrics that pertain to the specific item. 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 all of 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.

Rubric Creation

The constructed response items for Reading, Mathematics, Science, and Social Studies are developed with item-specific rubrics detailing the required demonstration of mastery for achieving each possible score point. At the time of item development, the rubrics are discussed among the content specialists for Pearson and KDE. For On-Demand Writing, however, a scoring rubric was created to meet the needs of judging writing proficiency and providing sufficient score information. The scoring rubric for the On-Demand Writing tasks was created through collaboration between Pearson and KDE. The scoring rubric is designed to be used throughout the life of the On-Demand Writing program.

The Writing tasks under the previous assessment program—KCCT—were scored analytically through three domains: content, structure, and writing conventions. The first two were scored using a 0-4 point scale while the third domain used a 1-4 point scale. For K-PREP, Pearson and KDE discussed transitioning to a holistic scoring model where each writing response would receive a single score that represents a particular level of writing. In addition, the number of score points to include in the rubric became a point of concern. The 6-point rubric model from the National Assessment of Educational Progress (NAEP) was used as the starting point for discussions on this aspect. However, concerns over scorer ("inter-rater") agreement

with a six-point scale as well as the potential for sparsely-used score points led to the adoption of a 4-point rubric for K-PREP Writing.

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 2011 (see chapter 6, "Item Analyses"). After the field test, however, the scoring rubric was revisited to address concerns on the emphasis of counterclaims in argumentative responses across the grade levels. Through discussions between Pearson and KDE, minor modifications were made to the scoring rubric that addressed the concerns. These changes, though, were not considered large enough to warrant rescoring of field-test responses. The scoring rubric is provided in an appendix of this technical manual.

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, the first time items are administered to students in a testing environment. Pearson staff, "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. Several training sets for each constructed response item are constructed during this process: anchor set, practice sets, and qualification sets. In addition, a supplemental set of responses is constructed with multiple responses to all score points. The anchor set consists of multiple responses per possible point and is arranged from low to high; the practice and qualification sets consist of a set number of responses randomly arranged.

Once the training sets have been constructed, they are reviewed by KDE. Pearson and KDE staff meet together to review and discuss the training sets. KDE staff validate the scores provided to the responses in each training set and may recommend removal of responses from a particular set; responses from the supplemental training set may be used as substitutes. Annotations for each response are captured during this meeting as well; statements describing how the response achieves the proposed score. All training sets are validated by KDE before use during scorer training.

Scoring Process

This section describes the process of utilizing scorers for the Kentucky scoring projects, from recruitment to training and quality control.

Recruitment

Recruiting scorers is the responsibility of Pearson, which keeps a database of individuals who have scoring experience. The recruiting of scorers is done by the Pearson's 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 in order to accommodate for some attrition during the project.

Training

Highly qualified scorers are essential to scoring students' responses to constructed response items and writing prompts. Thus, the careful selection of professional scorers to evaluate the constructed-response items and writing tasks is critical in scoring the Kentucky assessments. Pearson has compiled a personnel database containing the academic training and professional experience of more than 4,500 college graduates who have completed the stringent selection process for scorers. This process requires that each candidate successfully completes a personal interview, a written essay assignment and a grammar and editing or a mathematics and science test when appropriate. Such pre-screening of candidates promotes the selection of readers of the highest caliber. Also, 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 particular 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 the Kentucky assessment.

At the beginning of the Kentucky scoring project, all scoring supervisors and scorers assigned to the project complete training specific to the Kentucky 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. The PSC staff develops its training materials to facilitate learning through visual, auditory and kinesthetic channels.

Scoring supervisor training occurs first since 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.

On-line training of scorers takes place after supervisors have been trained. The on-line training agenda includes an introduction to the Kentucky assessment program. It is important for scorers to have an understanding of 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 paper selection about 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 on-line 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 a practice set. 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. 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 demonstrate readiness for live scoring.

Quality Control

As part of quality control, items are double-scored for score consistency analyses. For On-Demand Writing, all responses are double-scored; 10% of responses to the constructed response items (i.e., short answer and extended response) of the other subjects are double-scored. Also, validity scoring is conducted throughout scoring. Validity responses are usually solid score point responses and these exemplar responses 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. Those reports include:

- *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 on a daily basis and cumulatively for the entire scoring project. This report allows scoring supervisors and subject leaders to see whether scorers have a tendency 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. In order to document retraining efforts for scorers with low reliabilities, the PSC maintains a Scorer Intervention Log. Entries on this form describe the feedback given to a scorer regarding his or her problematic scoring and enumerate the interventions taken. Scorers are dismissed if they have been counseled, retrained, given every reasonable opportunity to improve and are still performing below the acceptable standard.

Appendix N of the *Yearbook* contains summaries of the interrater agreement rates and score point distribution for constructed response items—short answer, extended response, and writing prompts—from the 2018 tests. This summary includes statistics from the current test administration as well as statistics from a previous

test administration in which the items were administered (highlighted in blue). For Reading, Mathematics, and Social Studies, the previous administration for the items in the summary can be from 2012 through 2017 (operational or field test). For Writing, the previous statistics are from the 2011 stand-alone field test. Appendix O of the *Yearbook* contains a summary of *total* scores as well as interrater agreement rates for Writing, by grade.

Security

Scorers assigned to the Kentucky assessment program must sign a nondisclosure agreement before they can see any K-PREP test materials. Furthermore, all materials provided to scorers are secured via security guidelines and infrastructure by Pearson.

Finally, all operational scoring is conducted by using Pearson's image-based scoring system. This system is 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 Kentucky 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.

12. 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 K-PREP from situations that could affect the reputations of both Pearson and KDE. This chapter will highlight how quality control measures are implemented throughout the assessment program.

Test Construction

Guidelines of test development are outlined in chapter 2, "Test Development", beginning with item development and going through forms construction. These guidelines help test developers—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 as well as maintain statistical quality appropriate for the assessment.

Non-Scannable Documents

Pearson contracts with outside vendors for the printing of non-scannable documents due to the large volume of printed materials necessary for K-PREP. The following quality controls are implemented to facilitate the successful performance outside printing vendors.

- Pearson provides design and schedule requirements to print vendors well in advance of the delivery of copy materials so that the printing schedule can be arranged.
- Changes made to the print schedule by either Pearson or KDE are immediately to the print vendors.
- Corrections to print materials are submitted to the print vendors.
- All page proofs, final proofs and printed materials are proofread in their entirety by the forms support department and are submitted to KDE for review.
- Sample printed materials are examined for the required paper type, ink color, collation, and copy quality. If discrepancies are found, the print vendor is immediately notified so that corrections and reprints can be made.
- Whenever possible, electronic transfer of copy materials is used to minimize human error and to expedite the printing process.

Pearson conducts an additional quality check of all outside printed materials during materials packaging.

Data Preparation

For an accurate accounting of the volume of K-PREP assessment documents that Pearson receives, Data Preparation staff perform a series of receipt and check-in procedures. All incoming materials are carefully examined for a number of

conditions, including damage, errors, omissions, accountability and secured documents. When needed, corrective action is promptly taken according to specifications developed jointly be Pearson and KDE.

Production Control

Pearson uses the "batch control" concept for document processing. When documents are received and batched, each batch is assigned a unique identifying number. This identifier assists in locating, retrieving and tracking documents through each processing step. The batching identifier also guards against loss, regardless of batch size.

All K-PREP assessment documents are continually monitored by Pearson's Workflow Management System (WFM). This mainframe system can be accessed throughout Pearson's processing facility, enabling Pearson staff to instantly determine the status of all work in progress. WFM efficiently carries the planning and control function to first-line supervisory personnel so that key decisions can be made properly and rapidly. Since WFM is updated on a continuous basis, new priorities can be established to account for K-PREP assessment documents received after the scheduled due date, late vendor deliveries, or any other unexpected events.

Scanning and Editing

Stringent quality control procedures and regular preventative maintenance operations are implemented so that Pearson's high-speed scanners function properly at all times. In addition, application programs consistently include quality assurance checks to verify the accuracy of scanned student responses.

Over the years, Pearson has developed a refined system of validity checks, editing procedures, error corrections, and other quality controls for maximum accuracy in the reporting of results. During scanning, K-PREP assessment documents are carefully monitored by trained scanner operators for a variety of error conditions. These error routines identify faulty documents, torn and crumpled sheets, document mis-feeds and paper jams.

As K-PREP answer documents are scanned, the data are electronically transcribed directly to data files, creating a project database. After scanning, a three-step editing process is performed to verify that all data in the database is complete and accurate. During this process, the data are examined for omissions, inconsistencies, gridding errors, and other specified error-suspect conditions.

The first step in editing consists of a complete computer editing of the data to verify that all documents are accounted for and all possible "suspects" or omissions have been checked. In the second step, editing personnel review the errors detected during the first step and indicate the necessary corrections to be made. The editing staff inspects both the computer-generated edit log and the actual field or information that may be in error. The editing staff visually checks this particular piece of information against the source document. At this point, double grids, erasures and smudge marks are flagged. From this, one of three actions is taken:

- *Correctable error:* If an error is correctable by the editing staff according to editing specifications, then the corrections are handwritten on the edit log, checked by a lead staff member and the required changes are made by the Data Input department. These editing specifications are customized for requirements specified by KDE.

- *Error Not Correctable According to Specifications:* If an error is not correctable according to the specifications, the Project Director and KDE will be notified. The correction information will be obtained from KDE for the item in questions. The specifications for the types of error corrections requiring contact with KDE are developed jointly.
- *Non-correctable error:* If a "suspect" is found, but no alterations are possible according to the specifications, the proper procedure to allow this type of data to remain on the records is initiated, and no corrective action is necessary. An example of this would be an answer document containing double-gridded student demographic information.

Once the necessary corrections have been entered in the edit log and checked by a lead staff member, the batch is forwarded to the Data Input department, where corrections are key-entered and key-verified on data entry terminals. At this point, the updated batch files will contain only valid information. The data entry screens are designed to enhance operator speed and accuracy: fields to be entered are titled to reflect the actual source document. When all corrections for a batch have been entered and verified, then the correction file is submitted to the mainframe computer for updating of the batch data file.

The third step in editing process, "post-edit," takes place as the data file is being updated. During this step, the entire data file is again re-edited according to an editing procedure approved by KDE.

Performance Scoring

Quality control measures are implemented throughout all phases of the performance scoring process. These measures with start 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 those 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 11, "Performance Scoring."

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 final results:

- Student data is inspected for completeness and accuracy of data, 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 items with statistical values below the minimum acceptable value are reviewed to verify that the item 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—item calibrations and scaling—are performed by two independent replications of Pearson staff and one external ("third-party")

consultant. The results from these replications are compared for consistency. Any unexpected differences are resolved. In addition, conference calls are held daily during the psychometric analyses.

- A summary of the psychometric analyses is provided to KDE for review.

Scoring and Reporting

Before reporting, script and conversion programs with mock data are run to check that accurate reports are being produced. In addition, a random sample of reports are 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 is checked again using production data. Data files are provided to KDE prior to the release of the score reports. This data is used by KDE to confirm the reported data is correct as well as 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.

13. On Demand Writing

The K-PREP On-Demand Writing (ODW) assessment was first field tested in 2011 and implemented operationally beginning in 2012. Under the current test design, the writing test consists of one passage-based prompt and two stand-alone (short stimulus) prompts. Students must choose to respond to one of the two stand-alone prompts, and all students must respond to the passage-based prompt. In general, the test design includes different writing genres administered across years, within grade. Student responses are double scored based on a 4-point rubric. Final scores are determined as the sum of all ratings across the two prompts, resulting in a raw score range from 0 to 16. In line with the other K-PREP tests, scores are also reported by performance level (Novice, Apprentice, Proficient, and Distinguished), determined through raw score cut points from standard setting workshops in 2012.

Prior to 2015 test administration, the comparability of ODW scores across years has been premised on the basis that prompts are of similar difficulty and that administration conditions and scoring processes are the same from year to year[10]. Unlike the other K-PREP content areas, repeating ODW test items across years for purposes of "equating" is challenging. While overall performance comparisons across years has been reasonable, each year there have been instances in which a given grade's results are more marked. This is a result of the limitation of having no mechanism in place to help control for differences in prompt difficulty across administrations, as is done with the other K-PREP tests. In 2015 KDE decided to use a Rasch model based approach for ODW to establish a more stable reporting scale and equate scores over subsequent test forms.

Base Scale

Because no equating was possible between 2015 ODW and 2014, the reporting of 2015 ODW scores reflected a similar trend as previous years. That is, the raw score cuts used for reporting performance level information were used to determine the cuts on the new scale score metric and for 2015 score reporting. While no direct means of equating 2015 ODW to previous years existed, the process of establishing a base scale using the Rasch model in 2015 did provide a mechanism for the respective choice prompts to exist on the same scale through the common prompt.

To establish the 2015 ODW base scales, all prompt ratings within each grade-level data were calibrated according to the Partial Credit Model (Masters, 1982). Using a cumulative frequency distribution of IRT proficiency estimates across students, the performance level distribution of each grade was set using the raw score cut points from the 2012 standard setting workshops. Table 13.1 provides the raw score cuts, introduced in Chapter 5 ("Performance Standards"), and the 2015 performance distributions based on those raw score cuts. From these distributions, threshold points on the IRT scale were determined to take the place of the raw score cuts. In order words, the proficiency estimates that defined the 2015 performance distributions in Table 13.1 will determine the performance levels for the ODW test. These estimates are provided in Table 13.2.

[10] This raw score approach was implemented because of low numbers of 4's being observed on several prompts during field testing and the concern that this would undermine the use of the same model used to scale the other K-PREP tests.

Table 13.1 Raw Score Cut Points and 2015 Raw Score Impact

Grade	Raw Score Cut Points			2015 Impact			
	Apprentice	Proficient	Distinguished	N	A	P	D
5	7	10	14	12.9%	43.3%	40.0%	3.8%
6	6	9	13	13.1%	42.4%	39.2%	5.3%
8	7	11	14	15.0%	50.6%	28.1%	6.3%
10	7	11	14	11.4%	48.1%	33.8%	6.8%
11	7	10	14	10.0%	27.8%	52.1%	10.1%

Table 13.2 Derived ODW Cut Points on Theta Metric

Grade	Apprentice	Proficient	Distinguished
5	-4.1348	1.3367	6.5145
6	-3.7609	1.179	5.9678
8	-3.327	1.9726	5.4038
10	-5.4448	1.6067	5.9538
11	-4.5726	-0.3169	5.2257

Equating

In 2016, multiple test forms were used for the K-PREP On Demand Writing assessment to establish common-item comparability for test equating. An equating analysis was conducted through administration of previously administered writing prompts to a sample of the 2016 testing population thus producing a common-item design across two test administrations. In all, three test forms of writing prompts were administered across the 2016 testing population, with each student being administered one form. The ODW equating analysis consisted of: 1) computing equating constants through the common-item design; 2) placing all test forms, within a grade, onto the same scale by choosing one form to serve as the "base"; and 3) adjusting the measurement parameters of each form (i.e., adjusting the item parameters of the writing prompts) by the equating constants to produce equated student proficiency estimates.

Using the common-item data, equating constants were derived from the step difficulties obtained from the calibration of the writing prompt scorers. Since each prompt response is scored independently by two scorers, each prompt score was treated as a separate "item" during the calibration. The first "item" represented the first scores provided by human scorers and the second item represented the second, independent, scores provided by human scorers. Therefore, calibrated data for one grade includes six calibrated items: two calibrated items for the passage (mandatory) prompt, two calibrated items for choice (stand-alone) prompt #1, and two calibrated items for choice prompt #2. The raw scored data for each item ranged from 1 to 4, resulting in three step difficulties per item in the IRT measurement framework.

A statistical procedure outlined in Masters (1984) was used to calculate the equating constants across the common items' step difficulties. The constant, for each grade, was computed as:

$$t_{XY} = \frac{\sum_{i}^{n} \sum_{j=1}^{m_i} \left(d_{ijX} - d_{ijY}\right) / W_{ijXY}}{\sum_{i}^{n} \sum_{j=1}^{m_i} 1 / W_{ijXY}}$$

where d_{ijX} is the estimated step difficulty of the *j*th step in item *i* on Form X (calibration estimate from previous administration) and d_{ijy} is the estimated step difficulty from the same item on Form Y (calibration estimate from current administration). W_{ijXY} is a weight based on the calibration errors associated with the step difficulties. This weight is calculated as:

$$W_{ijXY} = e_{ijX}^2 + e_{ijY}^2$$

where e^2_{ijX} and e^2_{ijY} are the squared error estimates for d_{iX} and d_{iY}, respectively. Table 13.3 provides the computed equating constants.

Table 13.3 2016 ODW Equating Constants

Grade	Equating Constant
5	-0.0193
6	-0.0750
8	0.3615
10	0.1332
11	-0.0878

Prior to the application of the equating constants, each 2016 test form, within a grade, was calibrated separately and placed on a common scale with one form chosen as the base. The form chosen as the base was administered to the majority of the 2016 testing population in each grade. Placing each form on a common scale was done through adjusted mean and standard deviation of student proficiency estimates (i.e., theta) from the base form. These forms were then recalibrated with their adjusted mean and standard deviation estimates of student estimates.

Once the test forms, within a grade, were placed on a common scale. The equating constants were applied to the *overall* difficulty parameter estimates for each "item". Then, student proficiency estimates, for scale score generation, were computed for each form-by-choice-prompt status. Since students select one of their scored writing prompts, separate scoring tables are needed for both possible outcomes from a test form: passage-base prompt with choice (stand-alone) prompt #1 and passage-based prompt with choice prompt #2.

Scaled Scores

The ODW test scores are be reported on a scaled score metric that began with the 2015 test administration. Similar to the scaled scores for other K-PREP content areas, the scaled scores are derived through linear transformations 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 examinee response data.

The scaled score metric for ODW was chosen to range from 100 to 300, with 235 representing the minimum scaled score for passing *(Proficient)*. To achieve this score metric, the following linear transformation was proposed:

$$SS = m(\theta - \theta_p) + b,$$

where the slope (*m*) was set to 8.335, the intercept (*b)* was set to 235*,* and *θ* is the person proficiency estimate defined as before. This transformation also includes θ_p, the person proficiency estimate identified as the minimum value for *Proficient*. The values used for this term are the values listed for *Proficient* in Table 13.2.

Given that 1) multiple test forms were administered per grade in 2015-2016, 2) students can select one prompt as part of the ODW assessment and 3) the use of IRT in test scoring, it is necessary to create scoring tables that apply based on the test form and choice prompt selected by students. For example, "Form 1A" can be considered the choice prompt #1 *plus* the passage-based prompt and from one set of test forms and "Form 1B" as choice prompt #2 *plus* the passage-based prompt from that same set for forms. The possible scaled scores for each form and choice prompt combination may be slightly different due to differences in overall "form" difficulty. This phenomenon is similar to what occurs for the other K-PREP content areas (e.g., Reading, Mathematics, etc.) with different test forms across years. Performance ratings (scores) for ODW are assigned based on a performance rubric, which is not specific to test form. Therefore, although there are different scaled scores across forms within a grade, the standard by which performance is evaluated and reported is the same across forms.

Appendix G of the *Yearbook* contains the tables of derived scaled scores for the 2018 ODW assessments (grades 5, 8, and 11). Scaled score cumulative distributions for each grade are provided in Appendix I of the *Yearbook*. Descriptive statistics—mean, standard deviation, minimum, maximum—for the scaled scores of the ODW assessment are provided in Appendix K of the *Yearbook*. The descriptive statistics are provided for the overall testing population, as well as by subgroups— gender, ethnicity, free/reduced lunch status, and accommodations.

The scaled scores align to definitions of achievement—performance levels. Transforming the theta cut points in Table 13.2 into scaled scores, the performance levels—Novice, Apprentice, Proficient, and Distinguished—for ODW are defined by the scale score ranges presented in Table 13.4. Note: Cut scores for grades 6 and 10 are provided for historical references. Performance level distributions for the 2018 ODW assessments are provided in Appendix L of the *Yearbook*.

Table 13.4 ODW Scaled Scores by Performance Level

Grade	Novice	Apprentice	Proficient	Distinguished
5	100-188	189-234	235-277	278-300
6	100-193	194-234	235-274	275-300
8	100-190	191-234	235-263	264-300
10	100-175	176-234	235-270	271-300
11	100-199	200-234	235-280	281-300

Glossary of Terms

Classical test theory: 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: Test item that requires a form of written response by the examinee.

