Kentucky Performance Rating for Educational Progress

2012–13 Technical Manual

Version 1.3

Table of Contents

LIST OF TABLES **5**

1. BACKGROUND **6**

KENTUCKY INSTRUCTIONAL RESULTS INFORMATION SYSTEM (1992-1998) 6
COMMONWEALTH ACCOUNTABILITY TESTING SYSTEM (1998-2010) 6
UNBRIDLED LEARNING (2010-PRESENT) 7
ORGANIZATIONS AND GROUPS INVOLVED 7
Kentucky Department of Education 7
Kentucky Educators 8
School Curriculum, Assessment, and Accountability Council 8
National Technical Advisory Panel on Assessment and Accountability 8
Contractors 9
KENTUCKY PERFORMANCE RATING FOR EDUCATIONAL PROGRESS ASSESSMENT PROGRAM 9
Reading 10
Mathematics 10
Science 10
Social Studies 10
Language Mechanics 10
On-Demand Writing 10

2. TEST DEVELOPMENT **12**

K-PREP CONTENT AND COMMON CORE ALIGNMENT 12
ITEM DEVELOPMENT 13
Item Specifications 13
Item Writing 14
Content Advisory Committees 15
Bias and Sensitivity Review 15
Item Editing 15
SCORING GUIDES 15
FORMS DEVELOPMENT 15
Test Design and Blueprints 15
Statistical Guidelines 18
Field-testing 18
TEST BOOKLET DESIGN 19
BRAILLE AND LARGE PRINT TEST MATERIALS 20

3. TEST ADMINISTRATION **21**

TEST ADMINISTRATION WINDOW 21
TEST MAKE-UP PROCEDURES 21
ELIGIBILITY REQUIREMENTS AND EXEMPTIONS 21
ACCOMMODATIONS 22
TEST ADMINISTRATION PROCEDURES 22
District Assessment Coordinators 22
District and Building Assessment Coordinators' Manual 23
Test Administrators' Manual 23
Interpretive Guide 23
TEST SECURITY 23

4. REPORTS **24**

Appropriate Uses for Scores and Reports 24
Individual Student Report 24
Kentucky Performance Report 24
Description of Scores 24
Raw Score 24
Scale Score 24
Student Performance Level 25
National Percentile Rank 25
Lexiles and Quantiles 25
Description of Reports 26
Student Report 26
School Listing Report 26
Kentucky Performance Report 26
Cautions for Score Interpretations and Use 27
Understanding Measurement Error 27
Interpreting Scores at Extreme Ends of the Distribution 27
Limitations When Comparing Scale Scores at Reporting Group Levels 27
Inappropriateness of Comparing Scale Scores Between Content Tests 27
Program Evaluation 27

5. PERFORMANCE STANDARDS 29

Performance Level Descriptions and College/Career Readiness 29
K-PREP and College/Career Readiness 29
On-Demand Writing 30
Science and Social Studies 31

6. ITEM ANALYSES 33

Item Mean Scores 33
Item-Test Score Correlations 33
Differential Item Functioning 34
Item Response Theory 35
On-Demand Writing Item Analysis 36

7. SCALING 38

Rationale 38
Measurement Models 38
Process 38
Overview 39
Quality Control 39
Scaled Scores 40
Transformation of Raw Scores 40
Considerations and Limitations 42
Results 44
Lexiles and Quantiles 44

8. EQUATING 45

Rationale 45
Process 45
Linking Items 45
Analysis 46
Field-test Item Calibration 47

9. RELIABILITY 48

Definition of Reliability 48
Estimating Reliability 49
Test-Retest Reliability Estimation 49
Alternate Forms Reliability Estimation 49
Internal Consistency Reliability Estimation 49
Domain Reliability Estimation 50
Standard Error of Measurement 50
Use of the Standard Error of Measurement 50
Conditional Standard Error of Measurement 51
Scoring Reliability for Open-Ended Items 51
Reader Agreement 51
Score Resolutions 52
Reliability of Performance Level Categorization 52
Accuracy and Consistency 52
Calculating Accuracy 53
Calculating Consistency 54
Calculating Kappa 54

10. VALIDITY 56

Argument-Based Approach to Validity 56
Scoring 57
Generalization 57
Extrapolation 57
Implication 58
Validity Argument Evidence for the Kentucky Assessment 58
Scoring 58
Generalization 59
Extrapolation 60
Implication 61
Summary of Validity Evidence 61

