



# Linking the K-PREP Math Test with the Quantile<sup>®</sup> Framework

*A Study to Link the Kentucky Performance Rating for  
Educational Progress Math Test with The Quantile<sup>®</sup>  
Framework for Mathematics*

November, 2012  
Redacted

*Prepared by MetaMetrics for:*

**Kentucky Department of Education  
Office of Assessment and Accountability  
500 Mero Street, C.P.T., 18th floor  
Frankfort, KY 40601**



**MetaMetrics<sup>®</sup>**

1000 Park Forty Plaza Drive, Suite 120  
Durham, North Carolina 27713  
[www.MetaMetricsInc.com](http://www.MetaMetricsInc.com)  
[www.Quantiles.com](http://www.Quantiles.com)



## Table of Contents

<b>Introduction.....</b>	<b>1</b>
<b>The Quantile Framework for Mathematics .....</b>	<b>3</b>
Structure of the Quantile Framework.....	3
Quantile Item Bank Development.....	8
Calibration of Items on the Quantile Scale .....	21
QTaxon Quantile Measures.....	24
Validation of The Quantile Framework for Mathematics .....	26
<b>The K-PREP - Quantile Framework Linking Process.....</b>	<b>37</b>
Description of the Assessments.....	37
Study Design .....	41
Analysis of the K-PREP Math Test/Quantile Linking Test Sample.....	41
Linking the K-PREP Math Scale with the Quantile Scale .....	47
Validity of the K-PREP Math Test - Quantile Link .....	49
Principal Components Analysis .....	54
<b>Quantile Framework and Instruction .....</b>	<b>58</b>
<b>Notes .....</b>	<b>70</b>
<b>References.....</b>	<b>71</b>
<b>Appendix A .....</b>	<b>A-1</b>
<b>Appendix B.....</b>	<b>B-1</b>



## Introduction

Often it is desirable to convey more information about test performance than can be incorporated into a single primary score scale. When two score scales are linked, the linkage can be used to provide a context for understanding the results of one of the assessments. It is often hard to explain what mathematical skills and concepts a student actually understands based on the results of a mathematics test. Parents typically ask the question, “Based on my child’s test results, what math problems can he or she understand and how well?” Once a linkage is established with an assessment that is reported as specific concepts and skills, then the results of the assessment can be explained and interpreted in the context of the specific concepts and skills that a student will likely understand.

Auxiliary score scales can be used to “convey additional normative information, test-content information, and information that is jointly normative and content based” (Petersen, Kolen, and Hoover, 1989, p. 222). One such auxiliary scale is The Quantile® Framework for Mathematics, which was developed to appropriately match students with materials at a level where the student has the background knowledge necessary to be ready for instruction on new mathematical skills and concepts.

The Quantile Framework for Mathematics takes the guesswork out of mathematics instruction. It serves as a hands-on tool which demonstrates which mathematics skills a learner has likely learned and which ones require further instruction. Teachers can also use the Quantile® Framework to determine a student’s readiness to learn more advanced skills. Because the Quantile Framework uses a common, developmental scale to measure both student mathematical achievement and task difficulty, educators can also determine how well a student is likely to be able to solve more complex problems (if provided with targeted instruction). The Quantile Framework includes the Quantile® measure and the Quantile® scale. The Quantile Framework targets instruction, forecasts understanding, and helps improve mathematics instruction and achievement by placing the mathematics curriculum, the materials to teach mathematics, and the students themselves on the same scale.

The Quantile Framework for Mathematics can be used to:

- Monitor student mathematics progress.
- Forecast student performance on end-of-year assessments.
- Match students with appropriate materials at their level.
- Determine if a student is ready for a new mathematics skill or concept.
- Link big mathematical concepts with state curriculum objectives.
- Identify student strengths and weaknesses.

- Understand the prerequisite skills needed to learn more advanced concepts in mathematics.
- Adapt instructional methods in the classroom to ensure a greater level of understanding and application.

The Quantile Framework for Mathematics is a unique resource for accurately estimating a student's ability to think mathematically and matching him/her with appropriate mathematical content. With this valuable information in the hands of educators, instruction can be more accurately tailored to the mathematical achievement of individual students. The structure of the Quantile Framework is organized around two principles – (1) mathematics and mathematical achievement are developmental in nature and (2) mathematics is a content area.

Linking assessment results with the Quantile Framework provides a mechanism for matching each student with materials on a common scale. It serves as an anchor to which resources, concepts, skills, and assessments can be connected allowing parents, teachers, and administrators to speak the same language. By using the Quantile Framework, the same metric is applied to the materials the children use, the tests they take, and the results that are reported. Parents often ask questions like the following:

- How can I help my child become better at mathematics?
- How do I challenge my child to think mathematically?

Questions like these can be challenging for parents and educators. By linking the Kentucky Performance Rating for Educational Progress (K-PREP) Math scale with the Quantile Framework, educators and parents will be able to answer these questions and will be better able to use the results from the tests to improve instruction and to develop each student's level of mathematics understanding.

This research study was designed to determine mathematics achievement levels that can be matched with mathematical skills and concepts based on test results on the K-PREP Math Test. The study was conducted by MetaMetrics, Inc. in collaboration with the Kentucky Department of Education (KDE) (contract PON2 540 1000002689 2). The primary purposes of this study were to:

- provide tools (Math@Home, Quantile Teacher Assistant, and Math Skills Database) and information that can be used to answer questions related to standards, student-level accountability, test score interpretation, and test validation;
- create conversion tables for determining Quantile measures from the scores on the K-PREP Math Test; and
- produce a report that describes the linking analysis procedures.

## The Quantile Framework for Mathematics

Just as for reading, there are dozens of tests of mathematics ability measuring a common construct and all reporting the results in proprietary, non-exchangeable metrics. The benefits of having a common supplemental metric to describe mathematics ability include the following:

- (1) Individual growth trajectories spanning the educational experience can be developed because the Quantile scale is developmental in nature and spans this range.
- (2) Various state definitions of grade-level proficiency can be compared by re-expressing scores on a common scale.
- (3) Textbook publishers can build links between mathematics curricula and major mathematics tests.
- (4) Test publishers can develop classroom/interim assessments that can link to the major mathematics tests and forecast how likely the student is to meet the state performance standards.
- (5) The classroom teacher can link his or her day-to-day instructional needs to the year-to-year needs of a state-level accountability system.

The Quantile Framework consists of a common supplemental metric – the Quantile – that is employed to scientifically measure a student’s ability to think mathematically and his or her mathematics achievement and to locate the student in a taxonomy of mathematical skills, concepts, and applications. In order to develop the Quantile Framework, several tasks were undertaken: (1) develop a structure of mathematics that spans the developmental continuum from first grade content through Algebra I, Geometry, and Algebra II content, (2) develop a bank of items that have been field tested, (3) develop the Quantile scale (multiplier and anchor point) based on the calibrations of the field-test items, and (4) validate the measurement of mathematics ability as defined by the Quantile Framework.

### Structure of the Quantile Framework

In order to develop a framework of mathematical ability, first a structure needs to be established. The structure of the Quantile Framework is organized around two principles – (1) mathematics and mathematical ability are developmental in nature and (2) mathematics is a content area.

*Developmental Nature of Mathematics.* The developmental nature of mathematics over time describes the increase in sophistication of the problems that can be addressed and the increase in the integration of skills and content to address these problems. The

National Research Council (2001, 2002) described mathematical proficiency as “...five intertwined strands: (1) understanding mathematics; (2) computing fluently; (3) applying concepts to solve problems; (4) reasoning logically; and (5) engaging with mathematics, seeing it as sensible, useful, and doable” (p. 1). Geary and Hamson (2002) observed that much of mathematics can be understood as an interlocking triad of competencies: conceptual competence, procedural competence, and utilization competence. In short, these competencies refer, respectively, to (1) understanding the natural language of mathematics, (2) knowing how to read mathematical expressions and employ algorithms to solve decontextualized problems, and, finally, (3) knowing why the conceptual and procedural knowledge is important and how and when to apply it. The descriptions of these three competencies follow.

- A. *Vocabulary of Mathematics*. This aspect concerns the recognition of a concept either verbally or pictorially. Concepts are drawn from the mathematical content (e.g., alternate interior angles, mean, tangent) and the mathematical process (e.g., compare, estimate, etc.) strands of the National Council of Teachers of Mathematics (NCTM) framework, and include contexts (e.g., sales tax, commission) and measurement concepts (e.g., time, weight). The NCTM Standards describe this as the language of mathematics.
- B. *Procedures of Mathematics*. This aspect concerns being able to apply mathematical procedures in a controlled environment (decontextualized). Procedural items ask the student to perform operations and can include graphics. For example, (1) simplifying  $(3x + 2)(4x - 8)$ ; or (2) identifying which three angles could form a triangle knowing that the sum of the angles of a triangle equals  $180^\circ$ . Procedures of mathematics can also be described as algorithmic, symbolic computation, and skills.
- C. *Applications of Mathematics*. This aspect involves being able to apply a mathematical procedure to solve a problem (contextualized). Application items ask the student to apply operations and concepts and can include graphics. For example, (1) determining how many cars are needed to transport the class to the museum knowing that each car can hold four students; or (2) determining how much soil is needed for a garden plot that is 3 feet wide, 6 feet long, and 8 inches deep. Applications of mathematics can also be described as problem solving, reasoning, projects, and experiences.

MetaMetrics recognizes that in order to adequately address the scope and complexity of mathematics, multiple proficiencies/competencies must be utilized. Just as the “math wars” have brought to the forefront the various aspects of mathematics instruction, we must also address these same issues. On the issue of the “math wars,” Richard Riley stated “We are suffering here from an ‘either-or’ mentality. As any good K-12 teacher will tell you, to get a student enthused about learning, you need a mix of information

and styles of providing that information. You need to provide traditional basics, along with more challenging concepts, as well as the ability to problem-solve, and to apply concepts in real world settings” (Starr, 2002). The Quantile Framework is an effort to recognize and define a basis for this “mix of information and styles” in the developmental context of mathematics instruction.

*Content of Mathematics.* A strand is a major subdivision of mathematical content. The strands describe what students should know and be able to do. The five strands of the Quantile Framework are based on the five Content Standards in the National Council of Teachers of Mathematics framework (NCTM, 2000), which are as follows:

1. *Number and Operations.* The development of number sense. Students with number sense naturally decompose numbers, use particular numbers as referents, solve problems using the relationships among operations and knowledge about the base-ten system, estimate a reasonable result for a problem, and have a disposition to make sense of numbers, problems, and results. Includes computational fluency.