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

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

Differential item functioning: 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: Measurement theory that prescribes relationships of item difficulty and examinee proficiency for indices of test performance.

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

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

Norm-referenced test: Test that reports examinee performance according to the performance of other examinees.

Percentile rank: A numerical value indicating relative standing of performance among other examinees.

Performance level: A categorization of achievement from test performance.

Performance level descriptor: A description of the performance level, outlining the knowledge and skills typical for that achievement 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: 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.

Scale score: A score derived from a transformation of a raw score.

Scaling: Transforming scores into meaningful and comparable units.

Standard error of measurement: A statistic, in classical test theory, expressing the interval of an examinee'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, providing the number of test items by content and subdomain levels.

Test design: A general summary of test form layout.

True score: An examinee'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 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.

References

American Educational Research Association, American Psychological Association, & the National Council on Measurement in Education. Joint Technical Committee. (1999). *Standards for Educational and Psychological Testing.* Washington, DC: American Educational Research Association.

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*. Mahwah, New Jersey: Lawrence Erlbaum Associates.

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

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

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

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

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. Columbia, SC: 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). Available online: 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.* Maple Grove, MN: JAM Press.

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

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

Kingston, N. M., Kahl, S. R., Sweeney, K., & Bay, L. (2001). Setting performance standards using the body of work method. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp.219-248). Mahwah, NJ: Lawrence Erlbaum.

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

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

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

Masters, G. N. (1984). Constructing an item bank using partial credit scoring. *Journal of Educational Measurement*, 21(1), 19-32.

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.

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

Tabachnick, B.G., & Fidell, L.S. (2007). *Using multivariate statistics (5th ed.).* New York: 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. Reading Item Writer Training

Kentucky

Item Writer Training
for K-PREP Reading Assessment

Training Objectives

Objectives:

- To explain item writer responsibilities
- To analyze the Common Core State Standards to which items must align
- To familiarize item writers with item writing styles and requirements
- To introduce/review methods of creating
  - multiple choice items
  - short answer items
  - extended response items
- To review the process of communicating with Pearson and transporting of materials to Pearson

Item Writer Responsibilities

- Confidentiality

  Item writers must maintain the security of all test items, documents, and material being created. Item writers **must not** retain paper or electronic copies of any material created for the KY development after assignment has been completed.

- Nondisclosure

  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.

- Ownership

  All materials developed for this assessment program **must be original** and may not appear in any other source. They are the property of the state of Kentucky and **may not be used for any other purpose**.

Item Writer Responsibilities

- Schedule

  Item writers **must** adhere to the schedule and submit items according to the directions on or before the stated deadline.

- Source Documentation

  Item writers **must** submit source documentation for factual information used in the contexts of the items. (Since all items are based on the passages which have been verified by RL and new information cannot be introduced in the items, this will most likely be a non-issue.)

- Material Familiarity

  Item writers **must** read and study the Common Core State Standards (CCSS) PDFs that are provided in the training material packet. **All created items must align to these standards** and meet any specific guidelines from the Kentucky Department of Education.

ITEM WRITER GUIDELINES

1

PEARSON

General Considerations in Item Writing

Items must

- ask questions that are worthy of being asked.
- align accurately to Common Core State Standards assigned.
- use the language of the standards.
- cover a variety of standards.
- have varied difficulty levels.
- be clearly and concisely written.
- be grammatically correct.

Item Writing Considerations Continued

Items must

- be able to be answered correctly **using the text** and inferences from the text provided.
- use language at or slightly below grade level to avoid any misunderstanding of what is being asked.
- be clear and concise leaving no doubt as to what the question is asking and to which standard the question aligns.
- be specific to the passage.
- avoid using contractions and multiple meaning words as these can cause problems for some students.
- unanswerable without using the text.

Common Core State Standards
CCSS

The K-PREP assessment will be assessing students' reading comprehension in both literature and informational genres and students' vocabulary acquisition and use.

You will be given an assignment which lists exactly the number of items, types, and alignments needed for each passage. We will ask you to create your items to designated CCSS standards to which items align using the following naming conventions when completing the metadata for each item submitted:

RL (Reading literature) 3 (grade level) 4 (standard number) = RL.3.4
RI (Reading informational) 6 (grade level) 8 (standard number) = RI.6.8
L (Vocabulary acquisition and use) 8 (grade level) 5c (standard number) = L.8.5c

Standards Addressing Single / Multiple Texts

Literature	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
RL2	S	S	S	S	S	S
RL3	S	S	S	S	S	S
RL4	S	S	S	S	S	M
RL5	S	M	S	S	S	M
RL6	S	M	S	S	S	S
RL7	S	M	M	M	M	M
RL8	NA	NA	NA	NA	NA	NA
RL9	M	M	M	M	M	M
Informational	**Grade 3**	**Grade 4**	**Grade 5**	**Grade 6**	**Grade 7**	**Grade 8**
RI2	S	S	S	S	S	S
RI3	S	S	S	S	S	S
RI4	S	S	S	S	S	M
RI5	S	S	M	S	S	S
RI6	S	M	M	S	M	M
RI7	S	S	M	M	M	M
RI8	S	S	S	S	S	S
RI9	M	M	M	M	M	M

CCSS Reading Standards for Literature

Standard 1 in the cluster for Key Ideas and Details is a standard to which any item may align; therefore, it really is not extremely useful for assessment purposes.

NO items will be created to align with Standard 1:

Grade 3: Ask and answer questions to demonstrate understanding of a text, referring explicitly to the text as the basis for the answers.
Grade 4: Refer to details and examples in a text when explaining what the text says explicitly and when drawing inferences from the text.
Grade 5: Quote accurately from a text when explaining what the text says explicitly and when drawing inferences from the text.
Grade 6: Cite textual evidence to support analysis of what the text says explicitly as well as inferences drawn from the text.
Grade 7: Cite several pieces of textual evidence to support analysis of what the text says explicitly as well as inferences drawn from the text.
Grade 8: Cite the textual evidence that most strongly supports an analysis of what the text says explicitly as well as inferences drawn from the text.

Grade 3 CCSS Reading Standards

Key Ideas and Details

These standards are to be used with a single passage or one specific passage of a pair. They are not to be used with crossover items pertaining to both passages in a pair.

`

Literature

Standard 2: The focus of this standard is seeing how the key details in the passage relate or show the main idea, lesson or moral or message of the passage. Stories, folktales, fables, myths and stories of diverse cultures are included.

Standard 3: The focus of this standard is on what the characters are like, what they do, why they do what they do and how they feel in relation to what's happening around them. Because this standard talks of motivations, "why" or "how" will be useful interrogatives to use.

Informational

Standard 2: The focus of this standard is really split into two. Items may legitimately ask students to ascertain the main idea, or they may ask students to identify and explain how details within the passage support the main idea. Items that ask what sort of support is provided or which details in the passages offer support or the best support for a particular main idea will also align.

Standard 3: The focus of this standard is on the organization and relationship developed between ideas, events, steps in a procedure. Language will include that which pertains to time, sequence, and cause/effect relationships.

Grade 3 CCSS Reading Standards

Craft and Structure

These standards are to be used with a single passage or one specific passage of a pair. They are not to be used with crossover items pertaining to both passages in a pair.

Literature

Standard 4: This standard deals with students ascertaining the meaning of words and phrases chosen by the author to serve a certain purpose. The distinction between this and the vocabulary standard is in the idea that the chosen word or phrase will have a slightly different or figurative meaning relating specifically to the passage rather than the vocabulary item which will simply be asking students for the denotative meaning of the word or the context used to determine the meaning.

In paragraph XX of the story, what is the meaning of the phrase "jealousy awakens in her"?
In paragraph 3 of the story, which meaning of "swallowed" is used in the phrase "he swallowed his anger"?

Standard 5: The main focus, aside from using the actual literary terminology, is on the **interconnections between parts of the text**. Asking "what effect" or what is the author's purpose or intent with text organization and structure would also fit into this standard. Items should be about the structure of the passage and how the author has the parts fitting together to develop his/her ideas.

Informational

Standard 4: The focus on this standard is on word choices. Asking students to be able to determine meaning of words that are general academic or domain specific and relevant to topic or grade appropriate subject are applicable item types.

Standard 5: The focus of this standard is on any illustrations or graphics, charts or other text features used in the passage. Asking students how the text features aid in understanding or what information they elicit in a single passage are applicable items for this standard alignment.

Grade 3 CCSS Reading Standards

Craft and Structure

This standard is to be used with a single passage or one specific passage of a pair. It cannot be used with crossover items pertaining to both passages in a pair.

Literature

Standard 6: The emphasis of this standard is to show a distinction between the readers' thoughts and those of the narrator, speaker, or characters in the passage. This standard will eventually evolve into determination of points of view and analyses of how authors demonstrate points of view and use varying points of view within a passage as the grade levels increase. It is not enough at grade 3 to simply ask students to identify who is telling the story as this will not address the reader's perspective.

Possible ways of addressing this standard in item stems might be:

The author states that . . .
(Give the point of view then ask) Another opinion might be . . .
(Address the reader directly for an SA item): Do you agree or disagree with the author? Support your answer.
(Identify a part of the text then ask): The narrator wants the reader to understand that . . .

Informational

Standard 6: The focus of this standard for informational text has a similar emphasis to the literature standard. Items must distinguish the reader's thoughts and reactions from the author's point of view.

Note: Standard 6 is a part of the craft and structure, so author's craft is the emphasis!

Grade 3 CCSS Reading Standards

Integration of Knowledge and Ideas

Literature

Standard 7: This standard focuses on the illustrations or graphics within a text to support or provide information or clarity of characters, plots, settings, etc. in the passage. **This standard is used with single passages and cannot be used with a crossover item.**

Standard 8: Standard 8 is not applicable to literary passages.

Standard 9: The focus of this standard is on the comparisons of themes, settings, plots, etc. (likenesses and/or differences) between two texts. **This standard is the one standard in grade 3 which must be used with crossover items / multiple texts.** This standard may not be used for a single passage as reference to the paired text must be made. Although students may actually be responding concerning one of the texts, they must have to go back to both texts of the pair to synthesize their response.

Informational

Standard 7: Just like in literature standards, this standard focuses on the photos, illustrations, graphics, charts, maps, etc. and how these impact understanding of the text's information. **This standard is used with single passages only.**

Standard 8: The focus of this standard is on the type of connection or impact particular parts of the text have on others. Looking at how specific sentences or paragraphs create the cause/effect or comparison, or organization of information within a text is key. **This is used with a single passage only.**

Standard 9: This is the only standard to be used as a crossover in grade 3 with multiple texts. The focus of the items will be on comparisons of details and other important information between the two texts.

Grade 4 CCSS Reading Standards

Key Ideas and Details

These standards are to be used with a single passage or one specific passage of a pair. They are not to be used with crossover items pertaining to both passages in a pair.

Literature

Standard 2: The focus is on a theme of a play, poem, or drama in the first part of the standard. The second part of the standard is about asking students to summarize. As a general rule, the stem or options should contain one of those words to provide a clear direction for the student.

Standard 3: The focus is on an in-depth description of the character, setting, or events in a literary passage. The student should have to draw on details, such as a character's thoughts, words, or actions, in order to provide a correct response. This standard lends itself particularly well to a short answer item. It cannot be used for an extended response, since ER items may not align to any of the Key Ideas and Details Standards.

Informational

Standard 2: One focus may be on basing the identification of a main idea on support in the text. In addition, an item aligned to this standard may request a summary of a particular paragraph or event. Items may either ask for the detail which should be included in a summary (or supports a given main idea) or it may specify the detail and ask for the main idea it supports.

Standard 3: The focus is on the use of information in the passage to explain events, procedures, ideas, or concepts. Like its literary counterpart, this standard lends itself particularly well to a short answer item.

Grade 4 CCSS Reading Standards

Craft and Structure

Literature

Standard 4: This standard deals with students ascertaining the meaning of words and phrases, including mythological references (explained or footnoted), that are chosen by the author to serve a certain purpose. If there is a mythological reference which is not explained in the passage, then the stem has to provide context. For example:

In mythology, Hercules was known for his strength. Why does the author use the phrase "Herculean effort" in line xx of the passage?

The distinction between this and the vocabulary standard is in the idea that the chosen word or phrase will have a slightly different or figurative meaning relating specifically to the passage rather than the vocabulary item which will simply be asking students for the denotative meaning of the word or the context used to determine the meaning. This standard is to be used with a single passage or one specific passage of a pair. It is not to be used with crossover items pertaining to both passages in a pair.

Standard 5: The main focus is on comparing the structural elements of various forms of literature, particularly those that are unique to poetry and drama. Examples of these elements are: verse, rhythm, rhyme, meter, setting, cast of characters, dialogue, stage directions. This standard is to be used with multiple texts. It is to be used only with crossover items pertaining to both passages in a pair.

Informational

Standard 4: The focus of this standard is on academic words which are specific to a grade 4 subject area or domain. This standard is to be used with a single passage or one specific passage of a pair. It is not to be used with crossover items pertaining to both passages in a pair.

Standard 5: The focus of this standard is on the structure of a passage, or part of a passage, such as chronological, comparison, problem/solution, and cause/effect. This standard is to be used with a single passage or one specific passage of a pair. It is not to be used with crossover items pertaining to both passages in a pair.

[**Domain specific vocabulary** includes low-frequency content specific words appearing textbooks and other instructional material. **General academic vocabulary** words appear frequently across academic domains, e.g.: classify, declaration, etc.]

Grade 4 CCSS Reading Standards

Craft and Structure Continued

These standards are to be used with multiple texts. They must only be used with crossover items pertaining to both passages in a pair.

Literature

Standard 6: The emphasis of this standard is on comparing the narrator's point of view in two texts. This comparison may include the difference in first and third person narration.

Informational

Standard 6: The focus of this standard is on comparing and contrasting a first and second-hand account of the same event or topic. Examples could be an autobiographical versus biographical view of a person or a concurrent news report versus a historical report of an event.

Grade 4 CCSS Reading Standards

Integration of Knowledge and Ideas

Literature

Standards 7: If items were possible for this standard, they would have to be crossover items / multiple texts. Because the emphasis here is on a comparison between a written text and a presentation that is seen or heard, items aligning to this standard for a paper/pencil assessment are not possible.

Standard 8: This standard is not applicable to literature.

Standard 9: The focus of this standard is on the comparisons of literature from different cultures. The comparison should involve stories and myths from different cultures that address a similar theme or topic. This standard is to be used with multiple texts. It is to be used with crossover items pertaining to both passages in a pair.

Informational

Standard 7: This standard focuses on interpretation of timelines, charts, graphs, etc. and how these impact understanding of the text's information. **This standard is used with single passages only.**

Standard 8: The focus of this standard is on reasons and evidence presented in support of particular points. This may be used in terms of persuasive arguments, although examples may be found in purely descriptive or informational passages. **This is used with a single passage only.**

Standard 9: This standard requires that the reader integrate information from two texts on the same topic in order to answer the item. This standard is only to be used with multiple texts in crossover items.

Grade 5 CCSS Reading Standards

Key Ideas and Details

These standards are to be used with a single passage or one specific passage of a pair. They are not to be used with crossover items pertaining to both passages in a pair.

Literature

Standard 2: The focus here is on the determination of the theme that comes from the supportive details. Items that look at how characters behave, what their motivations are, or how they respond to situations, other characters, or thoughts are viable items to assess this standard. Items may focus on the theme or the support used to demonstrate or elucidate this theme.

Standard 3: The comparisons and/or contrasts in this standard focus on characters, settings, plots and the interactions between them.

Informational

Standard 2: The focus of this standard is actually split. Items may look for two or more main ideas within a single text and show the support given for these ideas within the text. A backdoor approach for an item might be to give the main ideas and then ask students to find the type of support given. A second focus for some items might be in the matter of summarizing the key points within the single text. Asking students to identify key details that may be needed to complete an accurate summary of the text or individual paragraph or asking students to recognize the most accurate summation are viable item types.

Standard 3: The focus of this standard is on interactions within a single text. How individuals, events, ideas or concepts in any informational text are connected or relate are viable item directions.

Grade 5 CCSS Reading Standards

Craft and Structure

These standards are to be used with a single passage or one specific passage of a pair. They are not to be used with crossover items relating to both passages in a pair. (Exceptions are standards 5 and 6 in the informational.)

Literature

Standard 4: The focus here is on the determination of meaning of words and phrases used by the author. These uses may be figurative. Looking at how an author uses words and phrases can also be a way of determining the figurative meaning. In addition, this standard goes beyond simply asking for denotative meanings which would be classified as a vocabulary item.

Standard 5: This standard is looking at the organizational structure of a single passage. Asking students to explain or recognize how parts of the text interact and fit together to impact the whole is a part of this assessment.

Standard 6: The emphasis in this standard is within a single passage. Students must be able to describe the impact or effect the narrator's or speaker's points of view have on the plot or the passage as a whole.

Informational

Standard 4: For informational pieces, this standard is used to assess students' understanding of the meaning of academic vocabulary and domain-specific vocabulary words or phrases relevant to the topic.

Standard 5: The focus here is on how the author has chosen to structure the information within the paired passages — what techniques are used to create chronology, cause/effect, problem solution etc. **This is to be used with crossover items only since it involves multiple texts.**

Standard 6: The emphasis is given to the multiple accounts of same event; this will be used only with crossover items/multiple texts. The comparison or contrast of the accounts will be looked at to assess students abilities to note likenesses, differences in the points of view and their impacts.

Grade 5 CCSS Reading Standards

Integration of Knowledge and Ideas

Literature

Standard 7: This standard will be used only with crossover items as multiple texts are involved. The difficulty will be that the multi-media portion of this standard cannot be assessed easily with a paper/pencil test. Using hypothetical examples of multi-media presentations is not acceptable as this is artificial. Therefore, this standard will be very difficult if possible at all to create items that will assess it adequately. If visual text, such as graphic novel types were paired asking students to explain how the visualization contributes.

Standard 8: This standard is not applicable to literary passages.

Standard 9: This is the standard that will be used for crossover items (multiple texts) whether they are multiple choice or constructed response items. The focus here is on a comparison and/or contrast between stories, their themes, characters, etc. in the same genre. Noting how the authors in each pair confront or demonstrate themes, how they use characters, etc. will be viable assessment for this standard. Extended response items can only come from this standard.

Informational

Standard 7: This standard will be used with crossover items (multiple texts). If the paired passages include a graphic of some sort, crossover items may align. The focus here is on the information that is included within these multiple print sources and how this information can be used to access information.

Standard 8: This standard must only be used with a single passage. Its focus is on the author's use of reasons and evidence in presenting an argument or making a claim. A second focus may also be an evaluation of the soundness, credibility and/or the effectiveness of the evidence given.

Standard 9: This standard may only be used with crossover items as it involves the use of multiple texts. The focus is on the integration or synthesis of the information provided in two informational texts in order to create a response.

Grade 6 CCSS Reading Standards

Key Ideas and Details

These standards are to be used with a single passage or one specific passage of a pair. They cannot be used with crossover items pertaining to both passages in a pair.

Literature

Standard 2: The emphasis of this standard is two-fold. First, a determination of the theme and how it is supported by details is a key. Items can assess this by either asking for the theme or asking for the support given. The second direction items assessing this standard may go is through summarization. While summary items are often best suited for SA items, it is possible to create MC items which ask students to identify the most accurate summary or asking students to determine which addition to an accurate summary is needed.

Standard 3: The emphasis of this standard is on the plot and character. Asking students to look at plot development and ascertaining how characters of stories or plays behave or respond in the particular plot situations are viable items to assess this standard.

Informational

Standard 2: As with the literary counterpart, this standard for informational passages asks students to determine the central message or idea of a passage and look at the detail support that is provided. A summation of the passage or part of the passage is also a focus.

Standard 3: The focus of this standard is on the key elements of the informational text and how those key elements are introduced and developed. These elements may be concepts or ideas presented, events, or individuals involved. Looking particularly at the author's method of introducing and developing the elements within the text is key.

Grade 6 CCSS Reading Standards

Craft and Structure

These standards are to be used with a single passage or one specific passage of a pair. They cannot be used with crossover items pertaining to both passages in a pair.

Literature

Standard 4: There are two different focuses for this standard. Ascertaining meaning from words, phrases, figurative language uses and connotative meanings is an important piece. To distinguish this standard more from the vocabulary acquisition standard, the emphasis here will be more on the meanings of figurative uses and connotations and their broader impact. The vocabulary standard is more apt to assess specific denotative meaning of words within a particular sentence. The second focus of this standard is on author's word choice and the impact these authorial decisions have on the meaning and tone.