11. PERFORMANCE SCORING 62

Rubric Creation 62
Rangefinding 63
Scoring Process 63
Recruitment 63
Training 64
Quality Control 65
Security 66

12. QUALITY CONTROL PROCEDURES 67

Test Construction 67
Non-Scannable Documents 67
Data Preparation 67
Production Control 68
Scanning and Editing 68
Performance Scoring 69
Equating 69
Scoring and Reporting 70

GLOSSARY OF TERMS 71

REFERENCES 73

List of Tables

Table 2.1 K-PREP Reading Test Blueprint 16
Table 2.2 K-PREP Mathematics Test Blueprint.....17
Table 2.3 K-PREP Science Test Blueprint.....17
Table 2.4 K-PREP Social Studies Test Blueprint.....17
Table 2.5 2012 K-PREP On-Demand Writing Test Blueprint.....18
Table 2.6 2013 K-PREP On-Demand Writing Test Blueprint.....18
Table 2.6 K-PREP Test Booklet by Grade.....19
Table 5.1 Reading and Mathematics Final Cut Points and Impact Data 30
Table 5.2 ODW Final Performance Level Cut Points 31
Table 5.3 ODW Final Round Impact Data 31
Table 5.4 2011 Science and Social Studies Performance Level Distribution (KCCT) .32
Table 5.5 2012 Science and Social Studies Cut Points and Impact Data (K-PREP) ...32
Table 6.1 Criteria for Item Fit Statistics 36
Table 7.1 Proficient Cut Points for Derived Scaled Scores 40
Table 7.2 Raw Score to Scale Score Conversion.....43
Table 7.3 Scaled Scores by Performance Level 43
Table 8.1 2013 Linking Items.....46
Table 8.2 Unstable Linking Items.....47
Table 9.1 Example Accuracy Classification Table 53
Table 9.2 Example Accuracy Classification Table for Proficient Cutpoint.....54
Table 9.3 Example Consistency Classification Table.....54

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-Present)

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 Core Academic Standards (KCAS) were adopted to outline the minimum content required for all students before graduating from high school. For reading, mathematics and writing, the content standards are 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 KCAS. K-PREP is a blend of norm-referenced and criterion-referenced test content that provides 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 consists of test content from the Stanford Achievement Test Series, Tenth Edition, hereafter Stanford 10, and uses existing score norms to report Kentucky student achievement on a national scale (see chapter 4, "Reports").

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 Unbridled Learning accountability model and its components. KDE consists of smaller organizations that provide

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

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 Assessment Design and Implementation (data and statistics) and the Division of Support and Research (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, Inc.
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 are rooted in the Common Core State Standards, which have been 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. Reading comprehension is also assessed through the Stanford 10 reading comprehension subtest.

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. In addition, the Stanford 10 problem solving subtest is included on the K-PREP mathematics assessments.

Science

For the Science tests, the standards are 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. The Stanford 10 science subtest is also included on K-PREP and emphasizes unifying themes and concepts of science.

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. The Stanford 10 social science subtest is also included on K-PREP to measure the concepts important for the development of citizenship.

Language Mechanics

The Stanford 10 language subtest is used as the K-PREP "Language Mechanics" test to measure word- and sentence-level skills and whole paper skills in mechanics and expression.

On-Demand Writing

On-Demand Writing (hereafter, writing) assesses writing skills through goals set forth through the Common Core State 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 Yearbook.

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 along with emphases on common core, content and statistical guidelines considered, and test booklet design.

K-PREP Content and Common Core Alignment

One emphasis during K-PREP content development—item and passage development—is alignment to the Kentucky Core Academic Standards (KCAS). KCAS is the umbrella for all state standards that includes Kentucky's adoption of Common Core State Standards, hereafter Common Core, in reading, mathematics, writing, and state standards for science and social studies. The K-PREP testing contractor began item development activities by evaluating items developed to assess Common Core State Standards (CCSS) by a previous Kentucky state assessment contractor. This evaluation was used to create item development plans to bolster the item pool such that the CCSS 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 CCSS. 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 CCSS 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 Common Core; the Common Core standards are included with key aspects highlighted for training purposes. Training on Common Core 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 Common Core may change, dictated by national perspectives.