Instructional programs should enable all students to –

- Understand numbers, ways of representing numbers, relationships among numbers, and number systems;
- Understand meanings of operations and how they relate to one another;
- Compute fluently and make reasonable estimates.

2. *Geometry.* The study of geometric shapes and structures; specifying their characteristics and relationships. A means to interpret and reflect on our physical environment and serve as tools for the study of other topics.

Instructional programs should enable all students to –

- Analyze characteristics and properties of two- and three-dimensional geometric shapes and develop mathematical arguments about geometric relationships;
- Specify locations and describe spatial relationships using coordinate geometry and other mathematical systems;
- Apply transformations and use symmetry to analyze mathematical situations;
- Use visualization, spatial reasoning, and geometric modeling to solve problems.

3. *Algebra/Patterns and Functions.* The relationships among quantities, the use of symbols, the modeling of phenomena, and the mathematical study of change. Instructional programs should enable all students to –

- Understand patterns, relations, and functions;

- Represent and analyze mathematical situations and structures using algebraic symbols;
- Use mathematical models to represent and understand quantitative relationships;
- Analyze change in various contexts.

4. *Data Analysis and Probability.* The collection, analysis, and interpretation of data.

Instructional programs should enable students to—

- Formulate questions that can be addressed with data and collect, organize, and display relevant data to answer them;
- Select and use appropriate statistical methods to analyze data;
- Develop and evaluate inferences and predications that are based on data;
- Understand and apply basic concepts of probability.

5. *Measurement.* The assignment of a numerical value to an attribute of an object.

Instructional programs should enable students to—

- Understand measurable attributes of objects and the units, systems, and processes of measurement;
- Apply appropriate techniques, tools, and formulas to determine measurements.

*The QTaxon.* Within the Quantile Framework, a “QTaxon” describes a specific mathematical skill and is used to annotate the Quantile scale. For example, a QTaxon under the Numbers and Operations strand is “Model and identify the place value of each digit in a multi-digit numeral to the hundredths place;” and a QTaxon under the Geometry strand is “Identify and distinguish among similar, congruent, and symmetric figures; name corresponding parts.” The content taxonomy of QTaxons used with the Quantile Framework was developed during the spring of 2003 for grades 1 through 8, Algebra I, and Geometry. The framework was extended to Algebra II and revised during the summer and fall of 2003. The first step in developing a content taxonomy was to review the curricular frameworks from the following sources:

- National Council of Teachers of Mathematics (NCTM).
- National Assessment of Educational Progress: 2005 Pre-Publication Edition.
- North Carolina Standard Course of Study (Revised in 2003 for grades kindergarten through high school).

- California Mathematics Framework and state assessment blueprints: *Mathematics Framework for California Public Schools: Kindergarten through Grade Twelve* (2000 Revised Edition); *Mathematics Content Standards for California Public Schools: Kindergarten through Grade Twelve* (December 1997); Blueprints document for the Star Program California Standards Tests: Mathematics (California Department of Education, adopted by SBE 10/9/02), and sample items for the California Mathematics Standards Tests (California Department of Education, January 2002).
- Florida Sunshine State Standards: Sunshine State Standards Grade Level Expectations for Mathematics, grade 2 through Mathematics. The Sunshine State Standards “are the centerpiece of a reform effort in Florida to align curriculum, instruction, and assessment.” They identify what students should know and be able to do for the 21<sup>st</sup> century. Publishers are required to correlate instructional materials submitted for state adoption to the standards.
- Illinois: Illinois teachers for Illinois schools developed The Illinois Learning Standards for Mathematics. Their Goals 6 through 10 emphasize the following: numbers and operations, measurement, algebra, geometry, and data analysis and statistics – *Mathematics Performance Descriptors, Grades 1-5 and Grades 6-12* (2002).
- Texas Essential Knowledge and Skills: Texas Essential Knowledge and Skills for Mathematics (TEKS) were adopted by the Texas State Board of Education and became effective on September 1, 1998. The Texas Essential Knowledge and Skills (TEKS), the state-mandated curriculum, was “specifically designed to help students to make progress ... by emphasizing the knowledge and skills most critical for student learning” (TEA, 2002b, p. 4).

The Texas Assessment of Knowledge and Skills (TAKS) was mandated by the 76th Texas Legislature in 1999 and was administered for the first time during the 2002-2003 school year (TEA, 2002a). The TAKS was developed to assess the TEKS and ask questions in more authentic ways. The TAKS test objectives, “ ‘umbrella statements’ generated by TEA staff with input from educators,” were used to develop the items (p. 2). These statements serve as headings under which the TAKS are meaningfully grouped. The TAKS measures the statewide curriculum in reading at grades 3-9; in writing at grades 4 and 7; in English Language Arts at grades 10 and 11; in mathematics at grades 3-11; in science at grades 5, 10, and 11; and in social studies at grades 8, 10, and 11. The Spanish TAKS is administered at grades 3 through 6. Satisfactory performance on the TAKS at Grade 11 is prerequisite to a high school diploma.

The review of the content frameworks resulted in the development of a list of QTaxons spanning the content typically taught in kindergarten through Algebra I, Geometry and

Algebra II. Each QTaxon is aligned with one of the five content strands. The QTaxons can be viewed and searched at [www.Quantiles.com](http://www.Quantiles.com). Each QTaxon consists of a description of the content, a content identification number (C\_ID), the grade at which it typically first appears (Grade), and the strand it is associated with (1 = Numbers and Operations, 2 = Geometry, 3 = Algebra/Patterns & Functions, 4 = Data Analysis & Probability, and 5 = Measurement).

The Quantile Framework map (Appendix A) presents a picture of the construct of mathematics ability. The map is organized by the five strands and describes the development of mathematics from basic skills to sophisticated problem solving. Exemplar QTaxons and problems are used to annotate the Quantile scale and the strands. QTaxons are located on the Quantile scale at the point corresponding to the mean of the ensemble of items addressing that QTaxon from two large, national studies (Quantile Framework field study and *PA Series* Math field study described later in this document). Items are located on the Quantile scale corresponding to their Quantile measure based on the Quantile Framework field study.

## **Quantile Item Bank Development**

The second step in the process of developing The Quantile Framework of Mathematics was to develop and field test a bank of items that could be used in future linking studies. Item bank development for the Quantile Framework went through several stages – content specification, item writing and review, field-testing and analyses, and final evaluation.

*Item Specification and Development.* Based on the list of QTaxons aligned to the five strands, QTaxons were identified as typically being taught at a particular grade level. The curricular frameworks from Florida, North Carolina, Texas, and California were synthesized to identify the QTaxons instructed and/or assessed at each grade level. If a QTaxon was included in any state framework it was included in the list of QTaxons for which items were to be developed for use with the Quantile Framework field study.

During the summer and fall of 2003, over 1,400 items were developed to assess the QTaxons associated with content in grades 1 through Algebra II. The items were written and reviewed by mathematics educators trained to develop multiple-choice items (Haladyna, 1994). The items for the pool were specified by both strand and QTaxon. At least three items were written for each QTaxon within each grade.

With the current increased focus on authentic assessment and solving problems in context using real-world applications, mathematics items now tend to require more reading. While the vocabulary specific to mathematical content is used (e.g., congruent), every attempt is made to have the non-content vocabulary below the grade level.

*Item Writer Training.* Item writers were experienced teachers and item-development specialists who had experience with the everyday mathematical ability of students at various levels. The use of individuals with these types of experiences helped to ensure that the items were valid measures of mathematics. Item writers were provided with training materials concerning the development of multiple-choice items and the Quantile Framework. The item writing materials also contained incorrect and ineffective items that illustrated the criteria used to evaluate items and corrections based on those criteria. The final phase of item writer training was a short practice session with three items.

Item writers were also given additional training related to “sensitivity” issues. Part of the item writing materials addressed these issues and identify areas to avoid when selecting passages and developing items. These materials were developed based on material published on universal design and fair access – equal treatment of the sexes, fair representation of minority groups, and the fair representation of disabled individuals.

Items were reviewed and edited by a group of specialists that represented various perspectives – test developers, editors, and curriculum specialists. These individuals examined each item for sensitivity issues and for the quality of the response options. During the second stage of the item review process, items were either “approved,” “approved with edits,” or “deleted.”

*Linking and Field-Test Design.* Tests were developed for ten levels: Levels 2 through 8 were aligned with the typical content taught in grades 2 through 8, Level 9 was aligned with the typical content taught in Algebra I, Level 10 was aligned with the typical content taught in Geometry, and Level 11 was aligned with the typical content taught in Algebra II. For each level, three forms were developed with each form containing 30 items.

First, each form consisted of 22 unique items that were targeted specifically for the grade level. Across the three grade-level forms, 66 unique items were identified. These items were selected from a pool of items that covered the content of a particular grade level. For grades 2 through 8, 22 items were from Strand 1 – Numbers and Operations and 11 items were from each of the other four strands (Strand 2 – Geometry, Strand 3 – Algebra/Patterns & Functions, Strand 4 – Data Analysis & Probability, and Strand 5 – Measurement). For Algebra I and Algebra II, the primary focus of the 66 items was Strand 3 – Algebra/Patterns & Functions (33 items, 50%) with the remaining items evenly distributed across the other four strands; and for Geometry, the primary focus of the 66 items was Strand 2 – Geometry (33 items, 50%) with the remaining items evenly distributed across the other four strands.

Next, for each grade level, 12 of the 66 grade-level items were designated “linking” items. For each grade level set, 4 items were from Strand 1 – Numbers and Operations and 2 items were from each of the other four strands (Strand 2 – Geometry, Strand 3 – Algebra/Patterns & Functions, Strand 4 – Data Analysis & Probability, and Strand 5 – Measurement). For Algebra I and Algebra II, 6 items (50%) were from Strand 3 – Algebra/Patterns & Functions with the remaining six items randomly selected from the other four strands. For Geometry, 6 items (50%) were from Strand 2 – Geometry with the remaining six items randomly selected from the other four strands. For Grade 1, only the “linking” set of items was included in the field-test item pool.

The linking set of items for a grade level was used to link (1) the field-test forms within the grade, (2) the field-test forms from the grade below, and (3) the field-test forms from the grade above. The final field tests were comprised of 658 unique items. Two grade 10 forms only had 29 items (one on-grade level item was dropped from each of two forms due to graphics problems).

A common-item test design was employed to vertically link the test levels. In this design, multiple tests are given to non-random groups, and a set of common items is included in the test administration to allow some statistical adjustments for possible sample-selection bias. This design is most advantageous where the number of items to be tested (treatments) is large and the consideration of cost (in terms of time) forces the experiment to be smaller than is desired (Cochran and Cox, 1957). The multiple test forms were developed using a domain-sampling model where items were randomly assigned within QTaxon to a test form.