Standard 5: Physical structure and organization is a focus here. Students must analyze how a particular part of the passage fits into the whole and impacts or contributes to the passage development particularly as the parts work to develop or impact theme, plot, or setting.

Standard 6: This standard assesses students' abilities to explain how an author uses or develops the point of view within a passage.

Informational

Standard 4: The focus of this standard is on the determination of meaning with emphasis on figurative, connotative rather than denotative, and technical.

Standard 5: Again, structure and organization of parts in relationship to the development of the whole is key.

Standard 6: This standard in informational pieces goes beyond point of view. Assessing students' abilities at determining point of view and also author's purpose or intent and being able to explain how the author develops and conveys this purpose is key.

Grade 6 CCSS Reading Standards

Integration of Knowledge and Ideas

Literature

Standard 7: If this standard were possible for a paper/pencil test, items would have to be crossover involving multiple texts. Because the comparisons in this standard are between a written text and one that is either viewed or heard, the opportunity to assess students using this standard is not applicable to a paper/pencil assessment.

Standard 8: This standard is not applicable to literary text.

Standard 9: This standard can only be assessed using crossover items relating to multiple texts. The emphasis is on the comparison of texts of different genres in terms of the way each approaches or develops a common theme and/or topic.

Informational

Standard 7: This standard may only be assessed using crossover items relating to multiple texts. As long as the visual component or format is printable, this standard may be assessed. Students will be asked to use knowledge gained from looking at one format and integrate that with the knowledge they have gained from another format to demonstrate their understanding of both. For example, information may come from an expository selection related to information in another selection using a very different format and structure such as a graph. Items are really encouraged for this standard.

Standard 8: This standard may only be assessed using single texts. It may not be aligned to crossover items. The emphasis here is on an author's arguments or claims made and evidence used to support such claims. Detailing these arguments or claims and assessing the soundness of the support is vital.

Standard 9: This standard may only be assessed using crossover items relating to multiple texts. Asking students to compare or contrast the presentation of the material by two different author's for the same or similar topic is key.

Grade 7 CCSS Reading Standards

Key Ideas and Details

These standards are to be used with a single passage or one specific passage of a pair. They are not to be used with crossover items pertaining to both passages in a pair.

Literature

Standard 2: The focus of this standard is on the determination of a theme or central idea and look at the author's development of this theme or central idea. A second focus is a summation of the text.

Standard 3: The focus is on how parts or elements of the passage relate or interact with one another. Structure of the writing may be assessed here as long as it shows or implies how the interaction of parts is happening. Looking at characters in relationship to other elements such as plot, other characters, setting, etc. will be viable assessments.

Informational

Standard 2: The focus of this standard will be on the determination and analysis of two or more central ideas within a single text. Looking at how these central ideas are developed throughout the entire text is a part of this standard. A summation of the information given is a second focus.

Standard 3: Items aligned to this standard must look at how individuals, ideas, or events are influenced and also look at the interactions happening between the individuals, ideas, or events.

Grade 7 CCSS Reading Standards

Craft and Structure

These standards are to be used with a single passage or one specific passage of a pair. They are not to be used with crossover items pertaining to both passages in a pair. (An exception is standard 6 for informational texts.)

Literature

Standard 4: This standard still assesses word choices and their impacts, but it goes beyond to include more visible employments of sound, repetitions and purposeful uses within poetry and drama as well. Looking beyond figurative language uses to also connotations of words and their effects or impacts on the text is a part of this standard.

Standard 5: The focus of this standard is specifically on the structure of drama or poetry. Asking students to recognize and analyze how the structure impacts or contributes to the meaning is essential.

Standard 6: The focus of this standard is on the author's use and development of points of view of different characters or narrators within a text. Noticing the contrasts and the impacts these contrasts have on the meaning of the text is a key component of this standard.

Informational

Standard 4: Word choice and meanings of words and phrases is a focus for this standard. Determining the impacts made by the author's word choices or uses of connotative or figurative meanings to the entire meaning of the text is vital.

Standard 5: Structure and organization of the text and the way author's create cohesiveness between parts of the text are aligned to this standard.

Standard 6: Because of the wording of this standard, it will be used only with crossover items (multiple texts). The focus of the standard is on one text; however, in responding to items in this standard, students are asked to go beyond determining an author's point of view or purpose by distinguishing this point of view or purpose from that of others. Reference to a paired passage will encompass this other point of view or purpose.

Grade 7 CCSS Reading Standards

Integration of Knowledge and Ideas

Literature

Standard 7: This standard will be difficult if it's possible at all for which to create items and **definitely must be used with crossover items as it involves multiple texts or media formats.** The difficulty with this standard is that the comparison/contrast must between the written passage and another type of formatted presentation. It is not appropriate to reference a non-print format hypothetically as this is artificial.

Standard 8: The standard 8 is not applicable for literature.

Standard 9: This standard must be used with cross over items dealing with multiple texts. The focus of this standard is the comparison and/or contrast between a fictitious account of a real event, place, situation, character happening historically. Items that ask students to make the comparisons based on how the author of the one has used or altered the reality will align to this standard.

Informational

Standard 7: This standard must be used with crossover items (multiple texts); however, just as with the literary standard 7, it will be a difficult standard if it's possible at all for which to create items. The point of comparison must be with the written text and another non-print presentation or delivery which is not appropriate for a paper/pencil assessment. References made to hypothetical speeches, audio or video presentations are considered artificial.

Standard 8: This standard is only used with single passages and will not be used with crossover items. The focus here is on the evolution of the author's arguments or claims. Items asking students to look at the reasons and evidence used by the author and noting how these are developed throughout are viable. A second focus is to ascertain the credibility or soundness of the arguments or claims being made.

Standard 9: This standard must be used only with crossover items (multiple texts). The focus of this standard is to look at how the authors of paired texts treat similar topics and ideas. Asking students to note the distinctions and similarities and points of emphasis are viable assessments for this standard.

Grade 8 CCSS Reading Standards

Key Ideas and Details

These standards are to be used with a single passage or a specific passage of a pair. They are not to be used with crossover items pertaining to both passages in a pair.

Literature

Standard 2: There are two focuses for this standard – either of which can be assessed. First of all, the determination of the theme or central idea from looking at its development throughout the text is a necessary part. Items could ask students to look at how the theme is related to or demonstrated by the characters, settings, plots, etc. Putting the focus on the characters, settings, plots and then from this focus asking students to determine the theme would be another approach. An additional focus of this standard is on summation of the text.

Standard 3: The focus of this standard is on specific parts – lines, dialogue, actions – and showing how these parts are used to move the action, motivate characters or reveal their feelings or motivations as well as instigating or provoking a decision or result.

Informational

Standard 2: The focus here is on the central idea and how it is supported within the passage. Items that approach this from the theme first then the support or vice versa are viable. Another focus is a summation of the key points.

Standard 3: Although this standard is to be used only with single texts, within the text the focus is on connections between individuals, ideas, or events.

Grade 8 CCSS Reading Standards

Craft and Structure

Literature

Standard 4: This standard will be used only with crossover items (multiple texts) with the emphasis on analogous relationships of words, tone, or allusions made in paired texts.

Standard 5: This standard will be used only with crossover items for multiple texts. The focus in this standard of on comparison of the structure used in multiple texts regarding the same topic. The effect of the structure differences on the meaning of the texts is important.

Standard 6: This standard can only be used with single passages. This standard deals with contrasting points of view of characters within a single text and the reading audience and how these differences impact the passage.

Informational

Standard 4: As in its literary counterpart, this standard will be used only with crossover items (multiple texts). The use of analogies or allusions created within paired passages will be viable material to use for item development to assess this standard.

Standard 5: This standard applies to a single text only. The emphasis here is on sentences or paragraphs within a single text and how these parts clarify, refine, or develop key concepts.

Standard 6: This standard must only be used with multiple texts with crossover items. The focus is on the determination of the author's point of view or intent/purpose and relate this to conflicting points of view or purposes which indicates allusions or references to other texts.

Grade 8 CCSS Reading Standards

Integration of Knowledge and Ideas

Literature

Standard 7: This standard is not applicable to a paper/pencil assessment. The focus of this standard is on filmed or live productions which are not possible to use as a comparison for a paper/pencil assessment. Hypothetical references to such productions or films is not appropriate.

Standard 8: This standard is not applicable to literature.

Standard 9: This standard can only be used with multiple texts with crossover items. The focus of this standard is on the comparison of themes, events, patterns, or characters in the passages draw on other works, such as myths, traditional stories. An emphasis is also on how the traditional characters, themes, etc. are altered or modified in the comparative text.

Informational

Standard 7: This standard must be used only with crossover items (multiple texts). The emphasis here is on the evaluation of how different mediums to present information works. Multi-media modes cannot be used; however, various forms of print material are viable material for crossover items. Presentations of information in graphs, or charts may be analyzed as to their effectiveness.

Standard 8: This standard must be used with single texts only. The emphasis here is on an author's arguments or claims made and evidence used to support such claims. Detailing these arguments or claims and assessing the soundness of the support is vital.

Standard 9: This standard may only be used with crossover items for multiple texts. The emphasis in this standard is on a comparison of multiple texts showing conflicting information on same topic. Asking students to analyze this comparison and also recognize where and how authors disagree is a part of the emphasis.

CCSS Vocabulary Standards

Vocabulary Use and Acquisition

There are only two vocabulary standards in the Vocabulary Use and Acquisition cluster to which items will be aligned on the K-Prep reading assessment for grades 3 - 8. These two standards for every grade are Standard 4 and Standard 5. These standards will be recorded as such: L.X.Xx (e.g., a vocabulary item in grade 5 that makes use of affixes to determine meaning would be coded as L.5.4b.)

Standard 4: This standard for each grade involves determining and clarifying meanings for unknown or multiple-meaning words and phrases. The standard is broken down into specific parts designated by alpha characters. The focus of each of these parts is clearly delineated and items created must adhere to these focuses.

Standard 5: This standard focuses more on the word relationships and variances/nuances of meaning. This will include figurative language and connotative meanings. Again each alpha part explicitly spells out the particular focus to which items must adhere.

NOTE: You MUST read the grade specific standard breakdowns as there are distinct nuances between the grades to which items must adhere!

CCSS

In your item writer materials, we have included PDFs of the Common Core State Standards for Reading for grades 3 – 8. It is essential that you print these off and have them visible when you are creating your items.

You can also find these standards online - http://www.corestandards.org

In addition, there is valuable information as well as exemplars given in both Appendices A and C. It would benefit you to spend time reading these and familiarizing yourself with all of the common core.

Depth of Knowledge

Depth of Knowledge (DOK) labels the difficulty level of the items relative to the cognitive functions or steps students must go through in order to respond correctly to the item. The DOK is adapted from the model used by Dr. Norman Webb, University of Wisconsin, to align standards to assessment tools.

- **DOK 1:** Recall and reproduction – recalling or locating information that is within the text.
- **DOK 2:** Skills and Concepts – involves interpretation, inferences, identification of patterns, classifications, predictions, etc.
- **DOK 3:** Strategic Thinking – development of an argument, critiquing, comparing and contrasting, drawing in-depth conclusions, connecting ideas to explain
- **DOK 4:** Extended Thinking – analyzing, synthesizing information

NOTE:
DOK 4 often necessitates use of extended periods of time which often makes this level of task inappropriate for a paper/pencil assessment in a finite time frame. **You will have no DOK 4 items.** The extended response items should be at the DOK 3 level.

The majority of the items on an assessment should be in the DOK 2 and DOK 3 ranges.

Item Types

2

PEARSON

Multiple Choice Items

When you are creating multiple choice items, please keep the following in mind:

- Options are created homogeneously.
- Single word options may sometimes be used although often are more accessible when enveloped in appropriate text.
- Options may require reader discernment, but cannot be purposely tricky. The rule is “may be difficult, but must be fair.”
- Avoid clang or cueing between options, stem and options, and item to item.
- Item stems should use **standard specific language** whenever possible.
- The words selected for vocabulary acquisition/use items are generally above grade level or are used in less common ways.
- Context must be found within the text for all vocabulary items.
- Items must cover the entire passage; avoid creating several items all concentrated on a single paragraph while leaving the rest of the passage without any item coverage.
- Answer options that are complete sentence must be punctuated accordingly.
- Answer options that are not complete sentences used with both open and closed stems should begin with lowercase.
- Options for open stems must use lowercase with periods at the end of each since they are completing the sentences.

Guidelines

Options

1. Each multiple choice (MC) item will have four answer options.
2. The key is the correct option of the four given.
3. Each distractor (three options that are not the correct answer) must be plausible but incorrect based on the context of the passage.
4. There must be **only one** clearly correct answer.
5. All options must use parallel construction in language, length, format, and specificity.
6. Answer options must be balanced across the set of items for a given passage.
7. No option, including the key, should visibly stand out from the rest.

Options to Avoid

Avoid

- "all of the above" and "none of the above."
- options that are subsets of another.
  (e.g., Option A. High school
  Option B. Grades 7 – 10)
- absolutes (e.g., "all," "none").
- outliers – options that are too obviously incorrect or do not have any relevance to the associated passage.
- options that essentially the same with slight wording variations.
- options that introduce brand new information.
- options that require knowledge beyond what is being assessed.
- options that use the passive voice.
- options that rely heavily on visual or auditory acuity.

Multiple Choice Rationales

Rationales will be written for each of the options in a multiple choice item. The rationales are of primary importance to you as item writers. These rationales are justifications for the item's viability.

- The rationale for the correct response will begin with "Correct:" followed by the reason it is the correct response.
- Rationales for the distracters must identify the plausibility of the option as well as indicating what makes it incorrect in the context of the passage.
- If you as the writer experience difficulties coming up with the justifications for each option, then that is your clue to rework the distracters or possibly change the item.

Short Answer Items

Short answer items will earn a maximum of two points. Rubrics must be provided that are specific to the item as part of the item creation. A short answer item is able to align to most all of the standards.

The rubrics for a short answer must be written for the maximum 2 points, 1 point, 0 points, and Blank. A text description of each must be given for each score point.

Sample Short Answer

Sample stem:

Which information in the passage shows why _____?
Explain how _____?

(The response to this question must have at least two parts or aspects that students may include.)

Sample rubric:

2 Points: Student completes all parts of the question and communicates ideas clearly. Student demonstrates an understanding of the concepts and/or processes. Student provides a correct answer using an accurate explanation as support.
1 Point: Student provides a partially correct answer to the question and/or addresses only a portion of the question. Student demonstrates a partial understanding of the concepts and/or processes.
0 Points: Student's response is totally incorrect or irrelevant.
Blank: Student did give any response at all.

Short Answer with Pair

If there is a paired passage, a short answer could possibly ask students to compare a concept/explanation/description across the two passages.

Sample:

In both passages, the main character interacts with his/her teacher. How was the interaction between the main character and his/her teacher different in each of the passages?

The rubrics need to capture the aspects of a complete response and then differentiate between all score points.

Extended Response Items

Grade 3 does not have extended response items. There will be one extended response written for each paired passage in grades 4 – 8. the extended response (ER) items are worth four points. These items **must** align to the **Integration of Knowledge and Ideas** standard of the CCSS, so can only be crossover items in a paired passage set.

The following slide shows a released ER item for the Kentucky Reading Assessment.

Sample Extended Response

Stem:

The last sentence of the passage says that Birbal “went to tell the barber the good news, with a little smile tugging at the corners of his mouth.”

a. Describe the good news Birbal was going to tell the barber.

b. Explain TWO reasons why Birbal might have been smiling.

Use examples from the passage to support your answer.

Scoring Guide (Rubric) :

4 Points: Student clearly describes the good news Birbal was going to tell the barber. Student clearly explains **two** reasons why Birbal might have been smiling. Response is supported with examples from the passage.

3 Points: Student generally describes the good news Birbal was going to tell the barber. Student generally explains **two** reasons why Birbal might have been smiling. Response is supported with examples from the passage.

Scoring Guide Continued
Rubric

2 Points: Student provides a limited description of the good news Birbal was going to tell the barber. Student provides a limited explanation of two reasons why Birbal might have been smiling. Response may be supported with few or no examples from the passage.

OR

Student provides a general explanation of the good news Birbal was going to tell the barber. **Part b** is missing or incorrect.

OR

Student provides a general explanation of two reasons why Birbal might have been smiling. **Part a** is missing or incorrect.

1 Point: Student demonstrates minimal understanding (e.g., student provides a limited description of the good news Birbal was going to tell the barber **or** a limited explanation of one reason why Birbal might have been smiling.)

0 Points: Student's response is totally incorrect or irrelevant.

Blank: No student response.

Universal Design Considerations

All items, regardless of type, must be accessible to all students. The items must NOT unfairly advantage or disadvantage any groups or populations of students.

All items must, therefore,

- respect the diversity of the assessment population.
- use text that avoids emphasis on visual acuity.
- are accessible to all (age, gender, ethnicity, personal limitations, socio-economic level, English language learners).
- avoid unnecessary repetition or word clutter.
- refrain from using multiple meaning words unless that is what is being assessed and for which context has been provided.
- avoid use of regionalisms, colloquial expressions, idioms unless context is provided.

Important General Reminders

1. It is very important that all items are of the highest quality possible. Therefore, we suggest **leaving enough time in your item writing to allow yourselves to step away from the items you've written for a day or so, then take another careful look at them.** This distance allows you to see things that may be elusive during the first working session regardless of the amount of times and effort spent initially.

2. After writing your set of items for a passage, go through the passage and check off paragraphs/sentences that are associated with items. This will help you to see the passage coverage your item set has.

3. **Check to make certain the items you've created adhere strictly to the standard alignments assigned and that the alignments are accurate.** The items must truly align and assess what the standard is requesting.

4. Make sure that you look at the items you've written for a passage all together. This will be a good check to see if you have overlap or clueing within an item set.

Submission Requirements and Contact Information

3

PEARSON

Submission Requirements

- Items must be submitted according to the schedule and specifications detailed on your assignment and Statement of Work.
- Item writers will receive assignments and feedback via the SFTP site.
- Pearson content specialists will notify item writers when assignments or feedback is posted.
- Item writers will email Pearson, copying all contact names listed, with confirmation that correspondence has been received.

NOTE: Never send passage titles or item specifics via email!

Requirements continued

- **Adhere to the schedules and meet all deadlines!**

  You will receive your assignments and feedback from any of the content specialists working on the Kentucky project. Whenever questions, comments, concerns or issues arise, please send an email immediately to the content specialist working directly with you and CC the other content specialists. Remember to NOT include anything confidentially specific to the passages or items in your email. Phone calls may be a more immediate and safer route to go!
- Post assignments on the SFTP site in the Kentucky/To Pearson folder
- Notify all content specialists immediately upon completion of the postings.

  It is imperative for all of us working on the Kentucky project to be kept abreast of the progress and process flow of these items.

Appendix B. Mathematics Item Writer Training

Pearson Item Writer's Training for the Kentucky K-Prep Field Test Items - *Mathematics*

Objectives

- To review Statement of Work and schedule of item submission
- To explain item writer responsibilities
- To familiarize item writers with the requirements of Pearson item writing guidelines for the Kentucky K-Prep Field Test Items
- To review how to submit items to Pearson
- To answer questions

Review of Statement of Work and Schedule of Item Delivery

- Review/Sign/Return needed documents to KY Project Manager (PjM Margot Hetrick)
- Review writing assignment
  - Contact PjM about contract questions and obligations.
  - Contact Math content specialist for clarifications concerning standards, item type, or other content related questions.
- Schedule of item delivery
  - Please submit half of the items ahead of schedule if time allows.
  - Send Pearson content specialist notification of possible delays via email or phone.

Item Writer Responsibilities

- Confidentiality
  - Item writers must not copy, discuss, or disclose in any manner the information or materials used during training, while writing items, or after the assignment has been completed.
- 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.

Item Writer Responsibilities

- Ownership
  - All materials developed for the Kentucky K-Prep Field Test must be original and may not appear in any other source. They are the property of the state of KY and may not be used for any other purpose.
- Schedule
  - Item writers must submit assignments according to the schedule and specifications detailed on the Statement of Work.
- Source Documentation
  - Item writers must provide source documentation for factual information used in the contexts of the items.

Kentucky K-Prep Field Test Items

- Written to the Common Core State Standards (CCSS)
  - Common Core Standards
    http://www.corestandards.org/the-standards/mathematics

- Depth of Knowledge
  - All MC (multiple choice) items should be written to a DOK of 2 or higher
  - All SA (short answer) items should be written to a DOK of 2 or higher
  - All ER (extended response) items should be written to a DOK of 3
  - KY document for Depth of Knowledge (DOK) levels provided in Item Writer materials

Common Core State Standards (CCSS)

- Items will be written to the common core.