During item writer training, the testing contractor presents the Common Core 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 Common Core 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 Common Core. 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 Common Core differently than what was intended by the item writers. The assessment content, alignments, and reviews by the testing contractor are prepared for review by KDE.

KDE reviews the assessment content and alignments to Common Core for appropriateness. Content specialists review each piece of assessment content and recommend modifications to Common Core alignments, as necessary. During this review, KDE and the testing contractor may discuss differences in interpretations of Common Core and appropriate solutions for assessing Kentucky's students. Once KDE has reviewed and approved the Common Core 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 Common Core alignment, in addition to reviews for content appropriateness. Changes to Common Core 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 Common Core alignment. Changes in Common Core alignment from committee review must be consistent within the general scope of Common Core alignment. Once changes in Common Core 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 Common Core 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 *Common Core State Standards*. For 2011-2012, Science and Social Studies had not been established within the common core framework; therefore, item development efforts did not take place for these subject areas.

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 common core standard;
- Items must be unique in approaches to assessing standards;
- Items must be grammatically (and/or mathematically) correct.

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.

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 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 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.

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 was included which added 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 2.1 through Table 2.4 provides the test blueprints used for constructing the K-PREP tests in Reading, Mathematics, Science, and Social Studies.

Table 2.1 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 2.2 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 2.3 K-PREP Science Test Blueprint

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 2.4 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 <u>and</u> within grade and will vary across years of the Writing

assessment. Table 2.5 and 2.6 show the test blueprint used for the 2012 and 2013 On-Demand Writing assessments.

Table 2.5 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 2.6 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

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.

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. 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.

Reading and Mathematics

Field-test items for Reading and Mathematics were included on the test forms administered in 2013. For Reading, field-test items were appended at the end of the operational test form—after the last operational item; for Mathematics, field-test items were placed in designated slots within the operational form. In both subjects, 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—were field-tested as required for maintaining a suitable pool of items for subsequent test forms. Performance data from the field-item items were 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. For example, third grade test booklet contains Reading and Mathematics only, but the fourth grade test booklet contains Reading, Mathematics, Science, and Language Mechanics. Table 2.6 shows the content areas and order of appearance in the test booklet by grade.

Table 2.6 K-PREP Test Booklet by Grade

Grade							
3	4	5	6	7	8	10	11
R	R	R	R	R	R	ODW	ODW
M	M	M	M	M	M		
	Sc	SS	LM	Sc	SS		
	LM	ODW	ODW		ODW		

R = Reading, M = Mathematics, Sc = Science, SS = Social Studies, LM = Language Mechanics, ODW = On-Demand Writing

For each content area, except for Language Mechanics and On-Demand Writing, the SAT10 items are presented first, followed by the items developed for K-PREP. Language Mechanics is only composed of SAT10 items, and On-Demand Writing does not contain SAT10 items.

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, 10, 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

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

- Those qualifying for an "extraordinary circumstance" exemption (see below).

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). 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 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. Also, the district and state summary reports can be used to determine accountability ratings for schools and districts.

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, except Writing. The range of scaled scores for the domains of each test is also set to 100-300. Scaled scores were not developed for Writing.

Student Performance Level

Student achievement on K-PREP is defined by *performance levels*, within a classification system of achievement from low ability to high ability. 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.

National Percentile Rank

K-PREP includes a norm-referenced component captured through Pearson's Stanford Achievement Test Series, *Tenth Edition* (SAT10). The content and make-up of this testing program is discussed in the background section of this report (chapter 1). For the current chapter, though, the focus will be on describing student achievement from this test. The Stanford tests provide many different kinds of scores serving many purposes, but the national percentile rank (NPR) is used on K-PREP score reports to further illustrate student achievement in each subject area. This is a norm-referenced score that indicates the standing of a student's achievement in relation to the performance of students across the nation.

Percentile ranks range from 1 to 99 where the value of 50 reflects average performance. These scores provide a relative standing of a student compared to students of the same level who took the Stanford 10 tests at the same time of year. The rank means that a student's individual performance is as good as or better than the performance of that percentage of students across the nation. For example, an NPR of 68 for a given subject means that the student scored as well or better than 68% of students from the national sample on that subject test.