*Quantile Framework Field Study – Sample.* The Quantile Framework field study was conducted in February 2004. Thirty-seven schools from 14 districts across six states (California, Indiana, Massachusetts, North Carolina, Utah, and Wisconsin) agreed to participate in the study. Data were received from 34 of the schools (two elementary and one middle-school did not return data). A total of 9,847 students in grades 2 through 12 were tested. The number of students per school ranged from 74 to 920. The schools were diverse in terms of geographic location, size, and type of community (e.g., suburban; small town, city, or rural communities; and urban). *Table 1* provides information about the sample at each grade level and by gender.

Table 1. Field-study participation by grade and gender.

Grade Level	<i>N</i>	Percent Female ( <i>N</i> )	Percent Male ( <i>N</i> )
2	1,283	48.1 (562)	51.9 (606)
3	1,354	51.9 (667)	48.1 (617)
4	1,454	47.7 (644)	52.3 (705)
5	1,344	48.9 (622)	51.1 (650)
6	976	47.7 (423)	52.3 (463)
7	1,250	49.8 (618)	50.2 (622)
8	1,015	51.9 (518)	48.1 (481)
9	489	52.0 (252)	48.0 (233)
10	259	48.6 (125)	51.4 (132)
11	206	49.3 (101)	50.7 (104)
12	143	51.7 (74)	48.3 (69)
Missing	74	39.1 (9)	60.9 (14)
Total	9,847	49.6 (4,615)	50.4 (4,696)

Students given Levels 2 through 11 were provided with rulers and students given Levels 3 through 11 were provided with protractors. For students given taking Levels 5 through 8 and 10 and 11, formulas were provided on the back of the test booklet. Administration time was approximately 45 minutes at each level. Students given Level 2 could have the test read aloud and mark in the test booklet if that was typical of instruction.

Table 2. Test-form administration by level.

Test Level	<i>N</i>	Missing	Form 1	Form 2	Form 3
2	1,283	4	453	430	397
3	1,354	7	561	387	399
4	1,454	17	616	419	402
5	1,344	3	470	448	423
6	917	13	322	293	289
7	1,309	6	463	429	411
8	1,181	16	387	391	387
9	415	4	141	136	134
10	226	5	73	77	71
11	313	10	102	101	100
Missing	51	31	9	8	3
Total	9,847	116	3,596	3,119	3,016

Table 2 shows the number of students by level and form. The final sample included 9,678 students with complete data. Data were deleted if test level or test form was not indicated or the answer sheet was blank.

*Field-Test Analyses.* The field-test data were analyzed using both the classical measurement model and the Rasch (one-parameter logistic item response theory) model. Item statistics and descriptive information (item number, field test form and item number, QTaxon, and answer key) were printed for each item and attached to the item record. The item record contained the statistical, descriptive, and historical information for an item; a copy of the item itself as it was field-tested; any comments by reviewers; and the psychometric notations. Each item had a separate item record.

*Field-Test Analyses – Classical Measurement.* For each item, the  $p$ -value (percent correct) and the point-biserial correlation between the item score (correct response) and the total test score were computed. Point-biserial correlations were also computed between each of the incorrect responses and the total score. In addition, frequency distributions of the response choices (including omits) were tabulated (both actual counts and percents). Items with point-biserial correlations less than 0.10 were removed from the item bank. Table 3 displays the summary item statistics.

Table 3. Summary item statistics from the Quantile Framework field study (February 2004).

Level	Number of Items Tested	Mean P-value (Range)	Mean Correct Response Point-Biserial Correlation (Range)	Mean Incorrect Responses Point-Biserial Correlation (Range)
2	90	0.583 (0.12 – 0.95)	0.322 (-0.15 – 0.56)	-0.209 (-0.43 – 0.12)
3	90	0.532 (0.11 – 0.93)	0.256 (-0.08 – 0.52)	-0.221 (-0.54 – 0.02)
4	90	0.552 (0.12 – 0.92)	0.242 (-0.21 – 0.50)	-0.222 (-0.48 – 0.12)
5	90	0.535 (0.12 – 0.95)	0.279 (-0.05 – 0.50)	-0.225 (-0.45 – 0.05)
6	90	0.515 (0.04 – 0.86)	0.244 (-0.08 – 0.45)	-0.218 (-0.46 – 0.09)
7	90	0.438 (0.10 – 0.77)	0.294 (-0.12 – 0.56)	-0.207 (-0.46 – 0.25)
8	90	0.433 (0.10 – 0.81)	0.257 (-0.15 – 0.50)	-0.201 (-0.45 – 0.13)
9	90	0.396 (0.10 – 0.79)	0.208 (-0.19 – 0.52)	-0.193 (-0.53 – 0.22)
10	88	0.511 (0.01 – 0.97)	0.193 (-0.26 – 0.53)	-0.205 (-0.55 – 0.18)
11	90	0.527 (0.09 – 0.98)	0.255 (-0.09 – 0.51)	-0.223 (-0.52 – 0.07)

*Field-Test Analyses – Bias.* Differential item functioning (DIF) examines the relationship between the score on an item and group membership while controlling for ability. The Mantel-Haenszel procedure has become “the most widely used methodology [to examine differential item functioning] and is recognized as the testing industry standard” (Roussos, Schnipke, and Pashley, 1999, p. 293). The Mantel-Haenszel procedure examines DIF by examining  $j \times 2 \times 2$  contingency tables, where  $j$  is the number of different levels of ability actually achieved by the examinees (actual total scores received on the test). The focal group is the group of interest and the reference group serves as a basis for comparison for the focal group (Dorans and Holland, 1993; Camilli and Shepherd, 1994).

The Mantel-Haenszel chi-square statistic tests the alternative hypothesis that there is a linear association between the row variable (score on the item) and the column variable (group membership). The  $\chi^2$  distribution has 1 degree of freedom and is determined as

$$Q_{MH} = (n - 1)r^2 \quad (\text{Equation 1})$$

where  $r$  is the Pearson correlation between the row variable and the column variable (SAS Institute, 1985).

The Mantel-Haenszel (MH) Log Odds Ratio statistic is used to determine the direction of differential item functioning (SAS Institute Inc., 1985). This measure is obtained by combining the odds ratios,  $\alpha_j$ , across levels with the formula for weighted averages (Camilli and Shepherd, 1994, p. 110):

$$\alpha_j = \frac{p_{Rj} / q_{Rj}}{p_{Fj} / q_{Fj}} = \frac{\Omega_{Rj}}{\Omega_{Fj}} \quad (\text{Equation 2})$$

For this statistic, the null hypothesis of no relationship between score and group membership, or that the odds of getting the item correct are equal for the two groups, is not rejected when the odds ratio equals 1. For odds ratios greater than 1, the interpretation is that an individual at score level  $j$  of the Reference Group has a greater chance of answering the item correctly than an individual at score level  $j$  of the Focal Group. Conversely, for odds ratios less than 1, the interpretation is that an individual at score level  $j$  of the Focal Group has a greater chance of answering the item correctly than an individual at score level  $j$  of the Reference Group. The Breslow-Day Test is used to test whether the odds ratios from the  $j$  levels of the score are all equal. When the null hypothesis is true, the statistic is distributed approximately as a  $\chi^2$  with  $j-1$  degrees of freedom (Camilli and Shepherd, 1994; SAS Institute, 1985).

For the gender analyses, males (approximately 50.4% of the population) were defined as the reference group and females (approximately 49.6% of the population) were defined as the focal group. The results from the Quantile Framework field study were reviewed for inclusion on later linking studies. The following statistics were reviewed for each item:  $p$ -value, point-biserial correlation, and DIF estimates. Items that exhibited extreme statistics were removed from the item bank (47 out of 685).

From the studies conducted with the Quantile Framework item bank (Palm Beach County [FL] linking study, Mississippi linking study, DoDEA/TerraNova linking study, and Wyoming linking study), approximately 6.9% of the items in any one study were flagged as exhibiting DIF using the Mantel-Haenszel statistic and the  $t$ -statistic from Winsteps. For each linking study the following steps were used to review the items: (1) flag items exhibiting DIF, (2) review items to determine if the content of the item is something that all students should know and be able to do, and (3) make decision to retain or delete the item.

*Field-Test Analyses – Rasch Item Response Theory.* Classical test theory has two basic shortcomings: (1) the use of item indices whose values depend on the particular group of examinees from which they were obtained, and (2) the use of examinee ability estimates that depend on the particular choice of items selected for a test. The basic premises of item response theory (IRT) overcome these shortcomings by predicting the performance of an examinee on a test item based on a set of underlying abilities (Hambleton and Swaminathan, 1985). The relationship between an examinee's item performance and the set of traits underlying item performance can be described by a monotonically increasing function called an item characteristic curve (ICC). This function specifies that as the level of the trait increases, the probability of a correct response to an item increases.

The conversion of observations into measures can be accomplished using the Rasch (1980) model, which states a requirement for the way that item calibrations and observations (count of correct items) interact in a probability model to produce measures. The Rasch IRT model expresses the probability that a person ( $n$ ) answers a certain item ( $i$ ) correctly by the following relationship:

$$P_{ni} = \frac{e^{b_n - d_i}}{1 + e^{b_n - d_i}} \quad (\text{Equation 3})$$

where  $d_i$  is the difficulty of item  $i$  ( $i = 1, 2, \dots$ , number of items);

$b_n$  is the ability of person  $n$  ( $n = 1, 2, \dots$ , number of persons);

$b_n - d_i$  is the difference between the ability of person  $n$  and the difficulty of item  $i$ ;

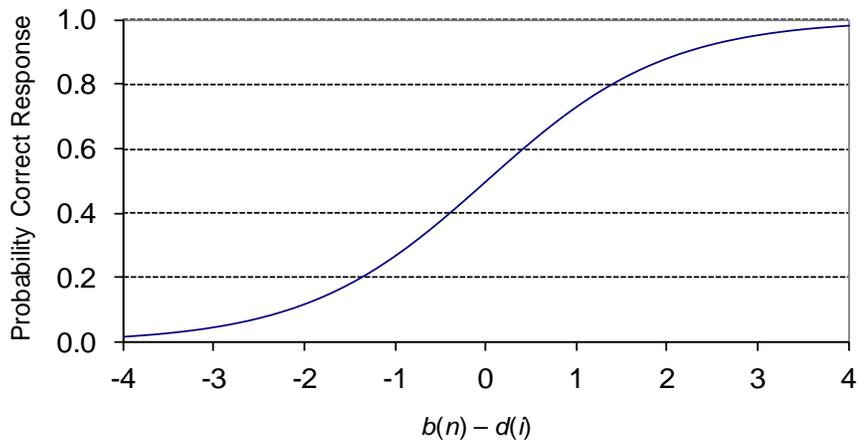
and

$P_{ni}$  is the probability that examinee  $n$  responds correctly to item  $i$

(Hambleton and Swaminathan, 1985; Wright and Linacre, 1994).