- How to read the grade level standards
  - Grade level
  - Domain: larger groups of related standards
  - Cluster: groups of related standards
  - Standard: what students should understand and be able to do

- CSS standard to write to
  - Standards have been selected for each item type assigned
  - SA and ER items may be written to a cluster

Common Core State Standards (CCSS)

An overview of the progression of the standards:

Domain	Abbrev.	Grade Level								
		K	1	2	3	4	5	6	7	8
Counting and Cardinality	CC	X								
Operations and Algebraic Thinking	OA	X	X	X	X	X	X			
Number and Operations in Base Ten	NBT	X	X	X	X	X	X			
Number and Operations-Fractions	NF				X	X	X			
Measurement and Data	MD	X	X	X	X	X	X			
Geometry	G	X	X	X	X	X	X	X	X	X
Ratios and Proportional Relationships	RP							X	X	
The Number System	NS							X	X	X
Expressions and Equations	EE							X	X	X
Statistics and Probability	SP							X	X	X
Functions	F									X

Common Core State Standards (CCSS)

- Please refer to the common core standards below and above your grade level for a better understanding of what is expected at your given grade.
- Standards are a balanced combination of procedure and understanding.
- Standards that start with the word “understanding” provide an opportunity to connect the practices to the content.
- In your item writer materials, we have included PDFs of the Common Core State Standards for Mathematics. It is essential that you print these off and have them visible when you are creating your items.

Useful links to Common Core State Standards (CCSS) additional information

- http://illustrativemathematics.org/standards/k8 Click on the right side to “Show Only Illustrated Standards.” Not all standards are illustrated yet, but there are some illustrations in each of the grades for which we are writing items. These provide great insight into the standards.
- http://math.arizona.edu/~wmc/ Click on Common Core State Standards in Mathematics under Recent Presentations to gain insight into grade 3 fractions and grade 6 statistics and probability on pages 9 through 13.
- http://ime.math.arizona.edu/progressions/ There are progressions at the bottom of the page helpful for understanding the intent of the standard.
- http://commoncoretools.me/ More CCSS resources.
- http://illuminations.nctm.org/ NCTM materials to understand the standards and write cognitively demanding items.

Depth of Knowledge

Depth of Knowledge (DOK) labels the difficulty level of the items, taking into consideration the cognitive functions or steps students must go through to respond correctly to the item. The DOK is adapted from the model used by Norman Webb, University of Wisconsin, to align standards to assessment tools.

- DOK 1: Recall and reproduction – recalling or locating information that is within the text.
- DOK 2: Skills and Concepts – involves interpretation, inferences, identification of patterns, classifications, predictions, etc.
- DOK 3: Strategic Thinking – development of an argument, critiquing, comparing and contrasting, drawing in-depth conclusions, connecting ideas to explain
- DOK 4: Extended Thinking – analyzing, synthesizing information

DOK 4 often necessitates use of extended periods of time, which often makes this level of task inappropriate for a paper/pencil assessment in a finite time frame. The extended response items should be at the DOK 3 level.

The majority of the items on an assessment will be in the DOK 2 and DOK 3 ranges. Please see DOK_SUPPORT_Mathematics document.

Style Guide

- No open-ended stems.
- Align numbers in options by decimals when applicable.
- All graphs and charts should have a title and appropriate labels for the horizontal and vertical axes.
- When referring to specific figures (parallelograms, circles, points, etc.), capitalize the term and italicize the term's letter
  - Point *Q,* Circle *R*, Triangle *ABC*, etc…
- Avoid answer options that are the same as the option letters.
- Try to avoid using the negative sense "not" and "never."
- Use "number cube" instead of "dice.'
- When unit of measurement labels are given in the stem, labels are not repeated in the answer options.
- Avoid using common brand names such as Kleenex, Coke, Jell-O, etc… Instead use tissue, soda, gelatin dessert, etc…

Style Guide continued

- Use “Which” for questions that have multiple potential answers and use “What” for questions that have only a single possible answer.
- Use the word “percentage” as a general term and use “percent” when asking a mathematical question or when accompanied by a number.
  - Percentage usage
    - A large percentage of students prefer math to P.E.
  - Percent usage
    - She expected gas prices to rise by 10 percent.
- When amounts greater than or equal to $1.00 and amounts less than $1.00 are used in an item, all amounts should be written with $ instead of mixing $ and ¢.
- Format when art is used in an item should be:
  - Stem/Art/Stem

General Considerations in Item Writing

- Be creative and innovative; try to approach the assessment of the standard in an interesting way. Do not overuse textbook style questions and contexts.
- All items should be unique in their approach to the standard.
- Make sure the item really does assess the standard listed.
- Item should be clear and unambiguous.
- Item should be mathematically and grammatically correct.
- Item should be able to be answered correctly by the students who have mastered the content of the standard being assessed.
- Use real world context where applicable.
- Use the grade below and above your grade level for assessment limits.

Item Writer Guidelines – Answer Options

- Include 4 answer options with 1 correct key or best answer
- Arrange numerical answer choices in either ascending or descending value.
- Use plausible distractors that demonstrate different common student mistakes or misconceptions that fit the standard being assessed.
- Options should be parallel in language, length, format, and specificity.
- Keys should be fairly balanced across items in an assignment.
- Avoid using intermediate steps as distractors

Item Writer Guidelines – Answer Options Things to Avoid

- Avoid “all,” “ none of the above,” and “not enough information”
- Avoid an option that includes all other options
- Avoid absolutes
- Avoid outliers
- Avoid options that mean the same thing
- Avoid content that requires knowledge other than what is being assessed

Item Writing Guidelines – Item Format

- Templates
  - Use the templates posted on the SFTP site. Templates are numbered and separated by item type.
  - **Most** metadata has been filled in for you.
    - Confirm the DOK
    - SA and ER items may have a secondary standard
- Use MathType or Equation Editor to write all math expressions, equations, symbols, etc.
- Submitting items
  - Make sure you save all templates in MS Word 97-2003.
    - For example: M5099.doc
  - Post all items to the SFTP site. Do NOT email items or art.
  - Send an email to Olga Garza and Lillian Butcher notifying us that you posted your items.
  - Pearson content specialist will send notification a week prior to your due date as a reminder.

Item Writer Guidelines - Rationales

- Include a rationale for each Multiple Choice answer choice to show how the student would have arrived at the answer choice. Intermediate steps should be avoided as distractors.
- For the correct response, begin the rationale with “Correct”
  - For example:
    - Correct. $d\pi = (5)\pi \approx 15.708$
- Avoid including incorrect math statements in the rationale
  - For example:
    - $2 + 4 = 6 - 5 =$
- Avoid using the following rationales.
  - For example:
    - Unfamiliar with concept
    - Incorrect answer

Item Writer Guidelines - Rationales

- Write rationales in past tense; phrases or complete sentences are acceptable.
  - For example:
    - Used the radius instead of the diameter.
    - Found the reciprocal of the slope and used the correct y-intercept.
- Rationales may show calculations instead of, or in addition to, a verbal description to show the method to find the answer.
  - For example:
    - Used 3 instead of 6 as the diameter. $(3)\pi$
    - Correct. $3.14(6^2)(5)$
- Avoid using “The student thought…”

Item Writer Guidelines – Short Answer (SA) Rubrics

- An SA item presents a scenario and directs the student to perform a task related to the scenario.
- Students are usually directed to show work and/or explain how they determined their answers.
- A typical SA item should take the student approximately 3 to 5 minutes to complete.
- Each SA item is worth 2 points, and students can earn scores of 0, 1, or 2 points

Item Writer Guidelines – Extended Response (ER) Rubrics

- An ER item presents a scenario and directs the student to perform tasks related to the scenario.
- Students are usually directed to show work and/or explain how they determined their answers.
- A typical ER item should take the student approximately 10 to 15 minutes to complete.
- Each ER item is worth 4 points, and students can earn scores of 0, 1, 2, 3, or 4 points.

Item Writer Guidelines – SA and ER Rubrics

- As much as possible, items should be written so that students can approach the tasks in more than one way.
- At least one example of a top score response must be provided.
- A brief descriptor of responses that might earn each other possible score point must be provided.
- Students write their answers to SA and ER items on grid paper.
  - See Rubric examples handout
    - Half a page for SA type items
    - Full page for ER type items
  - Please take into consideration the space allowed for a student to answer SA and ER items.

Item Writer Guidelines - Art

- Identify the location of the art in the stem or options by using the UIN.
  - For example:
    - In the stem use the tag <InsertART_M5099_1>
    - In option A use the tag <InsertART_M5099_A>, etc…
- Supply a drawing of the art.
  - Art can be drawn using a software program or hand-drawn, scanned, and placed on the correct art page of the item template.
  - Or, art can be hand-drawn and faxed to 210-339-5976 to the attention of either Olga Garza or Lillian Butcher.
  - Please make sure to include your name and UIN.

Item Writing Guidelines – Universal Design

- Respects the diversity of the assessment population
- Uses clear format for text, clear pictures and graphics, including only essential illustrations
- Uses appropriate grade-level vocabulary
- Is accessible to all test-takers (age, gender, ethnicity, disability, socio-economic level, English language learners, students with disabilities)
- Minimizes skills required beyond those being measured
- Avoids unnecessary word clutter and idioms
- Avoids content that might unfairly advantage or disadvantage any student subgroup
- Be cognizant of student names – Student's names included in items should not distract from the problem or task.

Item Review Checklist

1. Does this item ask something worth asking?
2. Has the intent of the standard been clearly assessed?
3. Is this item unnecessarily easy or difficult?
4. Is this item biased?
5. Is this item free of sensitive, emotionally charged issues?
6. Does this item require background knowledge?
7. Does this item break item-writing guidelines?
8. Does the item have only one correct answer?
9. Does the item measure what it is intended to measure?
10. Is the Depth of Knowledge level appropriate for the level of thinking skill required?
11. Is the item straightforward and direct with no unnecessary wordiness?

Item Review Checklist (continued)

12. Are there any clues or clang words used which may influence the student's responses to this item or other items?
13. *Is the intent of the question apparent and understandable to the student without having to read the answer options?*
14. Do all items function independently?
15. Are all items grammatically and mathematically correct and in complete sentences whenever possible?
16. Read the item aloud, slowly. Do all sentences make sense?
17. *Are all distractors plausible yet incorrect?*
18. Do answer choices only appear in the answer options and not in the stem?
19. Have you looked at the footnotes to make sure the item is grade level appropriate with the grade level restrictions?

Submission Requirements

- Items must be submitted according to the schedule and specifications detailed on your assignment and Statement of Work.
- Item writers will receive assignments and feedback via the SFTP site.
- Pearson content specialists will notify item writers when an assignment or feedback is posted on the SFTP site in the Kentucky/From Pearson folder.
- Item writers will email Pearson, copying all contact names listed, with confirmation that correspondence has been received.

NOTE: Never send item specifics via email.

Requirements continued

- Adhere to the schedules and meet all deadlines.
- You will receive your assignments and feedback from any of the content specialists working on the Kentucky project.
- Whenever questions, comments, concerns or issues arise, please send an email immediately to the content specialist working directly with you and CC the other content specialist. Remember to NOT include anything confidentially specific to items in your email. Phone calls may be a more immediate and safer route to go.
- Post assignments on the SFTP site in the Kentucky/To Pearson folder.
- Notify all content specialists immediately upon completion of the postings.

Files for Item Writing Reference on SFTP site:

- Item Writer Training Presentation PowerPoint
- Item Writing Assignment
- Templates
  - Type (MC, SA, and ER)
  - Metadata
    - Standard
    - DOK (Verify the DOK as each template has a “2” listed.)
    - Calculator
- Common Core Standards and Depth of Knowledge documents
- Ruler(s) – Grade 3, Grade 4 – 6, and/or Grade 7 & 8
- Reference sheet for grades 7 & 8
- Rubric examples

Appendix C. Item Development Review Checklist

ITEM REVIEW CHECKLIST

Individual items:

1. Does the item ask something that is worth asking?
2. Is the item unnecessarily easy or difficult? If it is difficult, is it fair (devoid of anything purposely tricky?)
3. Is the item free of any sort of bias or emotionally-charged issues?
4. Can the item be answered using the information provided within the passage without requiring background or prior knowledge?
5. Does the DOK level appropriately indicate the item's difficulty?
6. Does the item truly assess what it purports to assess (align with the standard)?
7. Is the item free of any cue or clang between the stem and the options?
8. Is there only one clearly correct answer?
9. Are all of the distractors plausible and yet identifiably incorrect?
10. Does the stem make it clear to the students what they are to answer? (Do students have to read all of the options before their task is apparent?)
11. Does the item function independently of all others?
12. Is the item clearly and concisely stated using correct usage, grammar and mechanics?

Complete item set:

13. Do all items within an item set for each passage offer a variety of alignments to different standards?
14. Is there any cueing or clueing between items within an item set?
15. Does the entire item set offer a good and complete coverage of the passage?

Appendix D. Reading Content Committee Review


Kentucky Reading
Content Advisory
Committee
Welcome!

Kentucky Performance Rating for Educational Progress (K-PREP)

Pearson provides assessments for grades 3-8 and writing on-demand at high school.

Pearson Kentucky Program Team: An Overview

- Housekeeping
- Committee Appropriate Materials
- Nondisclosure Agreements
- Expense Reimbursement Form
- Substitute Teacher Payment Form
- MS PowerPoint Presentation

Large-Scale Assessment

- **Standardized testing for thousands of students**

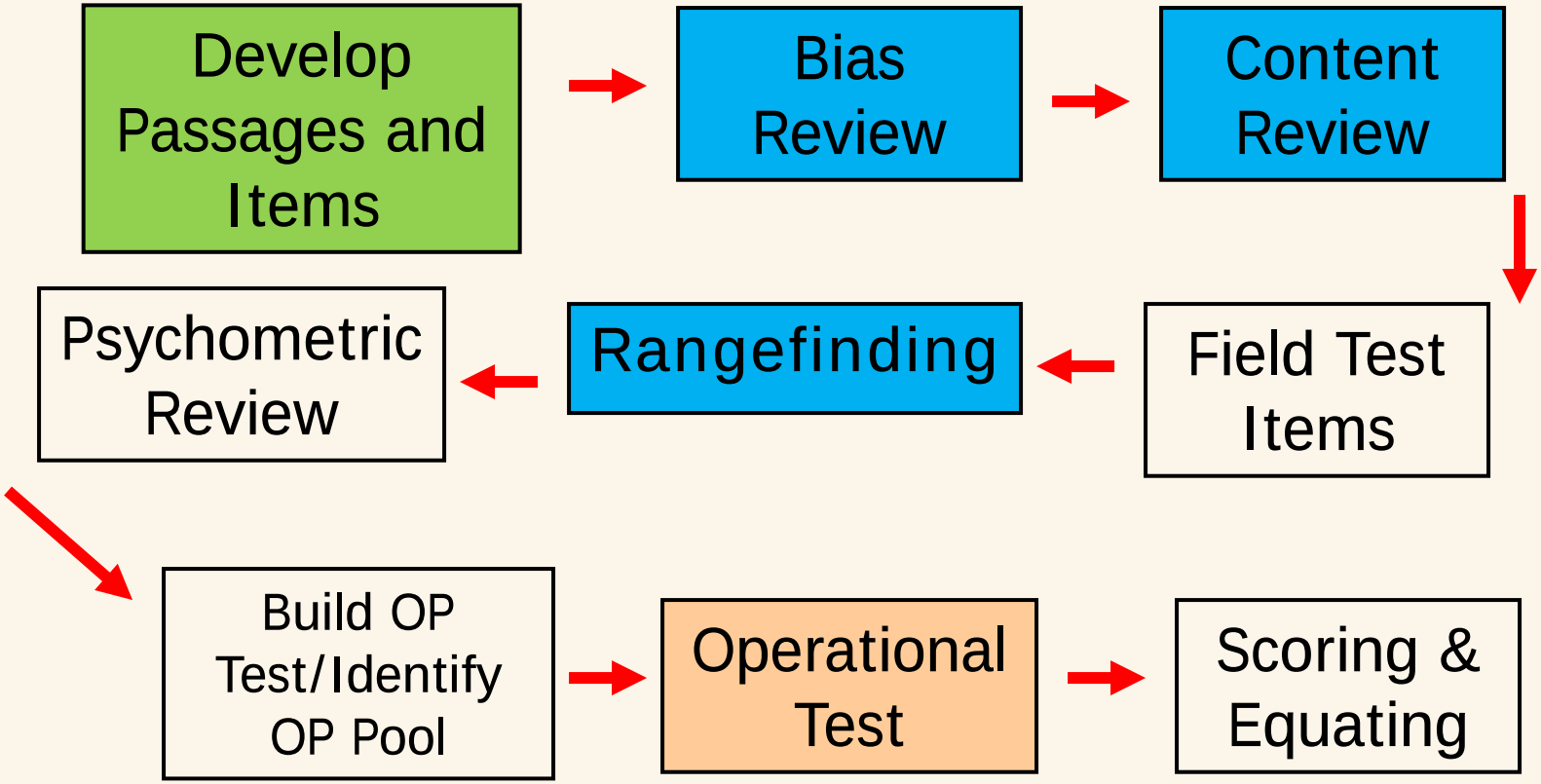
- **Designed to measure student knowledge and skills against pre-determined standards (Kentucky Core Academic Standards-KCAS)**

  - **Reading, Writing, Mathematics**
  - **Science, Social Studies**

Development Process

- Passage and Prompt Writer Training
- Technical Content Review
- Art Development
- Universal Design Review
- Fact Checking
- Copy Editing
- KDE Review
- Publication/Formatting

Test Development Process



Confidentiality

Certain measures are required for security purposes:

- **NO discussion with home districts or anyone else about items.**
- **NO "reproducing" of passages, prompts or items verbally, electronically or hard copy.**
- **NO cell phone use in the review rooms.**

Confidentiality is really about ensuring equity for all Kentucky students.

Housekeeping Reminder

- **Sign-in Sheet – Has everyone signed in?**
- **Blue folder Materials include:**
  - ✓Agenda
  - ✓Nondisclosure Agreement
  - ✓Expense Forms
  - ✓Related Committee Materials
- **Breakout locations**
- **Restrooms**

Common Core State Standards (CCSS)

The K-PREP will be assessing students' reading comprehension in both literature and informational genres and students' vocabulary acquisition and use.

The standards to which these items will align are the Common Core State Standards. There are four clusters for which each of the genres are divided.

CCSS Reading Standards

The standard expectations differ from grade to grade; however, for all grades the standards are divided into strands for each of the genres (literature and informational).

The strands being assessed on the K-PREP reading are the Key Ideas and Details, Craft and Structure, Integration of Knowledge and Ideas, and Vocabulary Acquisition and Use.

Cluster 1: Key Ideas and Details

This cluster includes three standards labeled by genre, grade level, then standard number (i.e., RL.3.1). There are three standards within this cluster.

The three standards that fall into this cluster cover theme, main ideas, supporting details, inferences, summaries, and the development of plot, character, idea and concept.

CCSS Reading Standards

Cluster 2: Craft and Structure

This cluster includes three standards. The standards that fall into this cluster assess students' ability to ascertain meaning of words: literal vs. nonliteral, figurative uses, allusions, connotations vs. denotations, and nuances.

The standards in this cluster also involve text type, distinctions between text types, overall structure and organizational development, author purpose and intent, the interactions and connections of structure to meaning, and point of view as it affects the meaning.

CCSS Reading Standards

Cluster 3: Integration of Knowledge and Ideas

This cluster contains three standards. For literature only one of the standards is assessable on a paper/pencil assessment. Standard 9 requires that the student compare and contrast themes, settings, characters, fiction versus factual either by same author, different texts or different genres.

For informational texts, the three standards assess relationships between illustrations or other graphics and the meaning of the text. The standards also assess students' abilities to identify connections within structure and contrasting views, compare and contrast key ideas, arguments, and different authors' presentations of similar topics, and also delineate arguments and logical reasoning of authors.

CCSS Reading Standards

Cluster 4: Vocabulary Acquisition and Use

This cluster has two standards that are being assessed in Reading. The first of these standards assesses students' abilities to determine or clarify the meaning of unfamiliar or multiple-meaning words using context, affixes, root words, resource materials.

The second standard addresses students' understanding of figurative language uses, word relationships and nuances in word meanings.