Percentile ranks are useful to show student performance as compared to other students in a particular reference group. However, they do not reflect actual amounts of student achievement. Additionally, they should not be treated as equal units across the scale. That is, the difference in achievement between NPRs of 10 and 20 is not the same as the difference between NPRs of 55 and 65. For this reason, NPRs are best used to interpret individual student position as it relates to the national sample of students.

Lexiles and Quantiles

Lexiles are measures used to describe a person's reading ability; 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 raw score is reported with the corresponding performance level and performance level descriptor. Like the scaled score for the other subject tests, the raw score is shown against the mean raw score at the school, district, and state levels.

The ISR also reports the individual's NPR along with a brief interpretation of the value obtained. 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 (or raw score for Writing), performance level, NPR, 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 SAT10 portion of K-PREP allows student achievement to be grouped into *quartiles* and compared against the national quartiles.

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. For achievement comparisons at the national level, this report provides the percentage of students in each percentile rank quarter at the school, district, and state level, based on the SAT10 portion of K-PREP.

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

[3] The equating of scores will begin with the 2013 test administration.

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.

5. Performance Standards

As part of adopting the common core state standards, hereafter common core, Kentucky joined the Partnership for Assessment of Readiness for College and Careers (PARCC) national consortium and 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 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 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.

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 ability skills 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) 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

The On-Demand Writing assessment does not carry the same explicit goal of college and career readiness as Reading and Mathematics and, therefore, used a different process for determining performance standards. 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 ability 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 ability 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 ability.

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 all of the work and feedback data that they reviewed and discussed throughout the process. Reviewing student work was not a planned part of this task; however educators were allowed to refer back 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.

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

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 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

[4] The score range is 0 to 16.

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.

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%

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 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" ability students. In general, the higher the correlation the better an item is at discriminating among high and low ability 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. 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. 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. DIF procedures include comparing students from the reference and focal groups *at the same ability level*. Therefore, statistical differences between the groups, found through DIF analyses, are not confounded by student ability.

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, however, only determines whether or not a difference exists in performance between the two groups. The Mantel-Haenzsel Delta DIF statistic (MHD: Dorans & Holland, 1993) is computed from the Mantel-Haenzsel 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.

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.

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 ability 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 ability 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 ability. However, this model applies to multiple-choice items only. Given that K-PREP includes constructed-response items, a separate model is required for estimating ability 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:

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

where $P_{ix}(\theta)$ is the probability that a person with ability 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}).

In addition to item difficulty, IRT provides other indices for item analyses, such 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 do 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 outliers and can be easy to diagnose and fix; *infit* mean-square statistics are influenced by the response patterns and are more difficult to diagnose and fix. The values of the mean-square statistics are classified into four groups for interpretation (see Table 6.1).

Table 6.1 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 then 1.0 indicate observed response patterns that are too predictable (model overfit). Values greater than 1.0 indicate unpredictable observed response patterns (model underfit). The IRT parameter estimates—item difficulty and item fit—are summarized in 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.

[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

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 ability parameters, θ. Under IRT, an ability 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 ability parameters are on the same logit scale, although the ability parameter often results in a wider range of values.

Under Rasch modeling, there is one-to-one correspondence of ability parameter to raw score value. In other words, for each possible raw score (total test score) value there is one person ability parameter estimated. For example, if there are 40 raw score points possible on a test, then there will be 40 person ability estimates generated one for each raw score. The ability estimates will also increase from lowest to highest value in relation to the ascending order of the raw scores. As will be demonstrated later in this chapter, the ability 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 ability 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 ability estimates both on a logit scale. Discussed in more detail later in this chapter, the ability 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 ability 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 ability estimate (θ) from Rasch modeling of examinee response data.

The scaled score metric for K-PREP was chosen to range from 100 to 300, for each subject except Writing, 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) was set to 210, and b is the person ability estimate defined as before. This transformation, however, includes θ_p, the person ability estimate identified as the minimum value for *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 and Social Studies, these values were determined by using 2011 KCCT performance data and finding similar performance patterns in the 2012 K-PREP test data. Scaling was not performed for Writing given that it is a two-item test, leaving the raw score—the sum of individual prompt scores—as the most sufficient score for this test. Therefore, the criterion raw score for *Proficient* in Writing was determined at standard setting. 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	0.3172
	7	0.1689
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 ability estimate, from total test performance, and θ_p is the minimum person ability 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 *Physical Science* subdomain in Science. Summary reports contained these comparisons for all reporting categories in Reading, Mathematics, Science, 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.