This measurement model assumes that item difficulty is the only item characteristic that influences the examinee's performance such that all items are equally discriminating in their ability to identify low-achieving persons and high achieving persons (Bond and Fox, 2001; and Hambleton, Swaminathan, and Rogers, 1991). In addition, the lower asymptote is zero, which specifies that examinees of very low ability have zero probability of correctly answering the item. The Rasch model has the following assumptions: (1) unidimensionality – only one ability is assessed by the set of items; and (2) local independence – when abilities influencing test performance are held constant, an examinee's responses to any pair of items are statistically independent (conditional independence, i.e., the only reason an examinee scores similarly on several items is because of his or her ability, not because the items are correlated). The Rasch model is based on fairly restrictive assumptions, but it is appropriate for criterion-referenced assessments. *Figure 1* graphically shows the probability that a person will respond correctly to an item as a function of the difference between a person's ability and an item's difficulty.

Figure 1. The Rasch Model – the probability person  $n$  responds correctly to item  $i$ .



An assumption of the Rasch model is that the probability of a response to an item is governed by the difference between the item calibration ( $d_i$ ) and the person's measure ( $b_n$ ). From an examination of the graph in *Figure 1*, when the ability of the person matches the difficulty of the item ( $b_n - d_i = 0$ ), then the person has a 50% probability of responding to the item correctly.

The number of correct responses for a person is the probability of a correct response summed over the number of items. When the measure of a person greatly exceeds the calibration (difficulties) of the items ( $b_n - d_i > 0$ ), then the expected probabilities will be high and the sum of these probabilities will yield an expectation of a high "number correct." Conversely, when the item calibrations generally exceed the person measure ( $b_n - d_i < 0$ ), the modeled probabilities of a correct response will be low and the expectation will be a low "number correct."

Thus, Equation 3 can be rewritten in terms of the number of correct responses of a person on a test

$$O_p = \sum_{i=1}^L \frac{e^{b_n - d_i}}{1 + e^{b_n - d_i}} \quad (\text{Equation 4})$$

where  $O_p$  is the number of correct responses of person  $p$  and  $L$  is the number of items on the test.

When the sum of the correct responses and the item calibrations ( $d_i$ ) is known, an iterative procedure can be used to find the person measure ( $b_n$ ) that will make the sum of the modeled probabilities most similar to the number of correct responses. One of the

key features of the Rasch IRT model is its ability to place both persons and items on the same scale. It is possible to predict the odds of two individuals being successful on an item based on knowledge of the relationship between the abilities of the two individuals. If one person has an ability measure that is twice as high as that of another person (as measured by  $b$  – the ability scale), then he or she has twice the odds of successfully answering the item.

Equation 4 possesses several distinguishing characteristics:

- The key terms from the definition of measurement are placed in a precise relationship to one another.
- The individual responses of a person to each item on an instrument are absent from the equation. The only information that appears is the “count correct” ( $O_p$ ), thus confirming that the raw score (i.e., number of correct responses) is “sufficient” for estimating the measure.

For any set of items the possible raw scores are known. When it is possible to know the item calibrations (either theoretically or empirically from field studies), the only parameter that must be estimated in Equation 4 is the person measure that corresponds to each observable count correct. Thus, when the calibrations ( $d_i$ ) are known, a correspondence table linking observation and measure can be constructed without reference to data on other individuals.

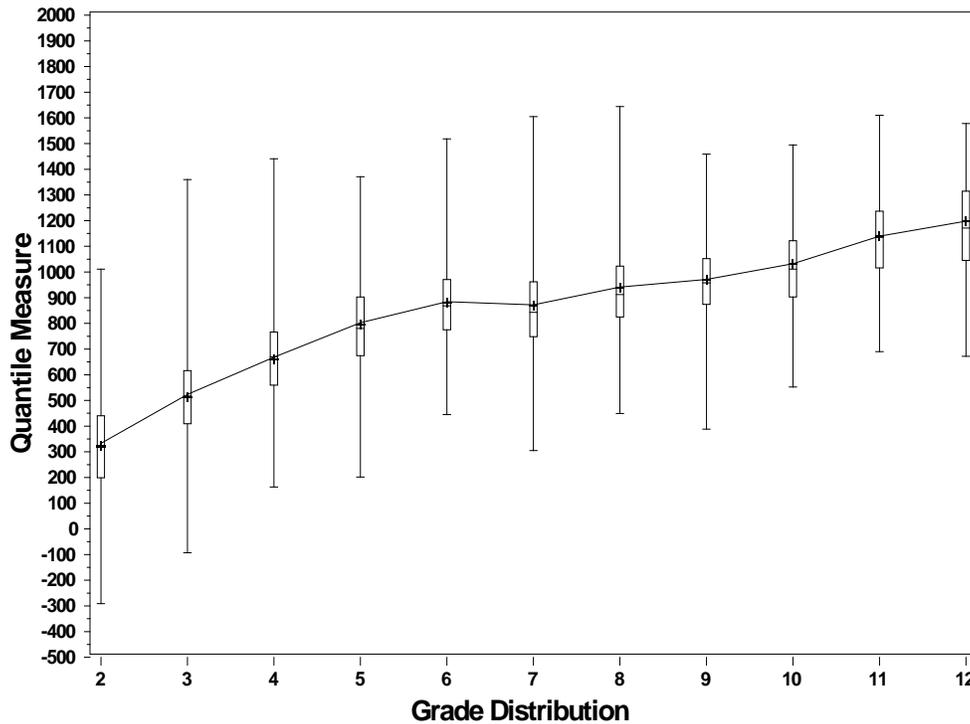
All students and items were submitted to a Winsteps analysis using a logit convergence criterion of 0.0001 and a residual convergence criterion of 0.001. Items that a student skipped were treated as missing, rather than being treated as incorrect. Only students who responded to at least 20 items were included in the analyses (22 students were omitted, 0.22%). *Table 4* shows the mean and median Quantile measures for all students with complete data at each grade level. While there is not a monotonically increasing trend in the mean and median Quantile measures in Grades 6 and 7, the measures are not significantly different. Results from other studies (e.g., *PA Series Math* described beginning on page 25 exhibit a monotonically increasing function).

Table 4. Mean and median Quantile measures for students with complete data (N = 9,656).

Grade Level	N	Mean Quantile measure (SD)	Median Quantile measure
2	1,275	320.68 (189.11)	323
3	1,339	511.41 (157.69)	516
4	1,427	655.45 (157.50)	667
5	1,337	790.06 (167.71)	771
6	959	871.82 (153.02)	865
7	1,244	860.52 (174.16)	841
8	1,004	929.01 (157.63)	910
9	482	958.69 (152.81)	953
10	251	1019.97 (162.87)	1005
11	200	1127.34 (178.57)	1131
12	138	1185.90 (189.19)	1164

Figure 2 shows the relationship between grade level and Quantile measure. The following box and whisker plots (Figures 2, 3, and 4) show the progression of the y-axis scores from grade to grade (the x-axis). For each grade, the box refers to the inter-quartile range. The line within the box indicates the median and the + indicates the mean. The end of each whisker shows the minimum and maximum values of the y-axis which is the Quantile measure. Across all students, the correlation between grade and Quantile measure was 0.76.

Figure 2. Box and whisker plot of the Rasch ability estimates of all students with complete data ( $N = 9,656$ ).



All students with outfit mean square statistics greater than or equal to 1.8 were removed from further analyses. A total of 480 students (4.97%) were removed from further analyses. The number of students removed ranged from 8.47% (108) in grade 2 to 2.29% (22) in grade 6 with a mean percent decrease of 4.45% per grade.

All remaining students (9,176) and all items were submitted to a Winsteps analysis using a logit convergence criterion of 0.0001 and a residual convergence criterion of 0.001. Items that a student skipped were treated as missing, rather than being treated as incorrect. Only students who responded to at least 20 items were included in the analyses. *Table 5* shows the mean and median Quantile measures for the final set of students at each grade level. *Figure 3* shows the results from the final set of students. The correlation between grade level and Quantile measure was 0.78.

Table 5. Mean and median Quantile measures for the final set of students ( $N = 9,176$ ).

Grade Level	N	Median Logit Value	Mean Quantile measure (Median)
2	1,167	-2.800	289.03 (292)
3	1,260	-1.650	502.18 (499)
4	1,352	-0.780	652.60 (656)
5	1,289	0.000	795.25 (796)
6	937	0.430	880.77 (874)
7	1,181	0.370	877.75 (863)
8	955	0.810	951.41 (942)
9	466	1.020	982.62 (980)
10	244	1.400	1044.08 (1048)
11	191	2.070	1160.49 (1169)
12	134	2.295	1219.87 (1210)

Figure 3. Box and whisker plot of the Rasch ability estimates for the final sample of students with outfit statistics less than 1.8 ( $N = 9,176$ ).