Depth of Knowledge

All items are given a DOK (Depth of Knowledge) level based on the number of cognitive functions required of the students to answer the question.

DOK 1: The items at a DOK 1 are generally quite literal. These items can generally be answered by simply going back to the passage and finding the information.

DOK 2: The items at a DOK 2 most generally require that students do some inferring, interpreting, or predicting using the information they have from the reading stimulus.

DOK 3: The items at a DOK 3 require students to do some analyzing and synthesizing of information. Often with paired passages students must make connections between the two passages which involves synthesis and/or drawing conclusions.

(DOK 4 items are generally inappropriate for use on a test given in a definite time frame as this level involves development over an extended period of time.)

The majority of items will be at the DOK 2 level. Generally speaking, for the Extended Response (ER) items and the Short Answer (SA) items, writers strive to keep these at a DOK 3 where possible.

Rationales and Rubrics

Rationales:

All multiple choice items will have rationales supplied for each of the answer options available. These rationales are primarily for the benefit of the item writers themselves. When they are asked to justify the plausibility and the accuracy of each distractor, they are more easily able to see if the item they have created is viable. The rationales attempt to show the plausibility of each option and yet also point out why, in the context of the reading stimulus, the particular distractor is incorrect.

Rubrics:

Scoring rubrics (writing expectations) are provided for all short answer (SA) and extended response (ER) items. Sample answers are not generally provided, but overarching general points of coverage are given to show a distinction between scoring levels.

Item Review Checklist

1. Does the item measure what it is intended to measure?
2. Does the item have only one correct answer?
3. Are all distractors plausible yet incorrect?
4. Check for clues or clang words which may influence the student's responses to other items. Do all items function independently?

Item Review Checklist-Cont.

5. Is the intent of the question apparent and understandable to the student without having to read the answer options?

6. Is the depth of knowledge level appropriate for the level of thinking skill required?

7. Is the item straightforward and direct with no unnecessary wordiness?

Orientation to Participant Materials

- **CCSS Reading Standards**
  - **Grades 3 through 5**
  - **Grades 6 through 8**
- **Depth of Knowledge**
- **Scoring Rubrics – Generic**
- **Item Review Checklist**
- **Bound Book with passages and items**

Organization of Bound Booklet

Organized by Passage, Standard, and Item Types

- **Multiple choice**
- **Short Answer**
- **Extended Response**

Item metadata (information about item)

UIN	KCAS Primary	Item Type	Key	DOK	Passage Title
R4239	L.4.4c	MC	C	2	Fighting the Sea

Ground Rules for Effective Group Work

- **Every opinion is important and valued!**
- **Be polite and respectful – please wait until a speaker has finished before making additional comments.**
- **Please place cell phones on vibrate.**
- **Please hold extended conversations outside the room while others are reading or reviewing.**
- **Please let staff know if you have a request or concern.**

Advisory Committee Purpose

- **Advisory Committees help us prepare and refine passages, items and prompts eligible to appear on the state assessment intended to measure Kentucky's chosen standards.**

- **Participate in thoughtful and meaningful discussions about grade-level content and/or passage, item, prompt appropriateness for Kentucky students.**

Questions?

Appendix E. Mathematics Content Committee Review


Kentucky Mathematics
Content Advisory
Committee
Welcome!

Kentucky Performance Rating for Educational Progress (K-PREP)

Pearson provides assessments for grades 3-8 and writing on-demand at high school.

Pearson Kentucky Program Team: An Overview

- Housekeeping
- Committee Appropriate Materials
- Nondisclosure Agreements
- Expense Reimbursement Form
- Substitute Teacher Payment Form
- MS PowerPoint Presentation

Large-Scale Assessment

- **Standardized testing for thousands of students**

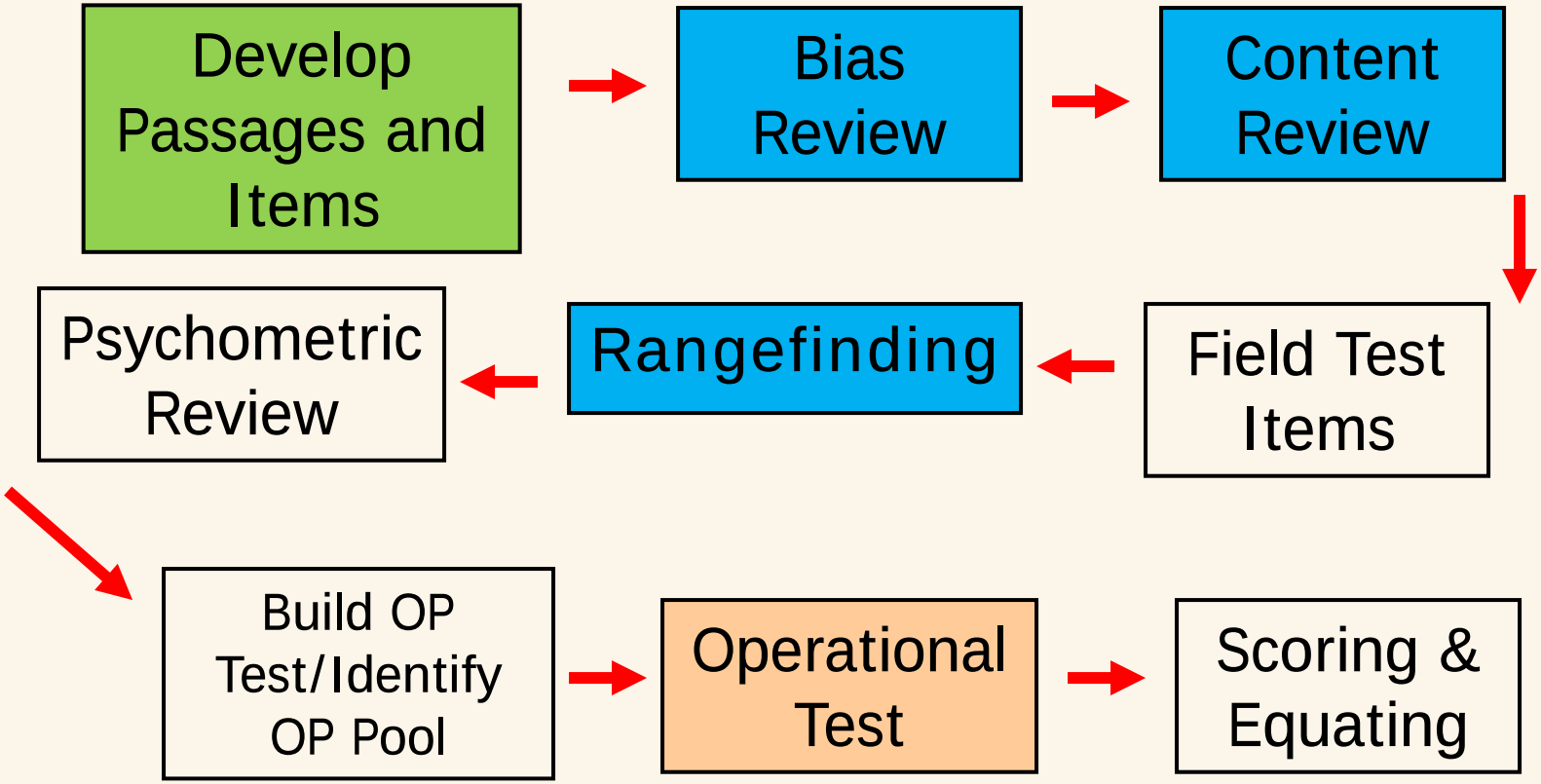
- **Designed to measure student knowledge and skills against pre-determined standards (Kentucky Core Academic Standards-KCAS)**

  - **Reading, Writing, Mathematics**
  - **Science, Social Studies**

Development Process

- Item Writer Training
- Technical Content Review
- Art Development
- Universal Design Review
- Fact Checking
- Copy Editing
- KDE Review
- Publication/Formatting

Test Development Process



Confidentiality

Certain measures are required for security purposes:

- **NO discussion with home districts or anyone else about items.**
- **NO "reproducing" of passages, prompts or items verbally, electronically or hard copy.**
- **NO cell phone use in the review rooms.**

Confidentiality is really about ensuring equity for all Kentucky students.

Housekeeping Reminder

- **Sign-in Sheet – Has everyone signed in?**
- **Blue folder Materials include:**
  - ✓Agenda
  - ✓Nondisclosure Agreement
  - ✓Expense Forms
  - ✓Related Committee Materials
- **Breakout locations**
- **Restrooms**

Common Core State Standards for Mathematics

Grade-Level Standards

- K-8 grade-by-grade standards organized by domain
- 9-12 high school standards organized by conceptual categories

Standards for Mathematical Practice

- Describe mathematical “habits of mind”
- Standards for mathematical proficiency: reasoning, problem solving, modeling, decision making, and engagement
- Connect with content standards in each grade

Common Core State Standards for Mathematics

The K- 8 standards:

The K-5 standards provide students with a solid foundation in *whole numbers, addition, subtraction, multiplication, division, fractions and decimals*

•The 6-8 standards describe robust learning in *geometry*, *algebra*, and *probability and statistics*

•Modeled after the focus of standards from high-performing nations, the standards for grades 7 and 8 include *significant algebra and geometry content*

•Students who have completed 7th grade and mastered the content and skills will be *prepared for algebra, in 8thgrade or after*

Common Core State Standards (CCSS)

An overview of the progression of the standards:

Domain	Abbrev.	Grade Level								
		K	1	2	3	4	5	6	7	8
Counting and Cardinality	CC	X								
Operations and Algebraic Thinking	OA	X	X	X	X	X	X			
Number and Operations in Base Ten	NBT	X	X	X	X	X	X			
Number and Operations-Fractions	NF				X	X	X			
Measurement and Data	MD	X	X	X	X	X	X			
Geometry	G	X	X	X	X	X	X	X	X	X
Ratios and Proportional Relationships	RP							X	X	
The Number System	NS							X	X	X
Expressions and Equations	EE							X	X	X
Statistics and Probability	SP							X	X	X
Functions	F									X

Participant Materials

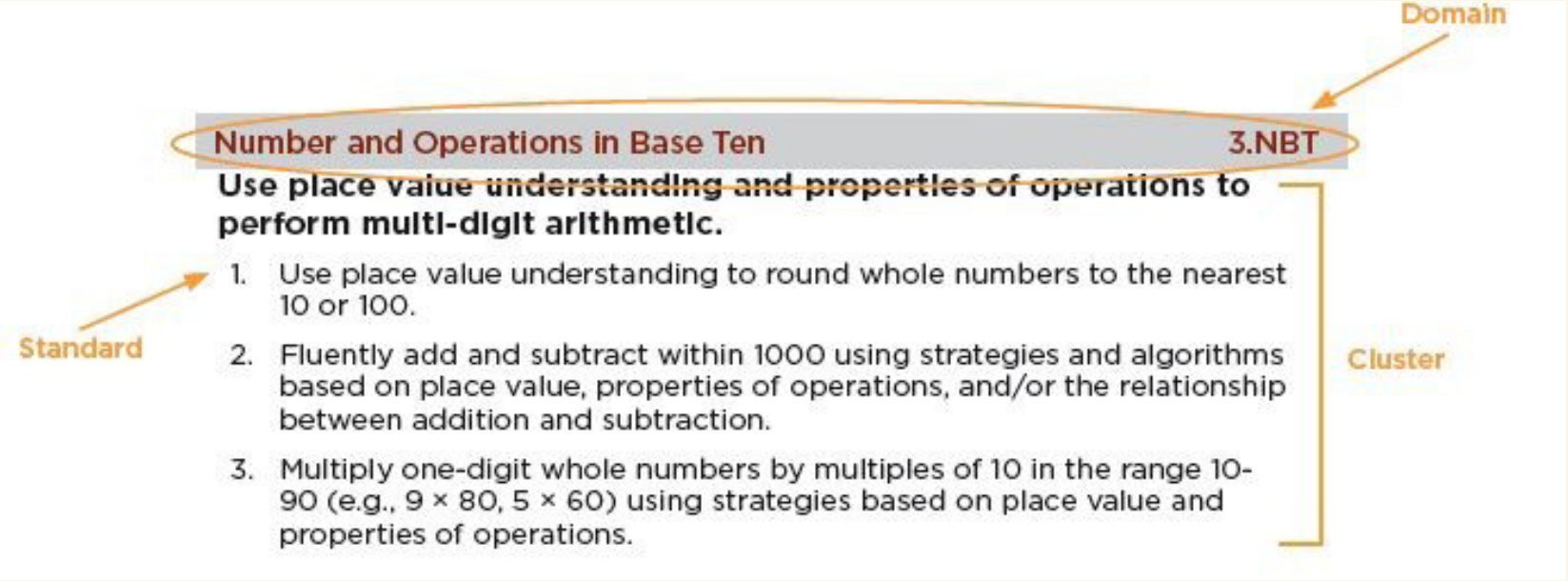
- Common Core Standards
- Depth of Knowledge
- Reference Sheet – Grades 7 & 8 only
- Scoring Rubrics – Generic
- Rulers
- Item Review Checklist
- Bound Book with items

Common Core Standards for Mathematics

Domains: overarching ideas that connect topics across the grades

Clusters: illustrate progression of increasing complexity from grade to grade

Standards: define what students should know and be able to do at each grade level



Organization of Bound Booklet

Organized by Standard and Item Type

- Item Types:
  - Multiple Choice (MC)
  - Short Answer (SA)
  - Extended Response (ER)

Item metadata (information about Item)

UIN	KCAS Primary	KCAS Secondary	Item Type	Key	DOK	Calculator
M5076	5.G.2		MC	2	1	NC

Item Types Developed

Multiple Choice (MC)

Short Answer (SA)

Extended Response (ER)

Multiple Choice (MC) Item

An MC item presents a question in which students have four various solutions from which to chose.

Students are provided four options; the correct answer and 3 plausible but incorrect options.

A typical MC item should take the student approximately 1 minute to complete.

Each MC item is worth 1 point and students can earn 0 or 1 point.

Short Answer (SA) Item and Rubric

An SA item presents a scenario and directs the student to perform a task related to the scenario.

Students are usually directed to show work and/or explain how they determined their answers.

A typical SA item should take the student approximately 3 to 5 minutes to complete.

Each SA item is worth 2 points, and students can earn scores of 0, 1, or 2 points

Extended Response (ER) Item and Rubric

An ER item presents a scenario and directs the student to perform tasks related to the scenario.

Students are usually directed to show work and/or explain how they determined their answers.

A typical ER item should take the student approximately 10 to 15 minutes to complete.

Each ER item is worth 4 points, and students can earn scores of 0, 1, 2, 3, or 4 points.

Depth of Knowledge

Depth of Knowledge (DOK) labels the difficulty level of the items, taking into consideration the cognitive functions or steps students must go through to respond correctly to the item. The DOK is adapted from the model used by Norman Webb, University of Wisconsin, to align standards to assessment tools.

- DOK 1: Recall and reproduction – recalling or locating information that is within the text.
- DOK 2: Skills and Concepts – involves interpretation, inferences, identification of patterns, classifications, predictions, etc.
- DOK 3: Strategic Thinking – development of an argument, critiquing, comparing and contrasting, drawing in-depth conclusions, connecting ideas to explain
- DOK 4: Extended Thinking – analyzing, synthesizing information

Item Review Checklist

1. **Does the item measure what it is intended to measure?**

2. **Does the item have only one correct answer?**

3. **Are all distracters plausible yet incorrect?**

4. **Check for clues or clang words which may influence the student's responses to other items. Do all items function independently?**

Item Review Checklist-Cont.

5. **Is the intent of the question apparent and understandable to the student without having to read the answer options?**

6. **Is the depth of knowledge level appropriate for the level of thinking skill required?**

7. **Is the item straightforward and direct with no unnecessary wordiness?**

Ground Rules for Effective Group Work

- **Every opinion is important and valued!**
- **Be polite and respectful – please wait until a speaker has finished before making additional comments.**
- **Please place cell phones on vibrate.**
- **Please hold extended conversations outside the room while others are reading or reviewing.**
- **Please let staff know if you have a request or concern.**

Advisory Committee Purpose

- **Advisory Committees help us prepare and refine items eligible to appear on the state assessment intended to measure Kentucky's chosen standards.**

- **Participate in thoughtful and meaningful discussions about grade-level content and/or item appropriateness for Kentucky students.**

Questions?

Appendix F. Item Content Committee Review Checklist

Item Content Review Committee

Checklist

1) Does the item measure what it is intended to measure?

2) Does the item have only one correct answer?

3) Are all distractors plausible yet incorrect?

4) Check for clues or clang words which may influence the student's responses to other items. Do all items function independently?

5) Is the intent of the question apparent and understandable to the student without having to read the answer options?

6) Is the depth of knowledge level appropriate for the level of thinking skill required?

7) Is the item straightforward and direct with no unnecessary wordiness?

Appendix G. Item Bias Committee Review


Kentucky Assessment
Advisory Committee
Welcome!
Item Bias Review

Kentucky Performance Rating for Educational Progress (K-PREP)

Pearson provides assessments for grades 3-8 and writing on-demand at high school.

Pearson Kentucky Program Team: An Overview

- Housekeeping
- Committee Appropriate Materials
- Nondisclosure Agreements
- Expense Reimbursement Form
- Substitute Teacher Payment Form
- MS PowerPoint Presentation

Large-Scale Assessment

- **Standardized testing for thousands of students**

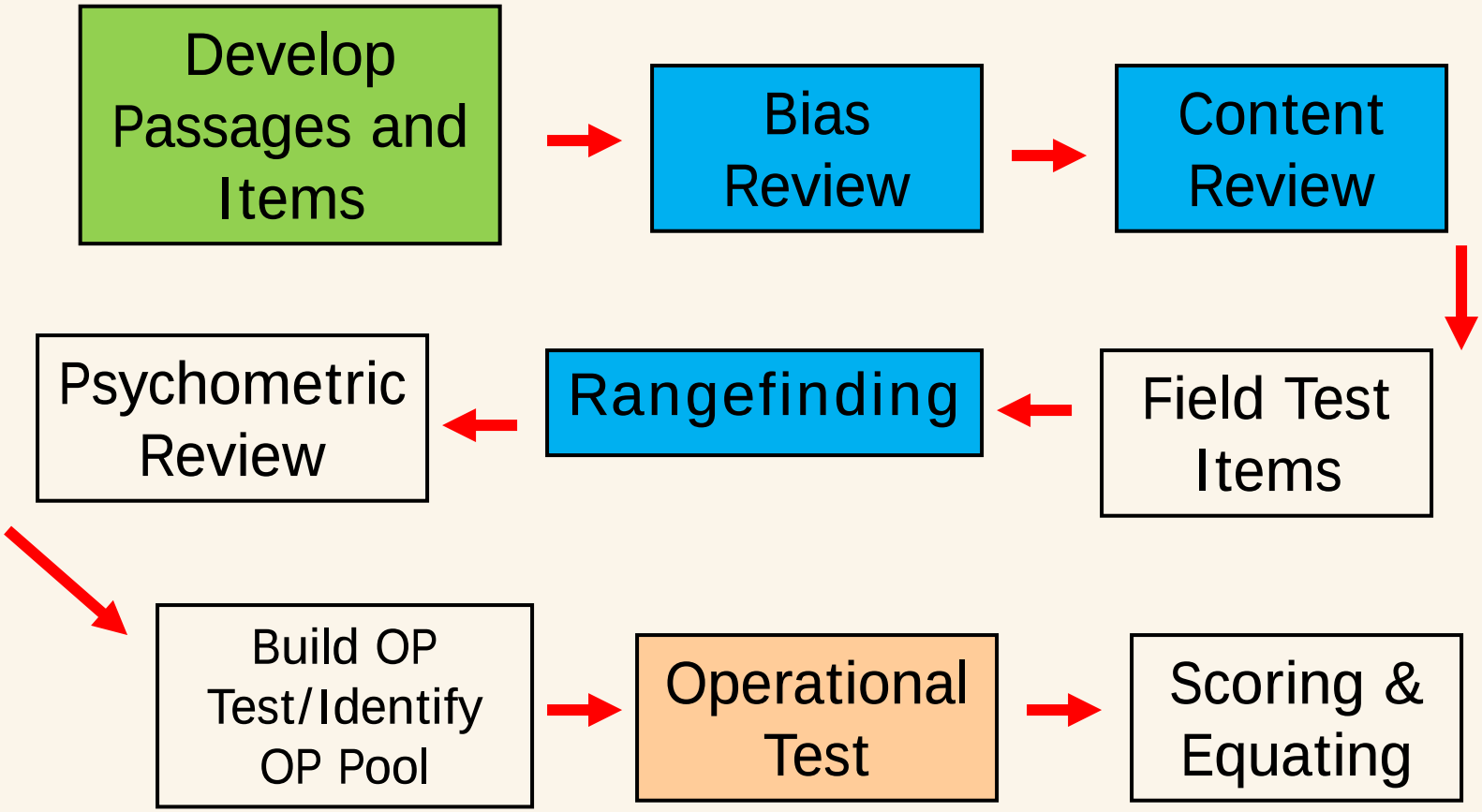
- **Designed to measure student knowledge and skills against pre-determined standards (Kentucky Core Academic Standards-KCAS)**

  - **Reading, Writing, Mathematics**
  - **Science, Social Studies**

Development Process

- Passage and Item Writer Training
- Technical Content Review
- Art Development
- Universal Design Review
- Fact Checking
- Copy Editing
- KDE Review
- Publication/Formatting

Test Development Process



Confidentiality

Certain measures are required for security purposes:

- **NO discussion with home districts or anyone else about items.**
- **NO "reproducing" of passages, prompts or items verbally, electronically or hard copy.**
- **NO cell phone use in the review rooms.**

Confidentiality is really about ensuring equity for all Kentucky students.