For 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 were aggregated by schools, districts, and state as in 2012, 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 from 2012 to 2013, 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 ability estimates from which they are derived. For example, a scaled score difference of 5 points can be the result of a small difference in ability 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	108
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	100-192	193-209	210-227	228-300
	7	100-194	195-209	210-230	231-300
Social Studies	5	100-192	193-209	210-233	234-300
	8	100-190	191-209	210-233	234-300

Results

The Yearbook contains the tables of derived scaled scores for each K-PREP assessment. Each table contains the raw scores, ability 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 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 also provided.

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 ability 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 ability and the difficulty of mathematical concepts. In Lexile and Quantile frameworks, the higher measure a person receives, the higher ability 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

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 K-PREP Reading, Mathematics, Science, and Social Studies, 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 the 2013 equating analyses, common items were chosen from the 2012 test forms; the 2012 test forms are the first operational test forms in K-PREP. 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.

Table 8.1 shows the number of common items that were selected for linking from the 2012 tests. 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 Science and Social Studies, there are multiple 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 2013 Linking Items

Subject	Grade(s)	Item Type		
		Multiple-choice	*Short Answer*	*Extended Response*
Reading	3	13	1	--
	4	13	1	
	5	18	1	
	6	17	1	
	7	19	1	
	8	15	1	
Mathematics	3	22	2	--
	4	20	2	
	5	23	2	
	6	23	2	
	7	24	2	
	8	25	2	
Science	4, 7	13	-	1
Social Studies	5, 8	13	-	1

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, 2011) 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. Table 8.2 shows the number of items that were deemed unstable and were dropped from the linking sets prior producing final item parameters and score conversion tables. The majority of linking items, across all grades and content areas, were considered stable for equating the 2013 test forms to the 2012 test forms.

Table 8.2 Unstable Linking Items

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

**One short answer item in grade 8 was removed from the linking set prior to the analysis.*

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

For the Reading and Mathematics test forms, items are included that serve 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 ability and knowledge consistently) has little or no value. Furthermore, the ability 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 ith 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 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 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. 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 ability, 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 $v's$ ability 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. 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.

Tables in the Yearbook give the score frequency distributions for On-Demand Writing. Also presented is the percent agreement among readers. As mentioned above, this check 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.

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.

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. 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 are used for the selection of items to go on the test (except for 2012) 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 given in 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.

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 ability 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. 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." The Yearbook has tables reporting the conditional SEM for each scale score point and 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.

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. As described in chapter 8, "Equating," 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 is evidence of students giving their full effort. 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, 10, 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 ability 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 the Yearbook.

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.

For the other K-PREP assessments, equating is performed to address variability in test form difficulty across years (see chapter 8). This is accomplished through the use of items common across two parallel test forms. For the On Demand Writing assessment, however, the test design does not include repeated items across two

test administrations. In addition, the test design includes different modes of writing (e.g., narrative, opinion, etc.) tested from year to year, by grade. In some respects, this presents a challenge in test forms being considered "parallel" in content and/or difficulty across years. Currently, equating analyses are not conducted for the writing assessment. Thus, test scores are not explicitly controlled for any variability of test difficulty (writing prompts) across test administrations. In lieu of equating, though, the quality control mechanisms listed above serve for monitoring and controlling scoring consistency across test administrations for the writing assessment.

The Yearbook contains an appendix summarizing the interrater agreement rates and score point distribution for constructed response items—short answer, extended response, and writing prompts—from the 2013 tests. This summary includes statistics from the current test administration as well as statistics from the previous test administration in which the items were administered (highlighted in blue). For Reading, Mathematics, Science, and Social Studies, the previous administration for the items in the summary is 2012 (operational or field test). For Writing, the previous statistics are from the 2011 stand-alone field test. The Yearbook also 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.

Glossary of Terms

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

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

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

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

Criterion-referenced test: Test that reports examinee performance according to established criteria.

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 ability 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: A measurement theory that prescribes relationships of item difficulty and examinee ability 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.

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: The process of 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.

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 unobservable quantity, in classical test theory, often hypothesized in estimates of reliability.

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

Validity: A framework of validating 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.

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.

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.

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. (2011). *WINSTEPS* Rasch measurement computer program. Chicago: Winsteps.com.

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

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

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