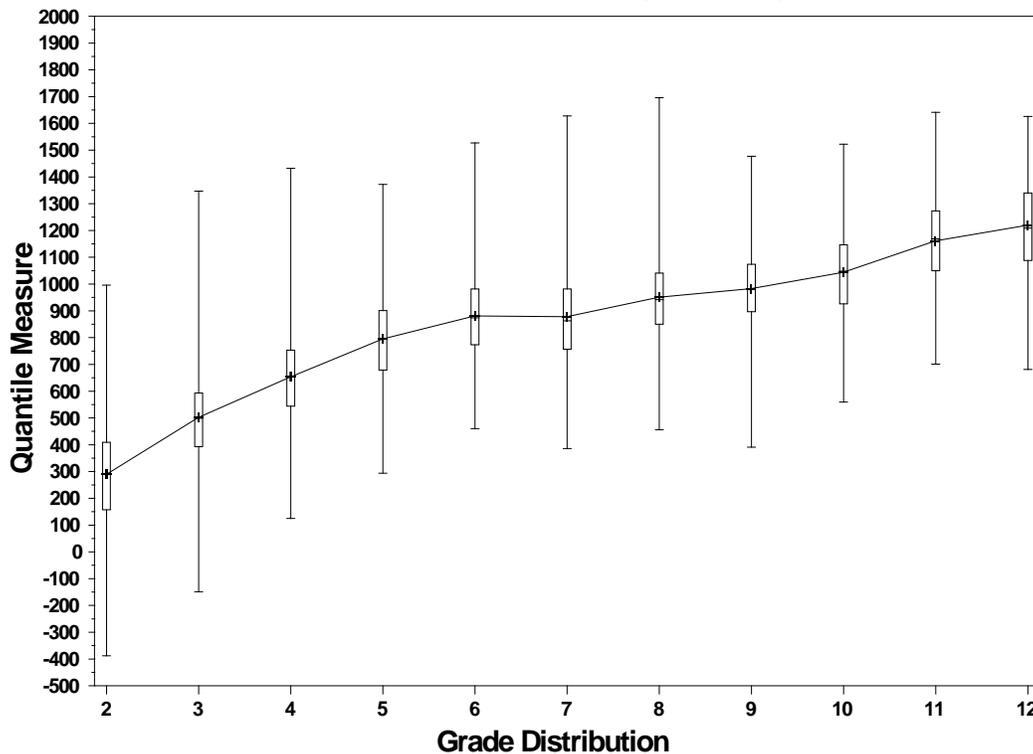
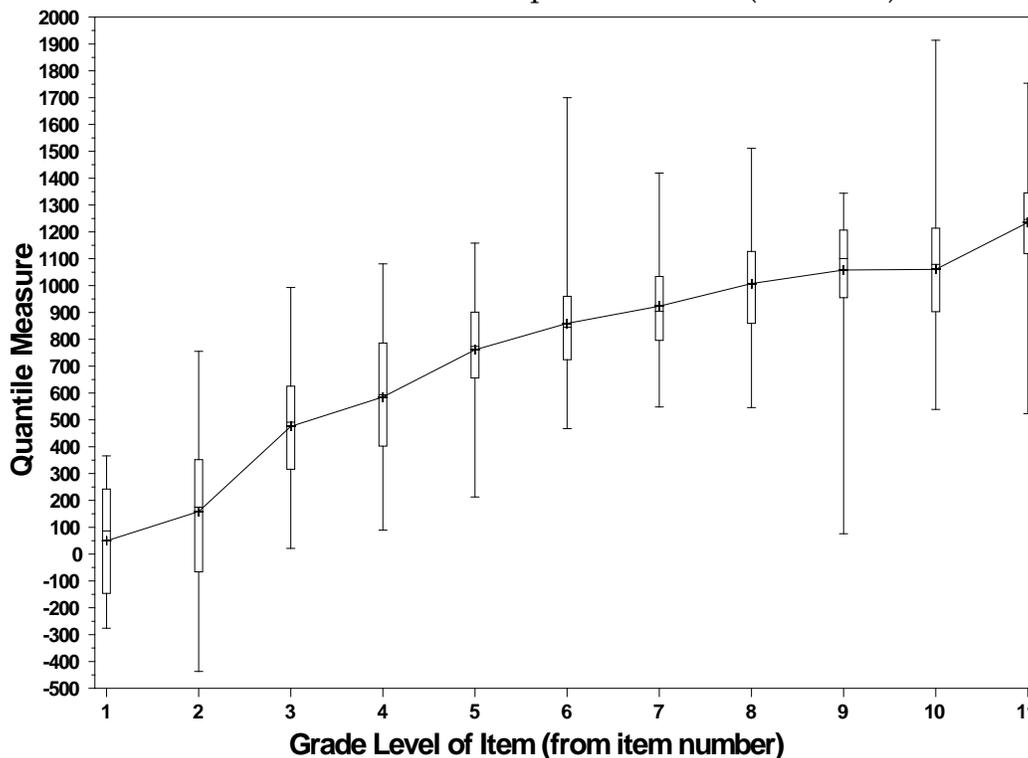


Figure 4 shows the distribution of item difficulties based on the final sample of students. For this analysis, missing data were treated as “skipped” items and not counted as wrong. There is a gradual increase in difficulty when items are sorted by level of test for which the items were written. This distribution appears to be non-linear, which is

consistent with other studies. The correlation between the grade level for which the item was written and the Quantile measure of the item was 0.80.

Figure 4. Box and whisker plot of the Rasch difficulty estimates of the 685 Quantile Framework items for the final sample of students ( $N = 9,176$ ).



### Calibration of Items on the Quantile Scale

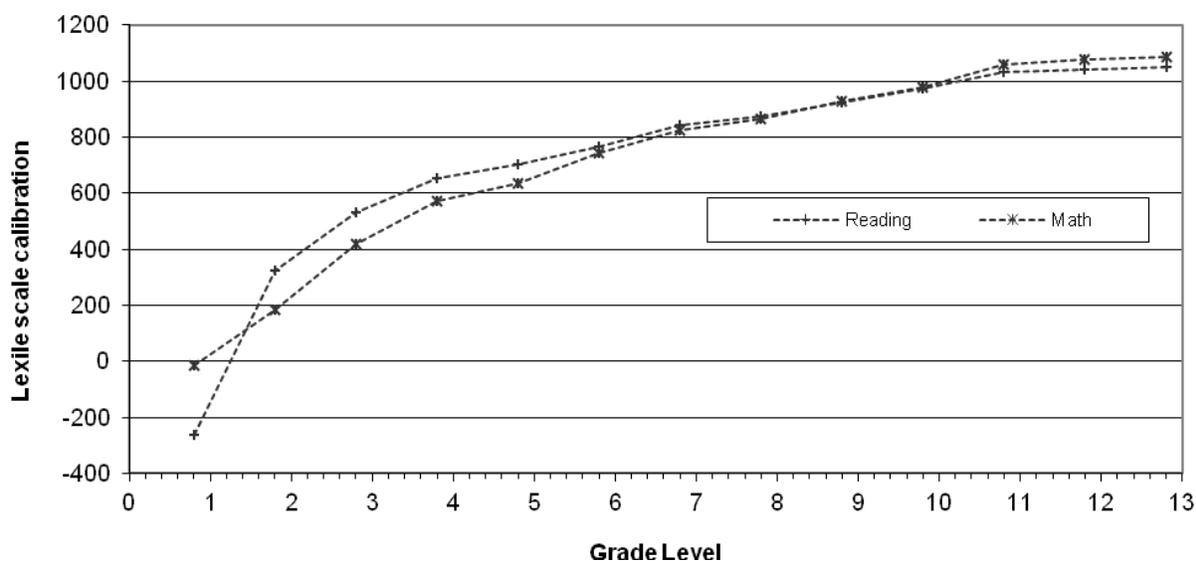
In developing the Quantile scale, two features of the scale were needed: (1) scale multiplier (conversion factor) and (2) anchor point. The Rasch item response theory model (Wright and Stone, 1979) was used to estimate the difficulties of items and the abilities of persons on the logit scale.

The calibrations of the items from the Rasch model are objective in the sense that the relative difficulties of the items will remain the same across different samples of persons (specific objectivity). When two items are administered to the same person it can be determined which item is harder and which one is easier. This ordering should hold when the same two items are administered to a second person. If two different items are administered to the second person, there is no way to know which set of items is harder and which set is easier. The problem is that the location of the scale is not known. General objectivity requires that scores obtained from different test administrations be tied to a common zero – absolute location must be sample independent (Stenner, 1990).

To achieve general objectivity, the theoretical logit difficulties must be transformed to a scale where the ambiguity regarding the location of zero is resolved.

The first step in developing the Quantile scale was to determine the conversion factor (CF) to be used to go from logits to Quantile measure. Based on prior research with reading and the Lexile scale, the decision was made to examine the relationship between reading and mathematics scales used with other assessments. The median scale score for each grade level on a norm-referenced assessment linked with the Lexile scale is plotted in *Figure 5* using the same conversion equation for both reading and mathematics.

*Figure 5.* Relationship between reading and mathematics scale scores on a norm-referenced assessment linked to the Lexile scale in reading.

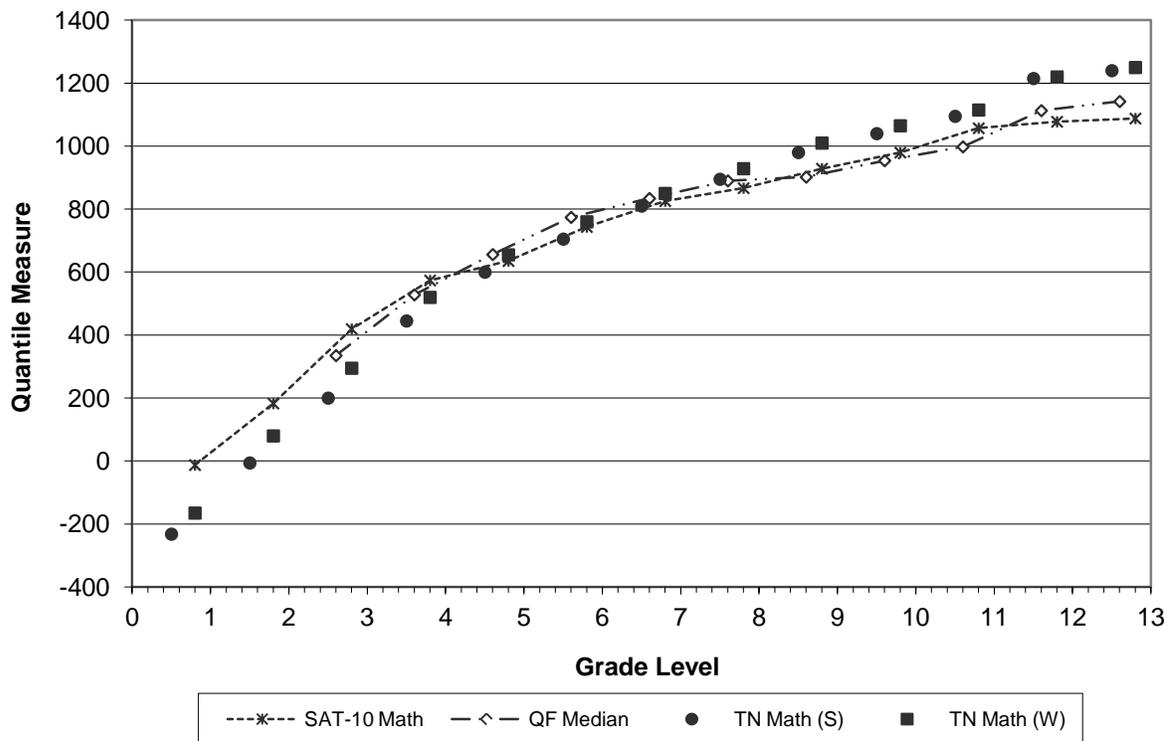


Based on an examination of *Figure 5*, it was concluded that the same conversion factor of 180 that is used with the Lexile scale could be used with the Quantile scale. Both sets of data exhibited a similar pattern across grades.