Housekeeping Reminder

- **Sign-in Sheet – Has everyone signed in?**
- **Participant folder Materials include:**
  - ✓Agenda
  - ✓Nondisclosure Agreement
  - ✓Expense Forms
  - ✓Related Committee Materials
- **Breakout locations**
- **Restrooms**

Bias and Sensitivity Review

Consideration

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

Construct Relevance & Language Appropriateness

- **Is the vocabulary grade appropriate?**
- **Is the language / vocabulary disadvantageous for English Language Learners?**
- **Does the passage use low frequency and / or ambiguous vocabulary?**
- **Does the passage require additional skills to those being measured?**

Gender Perspective

- What terms are used to refer to humanity at large?
- What activities are boys and girls involved?
- What emotions do characters display?
- What situations are characters placed?
- How are pictures or visuals used?

Racial, Ethnic or Cultural Perspective

- How are various ethnic groups or members of ethnic groups portrayed?
- Is there any stereotyping with respect to activities, emotions or characteristics?
- How varied are the pictures used to represent the diversity of the student population?
- Is any group over-included or under-included?

Economic or Social Class Perspective

- **How extensive is the use of luxury items or activities?**
- **How accessible to all children are the leisure activities portrayed?**
- **What values are presented in the passages and/or prompts?**

Regional Perspective

- **How common are the terms used?**
- **How accessible or familiar are the activities portrayed in the passages/prompts?**
- **What background/shared knowledge do passages/prompts expect students to have?**

Organization of Bound Booklet

- **Organized by Content**
  - Each group will review all items for all grades which includes grades 3-8

What to Look For?

Items 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.**

What to Look For? – Cont.

Items that…

- 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.

What to Look For – Cont.

Items that….

- 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.
- Are disadvantageous to English Language Learners.

Advisory Committee Purpose

- **Advisory Committees help us prepare and refine passages, items and prompts eligible to appear on the state assessment intended to measure Kentucky standards.**
- **Advisory Committees participate in thoughtful and meaningful discussions about grade-level content and/or passage, item, prompt appropriateness for Kentucky students.**

Ground Rules for Effective Group Work

- **Every opinion is important and valued!**
- **Be polite and respectful – please wait until a speaker has finished before making additional comments.**
- **Please place cell phones on vibrate.**
- **Please hold extended conversations outside the room while others are reading or reviewing.**
- **Please let staff know if you have a request or concern.**

Appendix H. Item Bias Committee Review Checklist

Bias Review Checklist
What to Look For

Items 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
- are disadvantageous to English Language Learners

Appendix I. Reading Passage Bias Committee Review


Kentucky Assessment
Advisory Committee
Welcome!
Reading Passage Review

Kentucky Performance Rating for Educational Progress (K-PREP)

- **Pearson provides all assessments for grades 3-8 and writing on-demand at high school.**

Pearson Kentucky Program Team: An Overview

- **Housekeeping**
- **Agenda**
- **Checklist**
- **Nondisclosure Agreements**
- **Expense Reimbursement Form**
- **Substitute Teacher Payment Form**
- **MS PowerPoint Presentation**

Large-Scale Assessment

- **Standardized testing for thousands of students**

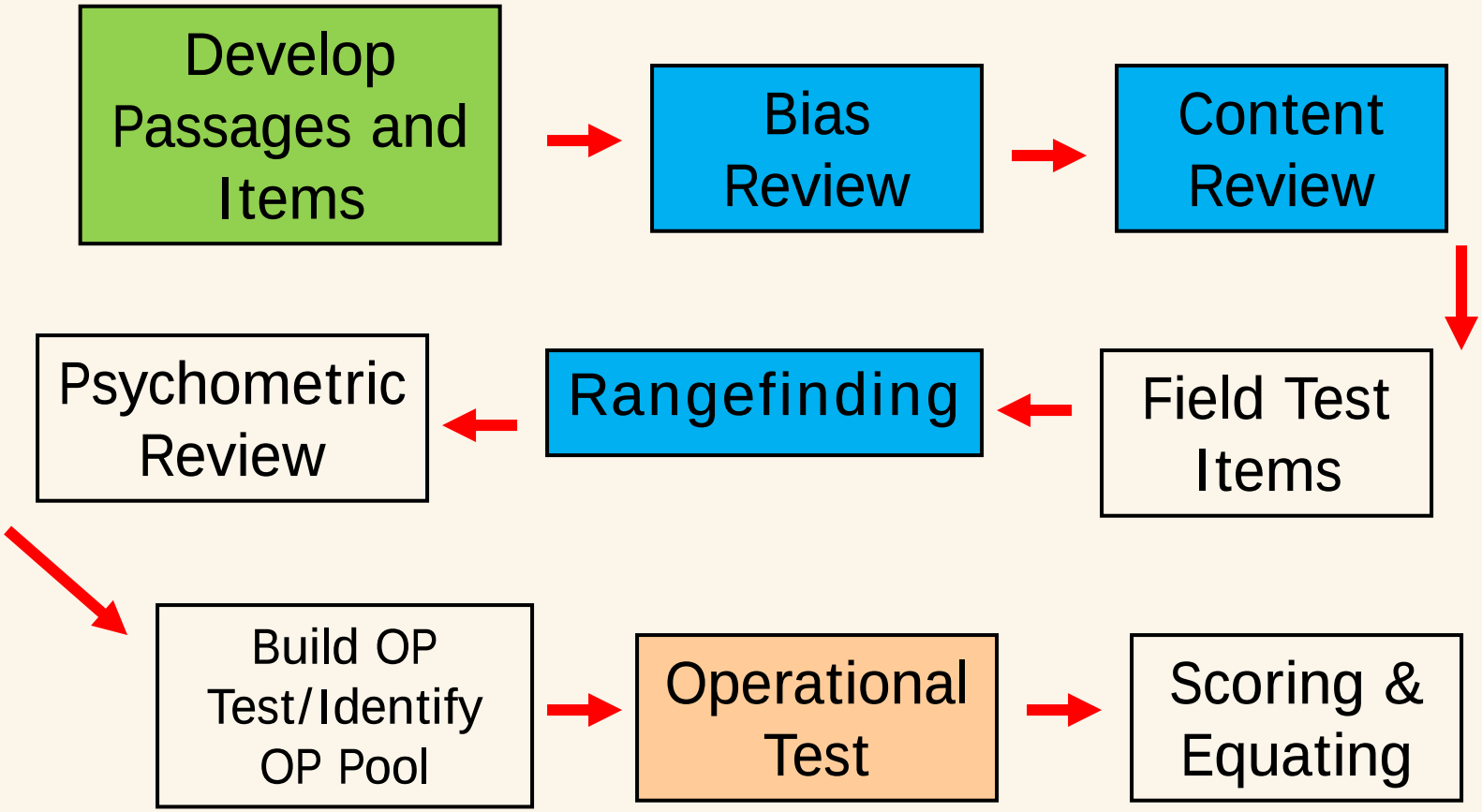
- **Designed to measure student knowledge and skills against pre-determined standards (Kentucky Core Academic Standards-KCAS)**

  - **Reading, Writing, Mathematics**
  - **Science, Social Studies**

Development Process

- Passage and Item Writer Training
- Technical Content Review
- Art Development
- Universal Design Review
- Fact Checking
- Copy Editing
- KDE Review
- Publication/Formatting

Test Development Process



PEARSON

Confidentiality

Certain measures are required for security purposes:

- **NO discussion with home districts or anyone else about items.**
- **NO "reproducing" of passages, prompts or items verbally, electronically or hard copy.**
- **NO cell phone use in the review rooms.**

Confidentiality is really about ensuring equity for all Kentucky students.

Housekeeping Reminder

- **Sign-in Sheet – Has everyone signed in?**
- **White folder Materials include:**
  - ✓Agenda
  - ✓Nondisclosure Agreement
  - ✓Expense Forms
  - ✓Related Committee Materials
- **Breakout locations**
- **Restrooms**

Bias and Sensitivity Review

Consideration

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

Construct Relevance & Language Appropriateness

- **Is the vocabulary grade appropriate?**
- **Is the language / vocabulary disadvantageous for English Language Learners?**
- **Does the passage use low frequency and / or ambiguous vocabulary?**
- **Does the passage require additional skills to those being measured?**

Gender Perspective

- What terms are used to refer to humanity at large?
- What activities are boys and girls involved?
- What emotions do characters display?
- What situations are characters placed?
- How are pictures or visuals used?

Racial, Ethnic or Cultural Perspective

- **How are various ethnic groups or members of ethnic groups portrayed?**
- **Is there any stereotyping with respect to activities, emotions or characteristics?**
- **How varied are the pictures used to represent the diversity of the student population?**
- **Is any group over-included or under-included?**

Economic or Social Class Perspective

- **How extensive is the use of luxury items or activities?**
- **How accessible to all children are the leisure activities portrayed?**
- **What values are presented in the passages and/or prompts?**

Regional Perspective

- **How common are the terms used?**
- **How accessible or familiar are the activities portrayed in the passages/prompts?**
- **What background/shared knowledge do passages/prompts expect students to have?**

Organization of Bound Booklet

- **Organized by Grade**
  - Group 1: Grades 3, 4, 5
  - Group 2: Grades 6, 7, 8

What to Look For?

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.**

What to Look For? – Cont.

Passages that…

- **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.**

What to Look For – Cont.

Passages that….

- 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.
- Are disadvantageous to English Language Learners.

Advisory Committee Purpose

- **Advisory Committees help us prepare and refine passages, items and prompts eligible to appear on the state assessment intended to measure Kentucky standards.**
- **Advisory Committees participate in thoughtful and meaningful discussions about grade-level content and/or passage, item, prompt appropriateness for Kentucky students.**

Ground Rules for Effective Group Work

- **Every opinion is important and valued!**
- **Be polite and respectful – please wait until a speaker has finished before making additional comments.**
- **Please place cell phones on vibrate.**
- **Please hold extended conversations outside the room while others are reading or reviewing.**
- **Please let staff know if you have a request or concern.**

Appendix J. Reading Passage Bias Committee Review Checklist

Bias Review Checklist

What to Look For

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
- are disadvantageous to English Language Learners

Appendix K. ODW Item Writer Training

KENTUCKY

On-Demand Writing

Prompt Writers' Training

2011

The Test

- Students in grades 5, 6, 8, 10, and 11 will take the writing test
- Each student will write to two prompts:
  - Choice of a stand alone prompt (two prompts will be presented; students will select one)
  - One mandatory passage-based prompt
- Students will be given the following time durations to complete each response

  - **Grade 5:**

    30 minutes for the choice prompt
    90 minutes for the passage-based prompt

  - **Grades 6, 8, 10 and 11:**

    40 minutes for the choice prompt
    90 minutes for the passage-based prompt

  In the time allotted, students must:
  - allow time to read the prompt and associated material
  - plan their writing
  - create their final response.

Considerations

- Kentucky has adopted the Common Core State Standards (CCSS). All items and writing prompts must assess and align with these standards.

  (Separate handout - http://www.corestandards.org/)
- The modes of responses for the KY Writing-on-demand will be:
  - **Opinion (grade 5 only)**
  - **Argument (grades 6, 8, 10, 11)**
  - **Informative / Explanatory (grades 5, 6, 8, 10, 11)**
  - **Narrative (grades 5, 6, 8)**

*** Narrative prompts will not be used at grades 10 and 11. The narrative standards in the common core for those grades will be assessed within the classroom rather than on the state assessment.**

*** We are using "Modes" to refer to the "types and purposes" from the common core.**

Prompt Topics

Prompts must

- allow writers the opportunity to use a variety of organizational structures.
- be accessible to all students and be free of bias and sensitivity concerns.
- be engaging and age-appropriate for students.
- have adequate depth and coverage to provide students the information they will need to respond.

Prompts should avoid topics that

- refer to or discuss taboo subjects
- may elicit inappropriate responses (e.g., drug usage, fearful situations, sexual activities, violence)
- are negative or emotionally charged
- may be perceived as personally invasive

Prompts

- are unique and grade level appropriate.
- are universally understood without extensive prior knowledge.
- give sufficient background information which aligns to the specific intent of the standard being assessed.
- clearly state the purpose.
- clearly specify the appropriate audience and tie the audience to the designated purpose.

* *

- In grades 5 and 6, the format for the response should be identified.
- In grades 8, 10, and 11, choices of formats within the mode should be provided for students (e.g., speech, editorial, letter, etc.)
- Common formats for responses should be used appropriate to the designated mode (e.g., articles, editorials, letters, narratives, etc.)

Language of Prompts

Prompts should

- use simple, direct language for easy understanding
- allow choices when possible
- use vocabulary below grade level
- use cue words to assist students in knowing the mode of their response
- clearly differentiate modes by not using the same verbs across modes
- specify the audience for the response; the audience must be appropriate for the purpose of the task
- clearly state an authentic purpose
- use brief reading passages, tables, charts, graphics, etc. to provide context for students
- clearly state the required mode and format of response; students will have a choice of formats in grades 8, 10, and 11
- have a writer role that is grade appropriate

Key words / phrases

There are key words or phrases that often help to signal the expected modes and types of responses for each prompt. Being aware of these key words and using them can be an aid to ensure appropriate student response.

- For **explanatory / informative prompts** with the purposes of informing, clarifying, explaining, defining or instructing, some cues are "why is/does," "how," "describe," "what."

- For **narrative prompts**, characterized by creativity, drama, suspense, humor, fantasy, and descriptive imagery, some cues are "tell about a time," "tell what happened," "write a story that shows." ("Why" is often avoided, as this word tends to cue and elicit expository writing.) (Narrative responses need to move beyond simply telling a story; there must be a relevant purpose to this telling that gets at significance and insight.)

- For **argument prompts** which tend to be persuasive with points of view or defense of sides of controversy, some cues are "convince," "persuade," "propose," "present an argument for" ("How" is often avoided as this word tends to elicit either explanation or narratives.)

- For grade 5 **opinion prompts**, some cues are words or phrases that elicit a student's thoughts or beliefs, such as "Do you agree," or "What do you think or believe."

Prompt constructs

Every stand alone prompt should contain a writing situation and directions for writing which will identify the required response mode :

(All narrative prompts are considered stand alones.)

- **Writing situation:**
  This directs students to write about a specific topic which will serve as the central theme and purpose of the written response. The intent of the situation is to expand, restate, or clarify the topic for the students, **NOT** to preclude student individual responses. It is possible for the situation to be something other than text, such as a table, a chart, or a graphic either with or instead of a prose text.

- **Directions for writing:**
  The directions will include a strategy statement suggesting the approach or mode the students should use in crafting their responses. The directions MUST BE SUCCINCT! The directions must state the purpose, mode ,format of delivery and the audience.

What are you writing?
Why are you writing?
For whom are you writing?

REMEMBER:

Grade 5: 30 minutes testing time (10 minutes to read and plan; 20 minutes to write response)
Grades 6+: 40 minutes testing time (10 minutes to read and plan; remaining time to write)

Sample writing prompt

Writing situation:

Many students in your community are involved in an intramural sports program which combines students from several different districts, allowing them to participate in both regulation sports as well as be introduced to lesser-known sports like cricket and jai alai. Unfortunately, state funding to schools has been reduced forcing schools to make sometimes painful decisions about which programs to cut. This intramural program, although very popular, is one schools are considering cutting. You want to prevent this cut from happening.

Writing directions:

Write a speech you intend to present at the next school board meeting to which students and other community members have been invited to attend. Explain your position on possible cuts to the intramural sports program giving reasons for your thinking.

Definitions

Explanatory / informative writing mode

- The purpose is to inform, clarify, explain, define, or instruct by giving information, explaining how or why, clarifying a process, or defining a concept – sometimes with examples.

- Well-written explanation has a clear, central focus developed through a carefully crafted presentation of relevant facts, examples or definitions that enhance the reader's understanding.

- Supporting elements are objective and not dependent on emotion. The writing, however, although objective and factual, may be engaging, lively, and passionate reflecting the writer's connection to the topic.

Informative / Explanatory

Students will be expected to:

- establish a thesis or topic
- show awareness of audience through content presented and tone of delivery
- provide explanation and insight through complete examination of the topic
- provide a good balance between generalizations and specific details in support for and development of ideas
- respond in a logically organized and coherent manner
- use language and tone that are appropriate to both the task and the audience

Therefore, the writing situation, whether stand alone or passage based, must be broad enough to give students the resources and/or jumping off point they will need in order to adequately respond.

Definitions

Narrative writing mode

- The purpose of narrative writing is to recount a personal or fabricated experience or to tell a story based on a real or imagined event.
- Unlike expository writing, the narrative uses emotion and more descriptive and figurative language.
- In well-written narration, a writer uses personal insight, creativity, drama, suspense, humor or fantasy to create a central theme or impression.
- The supporting details work together to develop an identifiable story line that is easy to follow and retell.

Narrative

Students will be expected to:

- Clearly convey the significance of an experience and all of its complexities, whether real or imagined
- Illustrate or recreate the experience with effective examples and sensory details for the audience
- Develop response with logical organization and reading flow

Narrative prompts will be stand alones rather than passage-based; although, the writing situations may be longer than other stand alones or use a short public domain folktale to create the springboard for the students' relating of the experience.

Definitions

Argument mode

- The purpose of developing an argument is to establish and support a position in a dual or multiple-positioned issue.
- The writer's position is personal but is supported and influenced by facts and other evidence.
- In the response, the writer establishes a position and then supports that position with relevant and sufficient facts, details, quotations which provide the reasons for his/her belief.
- Anticipation of counterclaims and alternate opinions or points of view are important to consider. These counterclaims may be used by the writer to clarify his/her position in the refutation of the counterclaims.

Argument

Students will be expected to:

- Establish an argument by introducing and focusing on a precise claim
- Show awareness of the audience in their use of tone
- Show awareness of the audience's knowledge level and needs
- Distinguish between the claims and counterclaims in their presentation of position
- Develop argument with relevant and effective support: reasons, examples, quotations, etc.
- Organize their argument effectively

Argument vs. Persuasion

When writing to persuade, students attempt to convince the audience by:

- Using sense of identity appeal
- Using emotional appeal
- Establishing their own credibility
- Establishing their knowledge and trust level

When writing a logical argument, students convince the audience by:

- Establishing a precise claim or position
- Assessing the validity of their thinking
- Anticipating audience knowledge of topic
- Anticipating counterclaims to their position/assertion
- Organizing arguments and support logically

Definition

Opinion mode (grade 5)

- In grade 5, students may be asked to present their opinion or belief about a certain situation or topic.
- Opinions are personal and often emotionally-charged.
- Support for opinions are often based on experiences and extraneous influences as well as on observations and some facts.
- The purpose is to make someone aware of the writer's reactions and feelings about something rather than to try to instigate any sort of action.
- The audience is important in this mode to allow the writer the appropriate tone and confidence.

Passage-based prompts

- For the Kentucky On-Demand Writing test, all students will be given a mandatory passage-based prompt along with their choice of a stand alone.
- Fairly brief stimuli which may either include or be tables, charts, graphics instead of all text to provide the context for the students' responses.
- The passage-based prompt will entail students' reading of a passage or passage set and then responding to the associated prompt.
- The testing time frame for the passage-based prompt will be ninety minutes. The text of the reading must be short enough to allow students to do reading and pre-planning in 20 to 30 minutes. They should have the remaining 60 – 70 minutes to write their response in the required mode and format.