The second step in developing the Quantile scale with a fixed zero was to identify an anchor point for the scale. Given the number of students at each grade level in the field study, it was concluded that the scale should be anchored at grade 4 or 5 (middle of grade span typically tested by state assessment programs). Median performance at the end of grade 3 on the Lexile scale is 590L. The Quantile Framework field study was conducted in February and this point would correspond to six months (0.6) through the

school year. Median performance at the end of grade 4 on the Lexile scale is 700L. To determine the location of the scale, 66Q were added to the median performance at the end of grade 3 to reflect the growth of students in grade 4 prior to the field study ( $700 - 590 = 110$ ;  $110 \times 0.6 = 66$ ). The value of 656Q was used for the location of grade 4 median performance. The anchor point was validated with other assessment data and collateral data from the Quantile Framework field study (see *Figure 6*).

*Figure 6.* Relationship between grade level and mathematics performance on the Quantile Framework field study and other mathematics assessments.



Finally, a linear equation of the form

$$[(\text{Logit} - \text{Anchor Logit}) \times \text{CF}] + 656 = \text{Quantile measure} \quad (\text{Equation 5})$$

was developed to convert logit difficulties to Quantile calibrations where the anchor logit is the median for grade 4 in the Quantile Framework field study.

## QTaxon Quantile Measures

In order to use the Quantile Framework to examine the difficulty of skills and concepts and the complexity of resources, the Quantile measure of each QTaxon must be estimated. The Quantile measure of a QTaxon estimates its solvability, or a prediction of how difficult the skill or concept will be for the learner with a Quantile measure of his or her own. The QTaxons fall into knowledge clusters along a content continuum.

The Quantile measures and knowledge clusters for QTaxons are determined by a group of three to five subject-matter experts (SMEs). Each SME has had classroom experience at multiple developmental levels, has completed graduate-level courses in mathematics education, and understands basic psychometric concepts and assessment issues.

*Knowledge Clusters.* Knowledge clusters are a family of skills, like building blocks, that depend one upon the other to connect and demonstrate how skills are founded, supported, and extended along the continuum. The knowledge clusters illustrate the interconnectivity of the Quantile Framework and the natural progression of mathematical skills (content progressions) needed to solve increasingly complex problems.

Each QTaxon was classified as having “prerequisite” and “supplemental” QTaxons or as being a “foundational” QTaxon by the SMEs. A *prerequisite QTaxon* is a QTaxon that describes a skill or concept that provides the foundation necessary for another QTaxon. For example, adding single-digit numbers is a prerequisite for adding two-digit numbers. A *supplemental QTaxon* is a QTaxon which describes supplementary skills or knowledge that assists and enriches the understanding of another QTaxon. An *impending QTaxon* describes the skills and concepts that will be built from a primary QTaxon and helps the teacher or parent to see a trajectory of knowledge across grades and content strands. The SMEs examined each QTaxon to determine where the specific QTaxon comes in the content progression based on classroom experience, instructional resources (e.g., textbooks), and other curricular frameworks (e.g., NCTM Standards). A QTaxon that is classified as “foundational” means this QTaxon describes a skill or concept that only requires readiness to learn. Readiness is based upon the learner’s cognitive experiences rather than knowledge of specific mathematical concepts. It is the basis upon which other QTaxons are built.

Once the knowledge cluster for a QTaxon was established, the information was used when determining the Quantile measure of a QTaxon (described below). If necessary, knowledge clusters were reviewed and refined if the Quantile measures of the QTaxons in the cluster were not monotonically increasing or there was not an instructional explanation for the pattern.

*Quantile measures of QTaxons.* To determine the Quantile measure of a QTaxon, actual performance by examinees was used. While expert judgment alone could have been used to scale the QTaxons, empirical scaling was more replicable. Items and resulting data from two national field studies were used in the process:

- Quantile Framework field study (685 items,  $N = 9,647$ , grades 2 through Algebra II) which is described earlier in this section; and
- *PASeries* Mathematics field study (7,080 items,  $N = 27,329$ , grades 2 through 9/Algebra I) which is described in the *PASeries* Mathematics Technical Manual (MetaMetrics, 2005).

The items initially associated with each QTaxon were reviewed by SMEs and accepted for inclusion in the set of items, moved to another QTaxon, or not included in the set. The following criteria were used:

- Psychometric (responded to by at least 50 examinees, administered at the target grade level, point-biserial correlation greater than or equal to 0.16);
- Matched grade level of introduction of concept/skill from national review of curricular frameworks (described on pages 3 and 4); and,
- Appropriate for instruction of concept (first night's homework; from the A and B sections of the lesson problems) based on consensus of the SMEs.

Once the set of items meeting the inclusion criteria was identified, the set of items was reviewed to ensure that the curricular breadth of the QTaxon was covered. If the group of SMEs considered the set of items to be acceptable, then the Quantile measure of the QTaxon was calculated. The Quantile measure of a QTaxon is defined as the mean Quantile measure of items that met the criteria. The standard deviation of the item difficulties was also calculated (mean standard deviation of item difficulties across QTaxons was 177.3Q). The final step in the process was to review the Quantile measure of the QTaxon in relationship to the Quantile measures of the QTaxons identified as prerequisite and supplementary to the QTaxon. If the group of SMEs did not consider the set of items to be acceptable, then the Quantile measure of the QTaxon was estimated and assigned a Quantile zone. By assigning a Quantile zone instead of a Quantile measure to a QTaxon, the SMEs were able to provide a valid estimate of the skill or concept's difficulty.

QTaxon Quantile measures are used in the calibration of resources (e.g., textbooks, instructional materials, supplemental materials, workplace documents, everyday documents) used with the Quantile Framework.

## Validation of The Quantile Framework for Mathematics

Validity is the extent to which a test measures what its authors or users claim it measures; specifically, test validity concerns the appropriateness of inferences “that can be made on the basis of observations or test results” (Salvia and Ysseldyke, 1998, p. 166). The 1999 *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education) state that “validity refers to the degree to which evidence and theory support the interpretations of test scores entailed in the uses of tests” (p. 9). For the Quantile Framework, which measures student understanding of mathematical skills and concepts, the most important aspect of validity that should be examined is construct validity. The construct validity of The Quantile Framework for Mathematics can be evaluated by examining how well Quantile measures relate to other measures of mathematics described in the following sections.

*Standardization set of items used with PASeries Mathematics.* PAseries Mathematics is a series of classroom-based, progress monitoring assessments designed for use in the US school market in grades 3 through 8 (MetaMetrics, 2005). Each PAseries Mathematics assessment measures a range of mathematics skills appropriate to a specific grade. For each grade, PAseries Mathematics includes a screener test (pre-test) and six progress assessments designed to be administered approximately every six weeks. Each assessment contains 30 items; an assessment can be administered in one typical class period. As the school year progresses, each assessment is designed to be at a higher Quantile level, resulting in progressively more challenging tests.

For the standardization set, the items in the Quantile Framework field study that were also in the PAseries Mathematics field study were examined. Only items that were presented in exactly the same form in both studies were retained. A total of 213 items were identified that were administered in both studies. One item was dropped because none of the responses were correct, five items were dropped because they were too easy, and five items were dropped because there were presentation differences between the studies. The final number of items in the standardization set was 207. The test level breakdown is presented in *Table 6*.

Table 6. Number of items in the Quantile Framework standardization set by grade level of the item content.

Content Level of Items (by Grade)	Number of Items in Standardization Set
1	6
2	32
3	25
4	29
5	27
6	26
7	27
8	19
9	15
10	1

The relationship between the calibrations of the standardization set of items used in the Quantile Framework field study and on *PA Series* Mathematics (the calibration of the *PA Series* Mathematics items will be described later in this technical manual) was examined. The correlation of the Quantile measures of the 207 items was 0.92. The mean difference was -186Q and the standard deviation of the differences was 153Q. The standardization set of items is validated by consistency of measures between the two studies. Characteristics of the items in the standardization set from the two field studies are presented in *Figures 7* and *8*.

Figure 7. Comparison of the difficulty (Quantile measure) of the standardization set of items across two field studies.

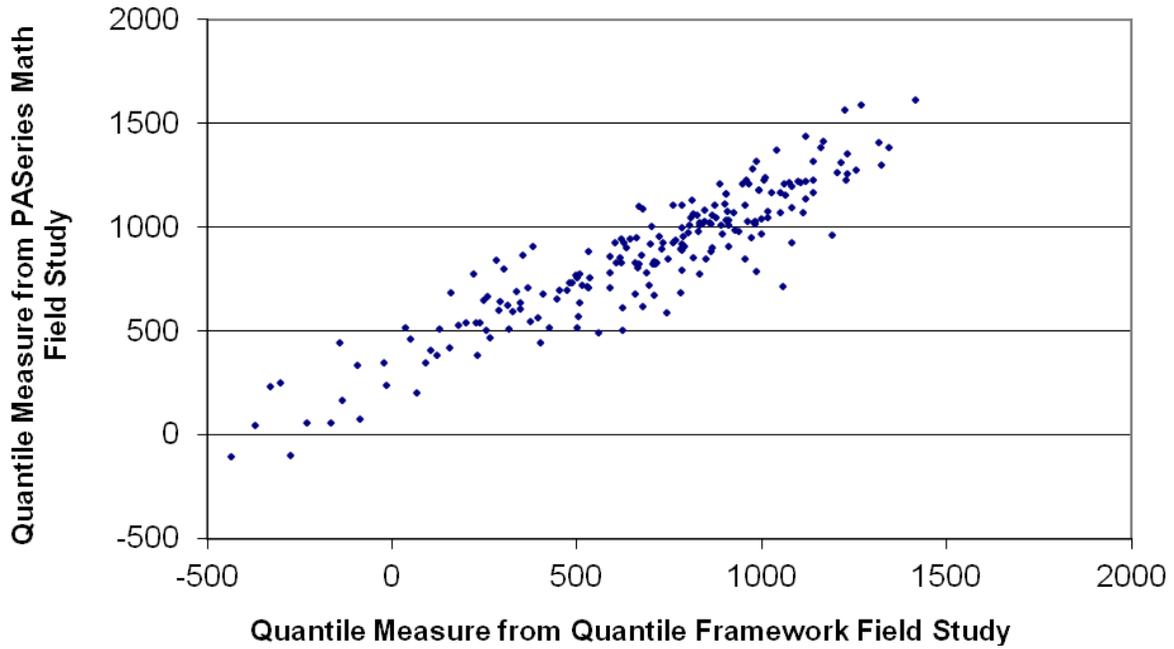
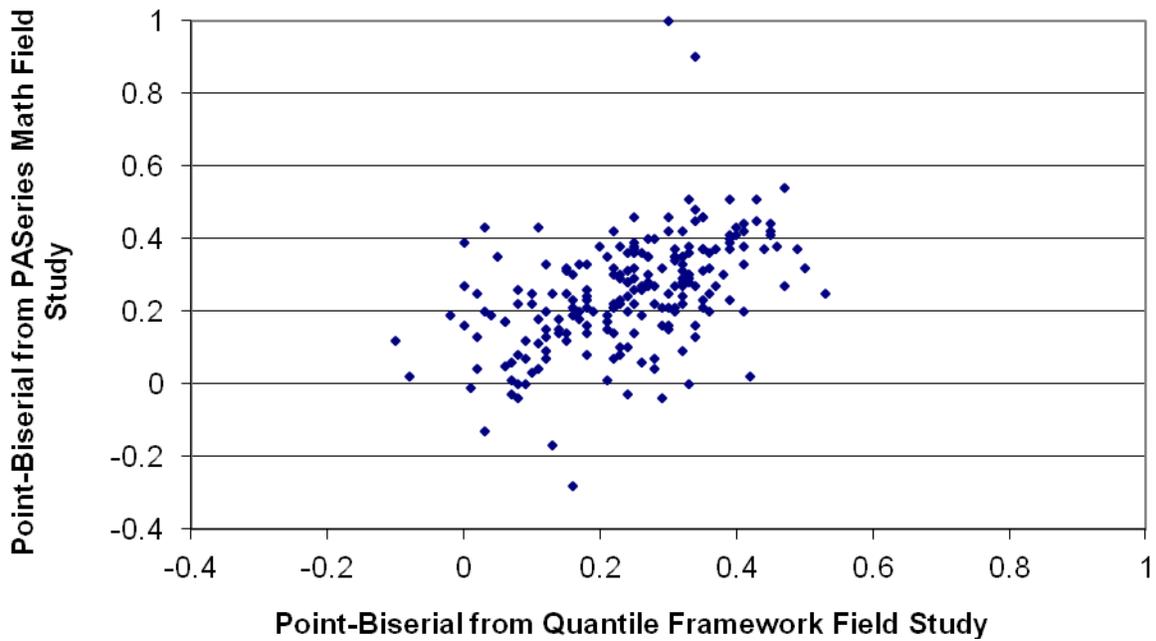


Figure 8. Comparison of the point-biserial correlations of the standardization set of items across two field studies.



The *PA Series* Math field study included 23,987 students who provided their grade level. *Table 7* shows the descriptive statistics for the sample by grade level. A monotonically increasing Quantile measure is observed across the grade levels.

*Table 7.* Mean and median Quantile measures for students with complete data from the *PA Series* Math field study ( $N = 23,987$ ).

Grade Level	N	Mean Quantile measure	Median Quantile measure
3	4,703	370.46	370
4	4,478	592.29	598
5	3,871	696.54	690
6	2,813	788.32	771
7	3,555	827.24	816
8	3,481	884.81	874
9	1,086	970.24	967

*Relationship of Quantile Measures to other Measures of Mathematical Ability.* Scores from tests purporting to measure the same construct, for example “mathematical ability,” should be moderately correlated (Anastasi, 1982). *Table 8* presents the results from field studies conducted with The Quantile Framework for Mathematics. For each of the tests listed, student mathematics scores were correlated with Quantile measures from the Quantile Framework field study.

*Table 8.* Results from studies conducted with The Quantile Framework for Mathematics.

Standardized Test	Grades in Study	$N$	Correlation Between Test Score and Quantile measure
RIT and Measures of Academic Progress (MAP by NWEA)	4 & 5	94	0.69
North Carolina End-of-Grade Tests (Mathematics)	4 & 5	341	0.73

*Quantile Framework Linked to other Measures of Mathematics Understanding.* The Quantile Framework for Mathematics has been linked to several standardized tests of mathematics achievement. When assessment scales are linked, a common frame of

reference can be used to interpret the test results. This frame of reference can be “used to convey additional normative information, test-content information, and information that is jointly normative and content-based. For many test uses ... [this frame of reference] conveys information that is more crucial than the information conveyed by the primary score scale” (Petersen, Kolen, and Hoover, 1989, p. 222).

*Table 9* presents the results from linking studies conducted with the Quantile Framework. For each of the tests listed, student mathematics scores can also be reported as Quantile measures. This dual reporting provides a rich, criterion-related frame of reference for interpreting the standardized test scores. When a student takes one of the standardized tests, in addition to receiving her or his norm-referenced test results, s/he can receive information related to the specific QTaxons that s/he is ready to receive instruction.

Table 9. Results from linking studies conducted with the Quantile Framework.

Standardized Test	Grades in Study	<i>N</i>	Correlation Between Test Score and Quantile measure
Mississippi Curriculum Test, Mathematics (MCT)	2 – 8	7,039	0.89
TerraNova (CTB/McGraw-Hill)	3, 5, 7, 9	6,356	0.92
Texas Assessment of Knowledge and Skills (TAKS)	3 – 11	14,286	0.69 to 0.78*
Proficiency Assessments for Wyoming Students (PAWS)	3, 5, 8, and 11	3,923	0.87
Progress Towards Standards (PTS3)	3-8 and 10	8,544	0.86 to 0.90*
Progress in Maths (PiM – GL Assessments)	1 – 8	3,183	0.71 to 0.81*
North Carolina End-of-Grade/End-of-Course Tests (NC EOG/NC EOC)	3, 5, 7, A1, G, and A2	5,069	0.88 to 0.90*
Kentucky Core Content Tests (KCCT)	3 - 8 and 11	12,660	0.80 to 0.83*
Oklahoma Core Competency Tests (OCCT)	3 – 8	5,649	0.81 to 0.85*
Iowa Assessments	2, 4, 6, 8, and 10	7,365	0.92
Virginia Standards of Learning (SOL)	3-8, A1, G, and A2	12,470	0.86 to 0.89*

Notes: \* TAKS, PTS3, PiM, NCEOC, KCCT, OCCT, and SOL were not vertically scaled; separate linking equations were derived for each grade/course.

*Multidimensionality of the Quantile Framework.* Test dimensionality is defined as the minimum number of abilities or constructs measured by a set of test items. A construct is a theoretical representation of an underlying trait, concept, attribute, process, and/or structure that a test purports to measure (Messick, 1993). A test can be considered to measure one latent trait, construct, or ability (in which case it is called unidimensional); or a combination of abilities (in which case it is referred to as multidimensional). The

dimensional structure of a test is intricately tied to the purpose and definition of the construct to be measured. It is also an important factor in many of the model(s) used in data analyses. Though many of the models assume unidimensionality, this assumption cannot be strictly met because there are always other cognitive, personality, and test-taking factors that have some level of impact on test performance (Hambleton and Swaminathan, 1985).

**Study 1 - Comparison of Mathematics with Reading.** The multidimensionality of the Quantile scale was examined using the Principal Components Analysis of Residuals in Winsteps (PRCOMP=S). The items were renamed with the strand number first for ease in review of the output. A three-step process was undertaken in order to examine the results and provide a context for interpreting the results.

The first step in the process was to run the Principal Components Analysis on all Quantile Framework field study items ( $N = 898$ ). Next, the residual matrix was factor analyzed. *Table 10* shows the output from the analysis. The variance that is unexplained by the first factor (the Rasch measurement model) is 0.2% of the residual variance or 2.5 items of information. Based upon this set of data, it cannot be concluded that mathematics achievement as measured by the Quantile scale is multidimensional. The results supported the use of a unidimensional item response model on the items.

*Table 10.* Principal components analysis and distribution of variance explained by the model with the Quantile Framework field-study mathematics items ( $N = 685$ ).

Source	Standardized Residual Variance (in Eigenvalue units)	Empirical	Modeled
Total Variance in Observations	1327.4	100.0%	100.0%
Variance Explained by Measures	642.4	48.4%	49.9%
Unexplained Variance (Total)	685.0	51.6%	50.1%
Unexplained Variance Explained by 1 <sup>st</sup> Factor of the Residual Matrix	2.5	0.2%	

Next, the items were ordered by factor loading. Based on an examination of the item names with strand listed first, there did not appear to be any effect of strand. Only 6 items out of the 685 unique items had loadings above 0.30 on the first residual factor.

These six items were all level 10 (Geometry) items and were from both strands 2 (Geometry) and 3 (Algebra).

To better understand the values produced in the first analysis, a second analysis was undertaken. The Level 5 (Grade 5) Quantile items were analyzed separately. The results are presented in *Table 11*.

*Table 11.* Principal components analysis and distribution of variance explained by the model with the Grade 5 Quantile Framework field-study mathematics items ( $N = 65$ ).

Source	Standardized Residual Variance (in Eigenvalue units)	Empirical	Modeled
Total Variance in Observations	118.1	100.0%	100.0%
Variance Explained by Measures	53.1	45.0%	45.9%
Unexplained Variance (Total)	65.0	55.0%	54.1%
Unexplained Variance Explained by 1 <sup>st</sup> Factor of the Residual Matrix	1.8	1.5%	

Three examples in the research literature describe the investigation of reading as a unidimensional construct: the 1940s Davis Study (Davis, 1944; Thurstone, 1946), the 1970s Anchor Study (Rentz and Bashaw, 1975, 1977; Jaeger, 1973; Loret, Seder, Bianchini, and Vale, 1974), and five 1980s and 1990s studies examining research conducted by ETS (Kirsch & Jungeblut and their colleagues, 1993, 1994; Reder, 1996; Salganik & Tal, 1989; Zwick, 1987). Other more recent examples include Harvey Goldstein’s research with PISA (November 17, 2003), research on the development of the North Carolina End-of-Grade Tests (NCDPI, 1996), and research with the 2003 Maryland School Assessment – Reading. All of the studies confirm the assumption of unidimensionality of the reading assessments. Since most research concludes that reading is a unidimensional construct, for comparison purposes, a set of reading grade 5 reading items was also analyzed. The results are presented in *Table 12*.

Table 12. Principal components analysis and distribution of variance explained by the model with Grade 5 reading comprehension items ( $N = 54$ ).