CAUTION:

This is a test to assess students' abilities to create a well-written response that achieves the intent of the standard mode being assessed. It is **NOT** a test of their reading comprehension.

Stimuli for passage-based

- Passages or other stimuli must align to the standard modes being assessed: narrative, informative/explanatory, and opinion/argument.
- All writing must be clear and coherent. In public domain documents, the language used must be accessible and relatable for students. Avoid those documents using archaic language and language structures.
- The stimuli for passage-based will align to either the informative/explanatory or argument/opinion modes.
- Stimuli must contain enough breadth and depth to allow students to adequately respond according to the specified directives of the prompt.
- The stimulus must be able to be read thoroughly in approximately five to ten minutes in order to allow students time to plan and construct their responses.

Paired Stimuli

- Paired stimuli which seem particularly well-suited to support an argument or a comparison/contrast development may be used.
- The combined length of the pair should fall into the same reading time allotment as used for a single passage (five to ten minutes reading time).
- For opinion/argument response, the passages must fairly present both sides of an issue
  - allowing students to take and support one side.

    OR
  - allowing students to juxtapose a more “middle of the road” position.

- For a pairing, it may be possible to use some sort of data set as the second stimulus.
- Data sets, charts, tables, graphics may either be used as a part of a pair or individually. They must contain applicable and sufficient information needed for the response expected.

Sample 1 with comments

Grade 5: Narrative

[Situation: topic-meeting a person who has had a positive impact
Writing directions: Describe for your readers how the meeting went and how that person has positively impacted your life today.]

Comments:

The format for the response and the specific audience are missing from this directive. “Describe” is too vague. For grade 5, it is fine to identify a specific format for the students.

Students need to know to whom they are writing.

“Impacted” may be unfamiliar to many students at this grade. A more accessible word such as “affected” would be more appropriate.

Revision:

Write a speech to present to your classmates about this person and how he/she has affected your life today.

Sample 2 with comments

Grade 6: Argument

[Situation: The Aesop tale “Town Mouse and Country Mouse” was used. (public domain)

Writing directions: Think about whether living a complicated life or a calmer, more sedate life suits you better. Write a paper in order to convince others to believe as you do. Be sure to state your position and support it with reasons and examples from the folktale.]

Comments:

The directions are way too wordy.

The specific format and audience are missing.

“Sedate” may be unfamiliar to many students.

A confusion exists concerning the appropriate mode being assessed.

It will be difficult to use specific examples from the folktale as support since the characters and situation are not directed to human beings.

Revision: In a letter to your family, tell them why living in a city or in the country is better for a future move. Using ideas from the folktale, support your argument.

Sample 3 with comments

Grade 8: Informative/explanatory

[Situation: A set up was given indicating the time students spend in school along with a quotation from author, Sam Ewing.

Writing directions: Write an explanation you would give to your younger friends and siblings of what Sam Ewing's means by these words and how they relate to time spent doing school work.]

Comments:

"Explanation" suggests a very short constructed response students might create for a single question instead of directing students to a format for a more extensive response.

A more specific, real-world format should be offered as a choice: (i.e., speech, letter, etc.).

The audience is appropriate and is stated clearly.

Revision: Write an explanation of this quotation and how it relates to the work in school to be posted on a school website welcoming students who are new to the district.

Sample 4 with comments

Grade 8: Argument

[Situation: An article about career and technical education in our schools was used.

Writing directions: Based on information in the passage, think about what you believe high schools should teach in order to prepare students for adult life. Write an editorial for a newspaper about the extent job-related skills should be taught in high school.]

Comments:

The directions are too cluttered. Be direct and to the point. Get immediately to the task at hand.

The format is stated; however, the audience is implied rather than directly stated which is sometimes appropriate.

Revision:

What should high schools teach to prepare students for jobs and careers? Write an editorial for your community newspaper stating your position and reasons for it.

Sample 5 with comments

Grade 10: Paired stimuli Argument

(Opening two paragraphs for first passage)

The catwalk is set. Nervous anticipation permeates the bevy of beauties as they prepare to “strut their stuff” in front of eager spectators and admirers. Hours upon hours of primping, clipping, tweezing or errant strands, powdering and polishing, combing tangles into shiny, flowing locks prefaced these impending minutes in the spotlight.

This could be Paris, Milan, Rome or New York – any of the hubs of the world of couture high fashion. Models with the slender figures decked out in designer couture, doing what they seemingly were destined to do – pleasing spectators by showcasing the designers’ talents. One has to wonder: Is this natural reality or is this nature redirected into largely unattainable dreams?

Comments for sample 5

Comments:

The language used will be largely unattainable for the average tenth grade reader. Since these are the opening paragraphs, those who cannot understand the vocabulary will already be defeated.

(Much better suited to a reading passage although even then, context will be needed to aid the reader.)

The actual issue being addressed is whether it is appropriate to physically alter a dog's appearance when it is not medically necessary, such as ear cropping and tail docking. These paragraphs set up an analogy and perhaps add some interest, but they do nothing at all to direct students to a purpose.

Revision: The extraneous information must be removed. It will shorten the pair and move more directly to the information the students will actually use for their own arguments.

The vocabulary will need to be simplified.

Submission Directions

- **All prompts/passages/items are secure and become the property of Pearson and may NOT be used for any other purpose by the writer.**

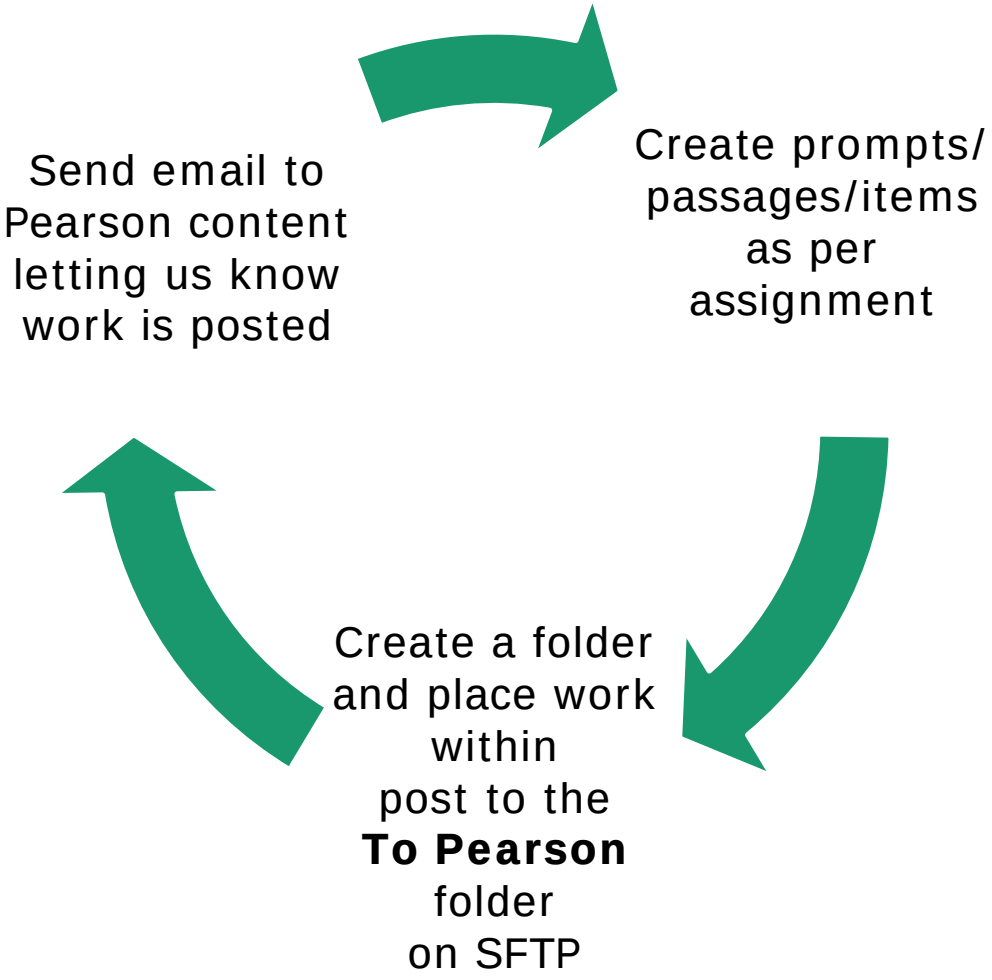
- **Deadlines must be met**; our schedule is tight and allows no room for delays
- Post the work in the To Pearson folder on the SFTP site
- Email the content specialist with whom you are working, copying the other specialists on the same email letting them know when work has been posted.
- Any revision requests or feedback will be posted in the From Pearson folder on the same SFTP site.

NOTE: Security is paramount!

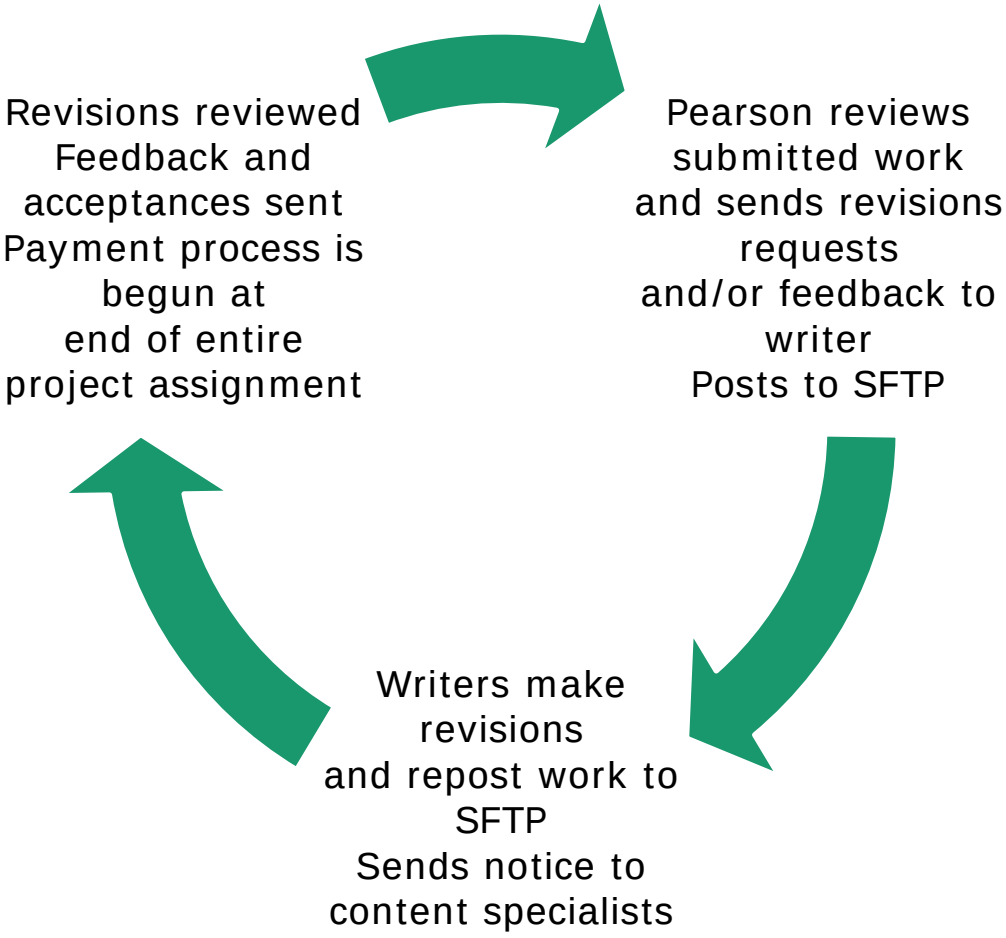
Everything transferred between contracted writer and Pearson MUST be placed on the secure SFTP site. Avoid using email to discuss anything specifically related to the passages or prompts. Use email ONLY to let Pearson know something has been posted on the SFTP site or to arrange a phone conversation. General questions are possible in emails.

**** We ask that writers initially post a list of topics and passage types (single or paired) upon receipt of assignments. This is to prevent overlap or duplication of topics.**

Work flow



Work flow continued



Appendix L. ODW Content Committee Review

Kentucky Assessment Advisory Committee

Welcome!

On-Demand Writing Prompt Review

The Call for a New Assessment System

- **Senate Bill 1 (SB 1), enacted in the 2009 Kentucky General Assembly, requires a new public school assessment program beginning in the 2011-12 school year.**

- **The new assessments will be called Kentucky Performance Rating for Educational Progress (K-PREP) tests.**

Who Will Provide the K-PREP Assessments?

1. **Pearson has been awarded the contract to provide all assessments for grades 3-8 and writing on-demand at high school.**

2. **Pearson currently provides large-scale assessment services in more than 25 states and for the U.S. Department of Education.**

Pearson Kentucky Program Team: An Overview

- **Housekeeping**
- **Agenda**
- **Checklist**
- **Reference Material**
- **Non-Disclosure Agreements**
- **Expense Reimbursement Form**
- **Substitute Teacher Payment Form**
- **MS PowerPoint Presentation**

Large-Scale Assessment

- **Standardized testing for thousands of students**

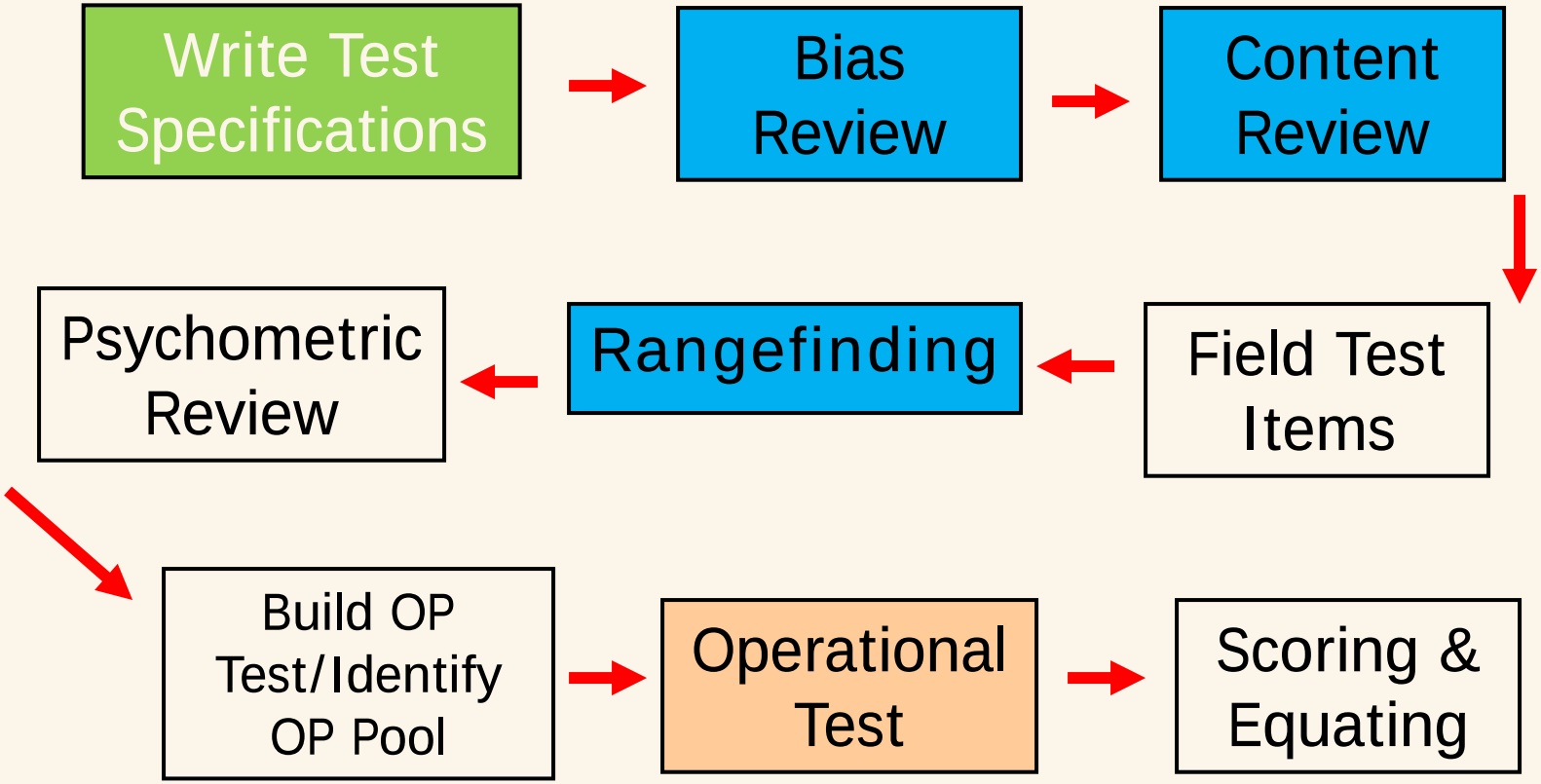
- **Designed to measure student knowledge and skills against pre-determined standards (Kentucky Core Academic Standards-KCAS)**

  - **Reading, Writing, Math**
  - **Science, Social Studies**

Development Process

- **Passage and Prompt Writer Training**
- **Technical Content Review**
- **Art Development**
- **Universal Design Review**
- **Fact Checking**
- **Copy Editing**
- **Publication – Formatting**
- **KDE Review**

Test Development Process



Confidentiality

Certain measures are required for security purposes:

- **NO discussion with home districts or anyone else about items.**
- **NO "reproducing" or passages, prompts or items verbally, electronically or hard copy.**
- **NO cell phone use in the review rooms**

Confidentiality is really about ensuring equity for all Kentucky students.

Housekeeping Reminder

- **Sign-in Sheet – Has everyone signed in?**
- **White folder Materials include:**
  - ✓Agenda
  - ✓Non-Disclosure Agreement
  - ✓Expense Forms
  - ✓Related Committee Materials
- **Breakout locations**
- **Restrooms**

Common Core Standards for Writing

In the common core, writing has been divided into three types and purposes:

Grade 5:

- **Opinion**
- **Informative/explanatory**
- **Narrative**

Grades 6, 8, 10, 11:

- **Argument**
- **Informative/explanatory**
- **Narrative**

Common Core Writing

While the Common Core standards are divided into types and purposes (referenced as Modes for this review), the expectation is that all writing will be clear and coherent. The development, organization and style will be appropriate to the task, purpose and audience for which it is written.

Grade 5

- **Opinion responses must support a point of view with reasons and information.**
  - Support may come from the Writing situation or the related passage. While personal feelings/experiences and prior knowledge cannot be eliminated from the responses, the prompts and related background information or passages must contain enough information to make it possible for all students to respond regardless of their own experiences.
- **Informative/explanatory responses are to examine a topic and convey ideas and information.**
- **Narrative responses develop real or imagined experiences or events using effective techniques, descriptive details, and event sequences.**

Grades 6 and 8

- **Argument responses require students to take a stand or position on an issue and then provide textual evidence for this position.**
  - The writing situation or passage must present an issue with opposing sides represented equitably allowing students to use textual evidence as support for their own positions as well as the opportunity to rebut counterarguments as additional support.
- **Informative/explanatory responses must convey information to an audience appropriate to the task at hand and the knowledge base and needs of that audience.**
- **Narrative responses require students to develop or recreate real or imagined events or experiences using effective techniques, language, and well-structured sequences.**

Grades 10 and 11

- **Argument responses require students to take a stand or position on an issue and then provide textual evidence for this position.**
  - The writing situation or passage must present an issue with opposing sides represented equitably allowing students to use textual evidence as support for their own positions as well as the opportunity to rebut counterarguments as additional support.
- **Informative/explanatory responses must convey information to an audience appropriate to the task at hand and the knowledge base and needs of that audience.**
- **Narrative responses are not being assessed on the On-Demand Writing assessment for grades 10 and 11.**

Orientation to Participant Materials

- **Writers' Reference Sheet – Draft**
- **Scoring Rubric – Draft**
- **Bound Book with passages and prompts**

Organization of Bound Booklet

Organized by Mode

- **Stand-alone prompts**
- **Passage-based prompts**
  - Some passages paired

Prompt Review Process

- **Read the passage and/or prompt.**
- **Think about possible responses.**
- **Match the prompt to the Common Core Standard**

Prompt 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?**

Prompt Review Checklist - Cont.

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?**

Prompt Review Checklist-Cont.

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?**

Prompt Review Checklist-Cont.

9. Is the prompt free of bias and sensitivity issues?

10. Is the passage or situation written in a clear and direct manner?

Ground Rules for Effective Group Work

- **Every opinion is important and valued!**
- **Be polite and respectful – please wait until a speaker has finished before making additional comments.**
- **Please place cell phones on vibrate.**
- **Please hold extended conversations outside the room while others are reading or reviewing.**
- **Please let staff know if you have a request or concern.**

Advisory Committee Purpose

- **Advisory Committees help us prepare and refine passages, items and prompts eligible to appear on the state assessment intended to measure Kentucky's chosen standards.**

- **Participate in thoughtful and meaningful discussions about grade-level content and/or passage, item, prompt appropriateness for Kentucky students.**

Group Discussion

Appendix M. ODW Prompt Review Checklist

Prompt 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 N. ODW Bias Committee Review

Kentucky Assessment Advisory Committee

Welcome!

On-Demand Writing Prompt Review

The Call for a New Assessment System

- **Senate Bill 1 (SB 1), enacted in the 2009 Kentucky General Assembly, requires a new public school assessment program beginning in the 2011-12 school year.**

- **The new assessments will be called Kentucky Performance Rating for Educational Progress (K-PREP) tests.**

Who Will Provide the K-PREP Assessments?

- **Pearson has been awarded the contract to provide all assessments for grades 3-8 and writing on-demand at high school.**

- **Pearson currently provides large-scale assessment services in more than 25 states and for the U.S. Department of Education.**

Pearson Kentucky Program Team: An Overview

- **Housekeeping**
- **Agenda**
- **Checklist**
- **Reference Material**
- **Non-Disclosure Agreements**
- **Expense Reimbursement Form**
- **Substitute Teacher Payment Form**
- **MS PowerPoint Presentation**

Large-Scale Assessment

- **Standardized testing for thousands of students**

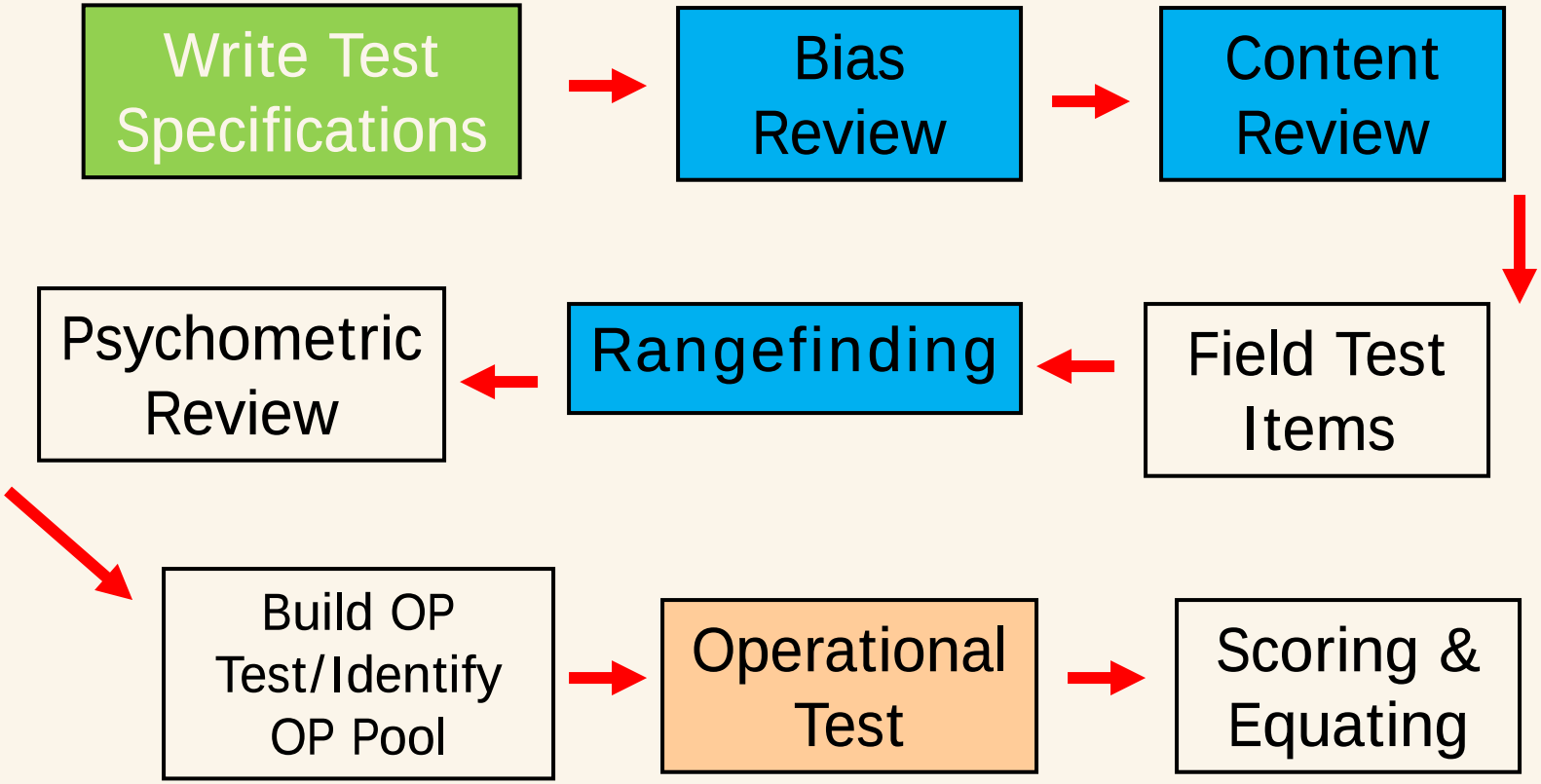
- **Designed to measure student knowledge and skills against pre-determined standards (Kentucky Core Academic Standards-KCAS)**

  - **Reading, Writing, Math**
  - **Science, Social Studies**

Development Process

- **Passage and Prompt Writer Training**
- **Technical Content Review**
- **Art Development**
- **Universal Design Review**
- **Fact Checking**
- **Copy Editing**
- **Publication – Formatting**
- **KDE Review**

Test Development Process



Confidentiality

Certain measures are required for security purposes:

- **NO discussion with home districts or anyone else about items.**
- **NO "reproducing" or passages, prompts or items verbally, electronically or hard copy.**
- **NO cell phone use in the review rooms**

Confidentiality is really about ensuring equity for all Kentucky students.

Housekeeping Reminder

- **Sign-in Sheet – Has everyone signed in?**
- **White folder Materials include:**
  - ✓Agenda
  - ✓Non-Disclosure Agreement
  - ✓Expense Forms
  - ✓Related Committee Materials
- **Breakout locations**
- **Restrooms**

Bias and Sensitivity Review

Consideration

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

Gender Perspective

- What terms are used to refer to humanity at large?
- What activities are boys and girls involved in?
- What emotions do characters display?
- What situations are characters placed in?
- How are pictures or visuals used?

Racial, Ethnic or Cultural Perspective

- **How are various ethnic groups or members of ethnic groups portrayed?**
- **Is there any stereotyping with respect to activities, emotions, or characteristics?**
- **How varied are the pictures used to represent the diversity of the student population?**
- **Is any group over-included or under-included?**

Economic or Social Class Perspective

- **How extensive is the use of luxury items or activities?**
- **How accessible to all children are the leisure activities portrayed?**
- **What values are presented in the passages and/or prompts?**

Regional Perspective

- How common are the terms used?
- How accessible or familiar are the activities portrayed in the passages/prompts?
- What background/shared knowledge do passages/prompts expect students to have?

Organization of Bound Booklet

- **Organized by Mode**
  - Stand-alone prompts
  - Passage-based prompts
    - Some passages paired

What to Look For?

Passages/Prompts that...

- **Reflect favoritism towards a gender or ethnic group.**
- **Are potentially offensive, inappropriate, or negative toward any group.**
- **Discriminate in any way against individuals with disabilities.**

What to Look For? – Cont.

Passages/Prompts that…

- **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.**

What to Look For – Cont.

Passages/Prompts that….

- **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.**

Advisory Committee Purpose

- **Advisory Committees help us prepare and refine passages, items and prompts eligible to appear on the state assessment intended to measure Kentucky's chosen standards.**
- **Participate in thoughtful and meaningful discussions about grade-level content and/or passage, item, prompt appropriateness for Kentucky students.**

Ground Rules for Effective Group Work

- **Every opinion is important and valued!**
- **Be polite and respectful – please wait until a speaker has finished before making additional comments.**
- **Please place cell phones on vibrate.**
- **Please hold extended conversations outside the room while others are reading or reviewing.**
- **Please let staff know if you have a request or concern.**

Appendix O. ODW Bias Review Checklist

Bias Review Checklist

What to 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. Science Bias Review Committee

Kentucky Department of Education
Kentucky Academic Standards
Science Assessment
Grades 4, 7 and HS Science

Bias Review
December 7-8, 2016


Namaste
مرحبا
Willkommen
Bem Vindo
Selamat Datang
Namaste
Croeso
Bienvenue
Bienvenue
Bienvenidos
Welcome
Bienvenidos
أهلا وسهلا
Benvenuti
Welcome
Welkom
Bem Vindo
Bienvenue
مرحبا
Welcome
Croeso
Welkom
Bienvenidos
Namaste
Welkom
أهلا وسهلا
مرحبا
Selamat Datang
Bienvenue
أهلا وسهلا
Welcome
Bem Vindo
Willkommen
Willkommen
Selamat Datang
Croeso
добре дошъл
Benvenuti
Willkommen
Καλώς ήλθατε
Benvenuti

Pearson

Welcome

Introductions

Housekeeping

Sign-In Sheet

Nondisclosure Agreement

Expense Forms

Checklist

Evaluation Form

Materials Security

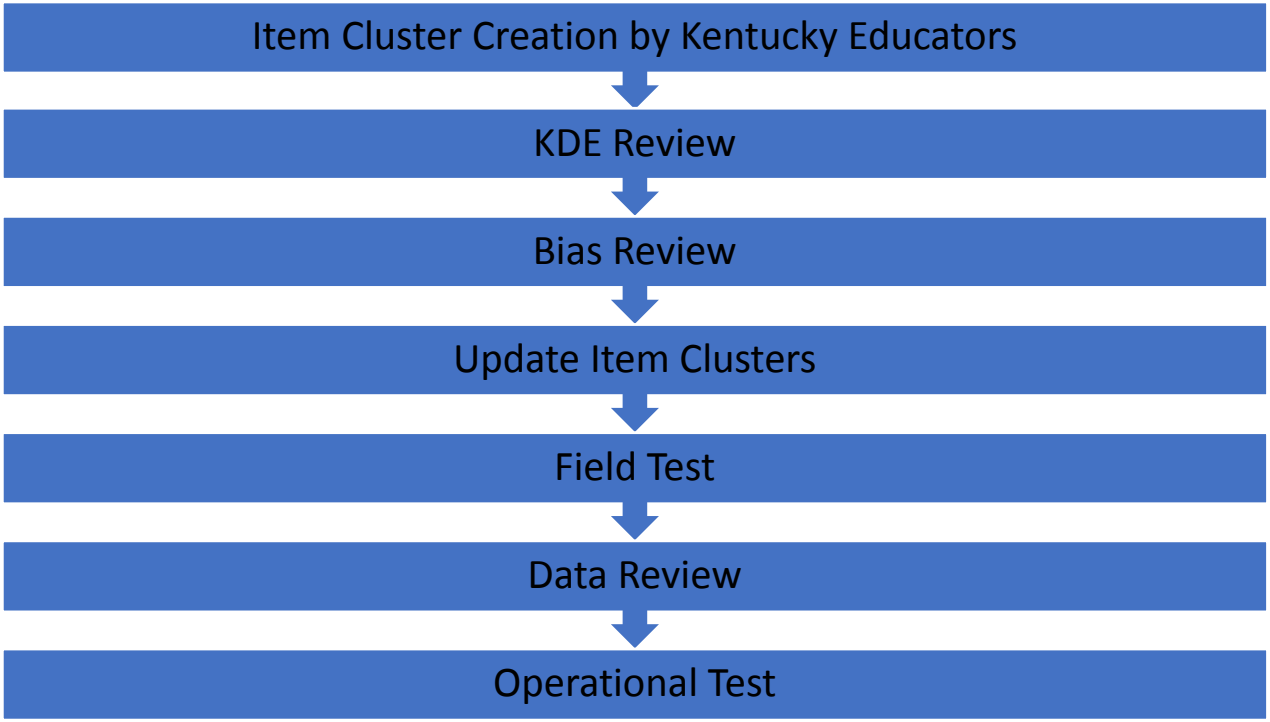
Wednesday, December 7, 2016

7:30 – 8:30	**Breakfast**	*Room 416*
8:30 – 9:30	**Orientation/Group Training**	*Room 416*
9:30 – 12:00	**Grade Level Breakouts**	
	Grade 4	*Room 540*
	Grade 7	*Room 540-A*
	High School	*Room 542-A*
12:00 – 1:00	**Lunch**	*Room 416*
1:00 – 2:30	**Bias Review/Grade Level Breakouts**	
2:30 - 2:45	**Break**	
2:45 - 4:00	**Bias Review/Grade Level Breakouts**	

Thursday, December 8, 2016

7:30 – 8:30	**Breakfast**	*Room 416*
8:30 – 12:00	**Grade Level Breakouts**	
	Grade 4	*Room 517*
	Grade 7	*Room 540 - A*
	High School	*Room 542-A*
12:00 – 1:00	**Lunch**	*Room 416*
1:00 – 4:00	**Bias Review/Grade Level Breakouts**	

Item Cluster Development Process or Life Cycle



Review Process

Panel members will:

Independently review item clusters

Evaluate each item cluster based on criteria

Discuss each item cluster as a group

Recommend to Accept or Revise each item cluster

Bias Review Committee Purpose

- Bias Review Committees help us prepare and refine item clusters that are eligible to appear on the state assessment intended to measure Kentucky standards.
- Bias Review Committees participate in thoughtful and meaningful discussions about bias or sensitivity appropriateness for Kentucky students.

Construct Relevance & Language Appropriateness

- Is the vocabulary grade-level appropriate?
- Is the language/vocabulary disadvantageous for English Language Learners?
- Does the cluster use low frequency and/or ambiguous vocabulary (excluding science vocabulary explicit in the content and/or standards)?
- Does the cluster require additional skills to those being measured?

Gender Perspective

- What terms are used to refer to humanity at large?
- What activities are boys and girls involved in?
- What emotions do people depicted in item clusters display?
- What situations are people depicted in item clusters placed in?
- How are pictures or visuals used?

Racial, Ethnic, or Cultural Perspective

- How are various ethnic groups or members of ethnic groups portrayed?
- Is there any stereotyping with respect to activities, emotions or characteristics?
- How varied are the pictures used to represent the diversity of the student population?
- Is any group over-included or under-included?

Economic or Social Class Perspective

- How extensive is the use of luxury items or activities?
- How accessible to all children are the leisure activities portrayed?
- What values are presented in the item clusters?

Regional Perspective

- How common are the terms used?
- How accessible or familiar are the activities portrayed in the item clusters?
- What background/shared knowledge do item clusters expect students to have?

To summarize ….

- Is there anything in the item cluster that will distract or upset the students which may hinder them from demonstrating what they know about science content and processes?

Confidentiality

Certain measures are required for security purposes:

NO discussion with home districts or anyone else about items.

NO “reproducing” of item clusters verbally, electronically or hard copy.

NO cell phone or iPad use in the review rooms.

Confidentiality is really about ensuring equity for all Kentucky students.

Secure Materials

Leave all materials in the room.

Do not discuss items with anyone.

Questions or comments?

Thank you.

Appendix Q. On-Demand Writing Scoring Rubric

4 Points:
Writers at this score point level display consistent, though not necessarily perfect, writing skill, resulting in effective communication.

- The writer establishes and maintains focus on **audience and purpose** and effectively engages the audience by providing relevant background information necessary to anticipate its needs.
- The writer consistently **develops ideas** with depth and complexity to provide insight, support, and clarification of the topic. The writer consistently develops ideas using appropriate and effective examples, details, facts, explanations, descriptions, or arguments. In grades 5 and 6, writers may address counterclaims in support of opinion and argument; in grades 8, 10 and 11, counterclaims are addressed effectively to help support arguments. The writer may use a variety of techniques or approaches.
- The writer consistently **organizes** the writing by using a logical progression of ideas that flows within and between paragraphs. The writer consistently uses a **variety of sentence lengths and structures.** The writing includes a variety of transitional words and phrases that connects ideas and guides the reader. The writer uses appropriate organizational techniques (e.g., comparison/contrast, cause/effect, order of importance, reasons/explanations).
- The writer maintains an appropriate voice or tone. The writer consistently **chooses words** that are appropriate to the intended audience and purpose of the writing. The writer consistently uses correct **grammar, usage, and mechanics** (e.g., spelling, punctuation, capitalization) to communicate effectively and clarify the writing.

3 Points:
Writers at this score point level display adequate writing skill, resulting in effective, though not consistent, communication.

- The writer adequately establishes focus on the intended **audience and purpose**, but may not consistently maintain this focus, losing sight of audience or purpose on occasion. The writer provides adequate background information that generally anticipates audience needs.
- The writer **develops ideas** with adequate support, and clarification of the topic through examples, details, facts, explanations, descriptions, or arguments. In supporting arguments and opinions, the writer in grades 5 or 6 may address counterclaims; the writer in grades 8, 10 and 11 addresses or considers counterclaims. The writer may use different techniques or approaches, but some are less successful than others; one technique may be prominent.
- The writer adequately **organizes** the writing by using a logical progression of ideas that generally flows from idea to idea, though connections between some ideas are less clear on occasion. The writer displays **variety in sentence lengths and structures.** The writing includes transitional words and phrases that generally guide the reader. The writer generally maintains organizational techniques, but organization and connection of ideas may become less clear on

occasion.

- The writer may have occasional lapses in language that cause voice or tone to weaken. The writer **chooses words** that are generally appropriate for the intended audience and writing purpose. The writer adequately demonstrates correct **grammar, usage, and mechanics** (e.g., spelling, punctuation, capitalization) to communicate A few errors may occur that do not impede understanding.

2 Points:
Writers at this score point level display developing writing skill, resulting in less effective communication.

- The writer identifies a generalized **purpose or audience** but does not maintain focus on both. Instead, the writer focuses more on the task (creating a letter, speech, etc.) than the actual purpose or intended audience. Irrelevant or inconsistent background information demonstrates a general lack of awareness of audience needs.
- The writer demonstrates inconsistent **development of ideas** often presenting facts (sometimes in isolation from one another) with little insight, interpretation, or clarification. The writer provides minimal or irrelevant examples and/or details for support. The writer in grades 8, 10, and 11 may attempt to address counterclaims in support of arguments or is unsuccessful in the attempt. If the writer attempts to use different techniques or approaches, their relation to the writing purpose may be unclear.
- The writer demonstrates some attempt at **organization**, but often places ideas in an unclear order that disrupts the natural flow or cohesion. The writer occasionally uses varied sentence structures, but these appear alongside mostly **simple sentences**. Transitions are simple and infrequent. The writer may use organizational strategies inappropriately or ineffectively, such as attempting to use a comparison when it is not warranted.
- The writer often uses language that causes voice or tone to weaken or emerge only on occasion. The writer occasionally chooses appropriate **words**, but these appear alongside language that is simple or inappropriate for the intended audience or purpose. Frequent errors in **grammar, usage, and mechanics** (e.g., spelling, punctuation, capitalization) appear alongside occasional control of these features and may impede understanding of the text.

1 Point:
Writers at this score level demonstrate little or no writing skill, resulting in mostly ineffective communication.

- The writer may identify a general topic but demonstrates little or no awareness of **purpose or audience**. The writer does not provide background or show awareness of the needs of the audience.
- The writer gives little or no purposeful **development of ideas,** interpretation, insight or clarification. The writer provides no examples and/or details for support or the support is inaccurate or irrelevant. The writer in grades 8, 10, 11 does not address counterclaims in support of argument or opinion.
- The writer offers little or no **organizational structure**, placing ideas in no logical order. The writer uses little if any **variety in sentence structures.** Ineffective or absent paragraph divisions create a lack of cohesion. Few, if any, transition words or phrases are used.
- The writer's tone or voice is either inappropriate or absent. The writer uses simple or inappropriate **words**. Errors that appear in **grammar, usage,** and **mechanics** (e.g., spelling, punctuation, capitalization) impede understanding of the text.