Source	Standardized Residual Variance (in Eigenvalue units)	Empirical	Modeled
Total Variance in Observations	137.1	100.0%	100.0%
Variance Explained by Measures	83.1	60.6%	62.1%
Unexplained Variance (Total)	54.0	39.4%	37.9%
Unexplained Variance Explained by 1 <sup>st</sup> Factor of the Residual Matrix	2.0	1.5%	

The Rasch model explains 60.6% of the variance in the reading comprehension items. Along with the results presented in *Tables 11* and *12*, these data are consistent with the use of a unidimensional item response theory model for each of the analyses (reading and mathematics).

Finally, items from strands 2 (geometry) and 3 (algebra) were analyzed. It was hypothesized, that if multi-dimensionality were to be evidenced in the data, this would be the most likely contrast. The Winsteps analysis using all 296 of the strand 2 and 3 items in all of the forms did not appear to have any connectivity (common item) problems.

Table 13. Principal components analysis and distribution of variance explained by the model with the Strand 2 and 3 Quantile Framework field-study mathematics items ( $N = 296$ ).

Source	Standardized Residual Variance (in Eigenvalue units)	Empirical	Modeled
Total Variance in Observations	644.7	100.0%	100.0%
Variance Explained by Measures	348.7	54.1%	55.5%
Unexplained Variance (Total)	296.0	45.9%	44.5%
Unexplained Variance Explained by 1 <sup>st</sup> Factor of the Residual Matrix	2.3	0.4%	

Given the larger number of items in the analyses (296 in *Table 13* compared to 65 when only the Grade 5 items were examined in *Table 11*), the Rasch model explains 54.1% of the variance in the geometry (strand 2) and algebra (strand 3) items. The results presented in *Tables 10* and *11* are consistent with the interpretation of a single construct for each of the analyses (reading and mathematics).

**Study 2 – Burg 2007.** A study conducted by Burg (2007) analyzed the dimensional structure of mathematical achievement tests aligned to the NCTM content standards. Since there is no consensus within the measurement community on a single method to determine dimensionality, Burg employed four different methods for assessing dimensionality: (1) exploring the conditional covariances (DETECT), (2) assessment of essential unidimensionality (DIMTEST), (3) item factor analysis (NOHARM), and (4) principal component analysis (WINSTEPS). All four approaches have been shown to be effective indices of dimensional structure. Burg analyzed Grades 3 through 8 data from the Quantile Framework field study previously described.

Each set of on-grade items for a test form from Grades 3 through 8 were analyzed for possible sources of dimensionality related to the five mathematical content strands. The analyses were also used to compare test structures across grades. The results indicated that although mathematical achievement tests for Grades 3 through 8 are complex and exhibit some multidimensionality, the sources of dimensionality are not related to the content strands. The complexity of the data structure, along with the known overlap of mathematical skills, suggests that mathematical achievement tests could represent a fundamentally unidimensional construct. Therefore, while these sub-domains of mathematics are useful for organizing instruction, developing curricular materials such

as textbooks, and describing the organization of items on assessments, they do not describe a significant psychometric property of the test or impact the interpretation of the test results. Mathematics, as measured by the Quantile Framework, can be described as one construct with various sub-domains.

Furthermore, these findings support the NCTM Connections Standard, which states that all students (prekindergarten through Grade 12) should be able to make and use connections among mathematical ideas and see how the mathematical ideas interconnect. Mathematics can be best described as an interconnection of overlapping skills with a high degree of correlation across the mathematical topics, skills, and strands.

# The Kentucky Performance Rating for Educational Progress - Quantile Framework Linking Process

## Description of the Assessments

*Kentucky Performance Rating for Educational Progress (K-PREP)*. Senate Bill 1 (SB 1), enacted in the 2009 Kentucky General Assembly, required a new public school assessment program beginning in the 2011-12 school year. These assessments are collectively named the Kentucky Performance Rating for Educational Progress (K-PREP) tests. The Kentucky School Testing System is designed to improve teaching and student learning in Kentucky.

The K-PREP Math Tests are designed to use the Kentucky Core Academic Standards (KCAS) as the foundation which are the Common Core State Standards for Mathematics. The K-PREP Math Tests for Grades 3 through 8 are a blended model built with norm-referenced test (NRT) and criterion-referenced test (CRT) items which consist of multiple-choice (MC), extended-response (ER), and short-answer (SA) items. The NRT portion is a purchased test with national norms and the CRT portion is customized for Kentucky. The test is administered in three parts, each given in a separate session.

For Grades 3 through 5, the following content domains are addressed: Operations and Algebraic Thinking, Number and Operations in Base 10, Number and Operations – Fractions, and Measurement and Data/Geometry. For Grades 6 and 7, the following content domains are addressed: Ratios and Proportional Relationships, The Number System, Expressions and Equations, Geometry, and Statistics and Probability. For Grade 8, the following content domains are addressed: The Number System and Expressions and Equations, Functions, Geometry, and Statistics and Probability. Each content domain is represented approximately equally in terms of the number of available points on the test.

The K-PREP Blueprint provides targets for test development and helps to inform instruction for classroom teachers. The following table illustrates the plan for distributing the material at each grade level.

Table 14. Summary of the K-PREP Math Test blueprint targets for test development.

Subdomain	Target %		
	Grade 3	Grade 4	Grade 5
Operations and Algebraic Thinking	20-25	20-25	20-25
Number and Operations in Base Ten	20-25	20-25	20-25
Number and Operations – Fractions	25-30	20-25	20-25
Measurement and Data, Geometry (MDG)	25-30	25-30	25-30
Non-Calculator Test Percentage	20-25		

Subdomain	Target %	
	Grade 6	Grade 7
Ratios and Proportional Relationships (RP)	18-23	18-23
The Number System	18-23	18-23
Expressions and Equations (EE)	18-23	18-23
Geometry	18-23	18-23
Statistics and Probability	18-23	18-23
Non-Calculator Test Percentage	20-25	18-23

Subdomain	Target %
	Grade 8
The Number System and Expressions & Equations (NS/EE)	25-30
Functions (F)	20-25
Geometry	25-30
Statistics and Probability	20-25
Non-Calculator Test Percentage	20-25

Scaling and equating of K-PREP raw scores is accomplished via item response theory (IRT), using the 1-parameter logistic model for multiple-choice items and the partial-credit model for open-response items. K-PREP Math test scores are calibrated by grade level and will be psychometrically equated between test administration years. The scale ranges from 100 to 300 in each grade.

*The Quantile Framework for Mathematics.* The Quantile Framework was developed to assist teachers, parents, and students in identifying strengths and weaknesses in mathematics and forecast growth in overall mathematical achievement. Items and mathematical content are calibrated using the Rasch IRT model. The Quantile scale ranges from “EM” (Emerging Mathematician, 0Q and below) to above 1600Q. The Quantile Framework was developed to assess how well a student (1) understands the natural language of mathematics, (2) knows how to read mathematical expressions and employ algorithms to solve decontextualized problems, and, (3) knows why conceptual and procedural knowledge is important and how and when to apply it. The Quantile Framework Item Bank consists of multiple-choice items aligned with first grade content through Geometry, Algebra II, and Pre-calculus content and field tested with a national sample of students during the winter of 2004.

The Grade 3 Quantile Linking Test contained 40 multiple-choice items, Grade 4 had 42 items, Grade 5 had 47 items, Grade 6 had 44 items, Grade 7 had 43 items, and Grade 8 had 49 items. Each test was divided into two parts: calculator active and calculator inactive. During the calculator inactive portion of the tests, a calculator was not allowed to be used by the students to assist with answering the questions. During the calculator active portion of the tests, approved calculators are allowed on the tests.

The Quantile Linking Tests were constructed to be aligned to the K-PREP Math Tests in terms of their item content and difficulty. To achieve these goals, each item of the K-PREP Math Tests was categorized according to its content demand and assigned to one of the five content strands in The Quantile Framework for Mathematics (i.e., Numbers and Operations, Geometry, Measurement, Algebra/Patterns & Functions, and Data Analysis & Probability). In addition, each item on the K-PREP Math Tests was categorized according to its specific content within each strand and assigned a QTaxon. The Quantile Linking Test was built to match the proportion of items in each strand on the K-PREP Math Tests. The specific items on the Quantile Linking Tests also matched the QTaxon skills from the K-PREP Math Tests. Each test had a mean Quantile measure that aligned with the K-PREP Math Tests content (Grade 3, 421Q; Grade 4, 571Q; Grade 5, 710Q; Grade 6, 773Q; Grade 7, 819Q; Grade 8, 974Q). To the extent possible, the grade level each item on the Quantile Linking Test was initially calibrated matched the grade level of the K-PREP Math Tests. An exception to this guideline occurred when an item was to be used as an across-grade linking item and was selected from a higher or lower grade level.

*Evaluation of the Quantile Linking Tests.* After administration, the Quantile Linking Tests items were reviewed. The raw score descriptive statistics for all items and all students that took the Quantile Linking Tests are presented in *Table 15*.

Table 15. Descriptive statistics for the Quantile Linking Tests raw scores.

Grade	N*	Raw Score Mean (SD)	Minimum Score		Maximum Score	
			Observed	Possible	Observed	Possible**
3	908	27.94 (6.8)	0	0	40	40
4	1,170	24.65 (7.7)	6	0	41	42
5	1,029	25.10 (8.3)	6	0	46	46
6	1,330	25.56 (7.0)	7	0	43	44
7	1,478	24.84 (8.0)	5	0	43	43
8	944	25.86 (8.9)	3	0	46	47
<b>Total</b>	6,859					

\* N size reflects the removal of 54 students for missing, unusable, or duplicate student IDs.

\*\* One item removed from Grade 5 and two items from Grade 8.

Selected item statistics for the Quantile Linking Tests are presented in *Table 16*. Based on the item examination, one item was removed from the Grade 5 Quantile Linking Test and two items were removed from the Grade 8 Quantile Linking Test. These items exhibited low point-biserial correlations. These items were removed from any further analyses. While some items retained on the tests had low point-biserial correlations, the items performed adequately (average ability measure for the correct answer was highest compared to the average ability measures of the three distractors from Winsteps analyses).

Table 16. Item statistics from the development of the Quantile Linking Tests.

Grade	N* (Persons)	N** (Items)	Percent Correct Mean (Range)	Point-Biserial Range	Coefficient Alpha
3	908	40	70 (26 - 96)	0.18 - 0.58	0.87
4	1,170	42	59 (11 - 92)	0.11 - 0.50	0.88
5	1,029	46	55 (19 - 92)	0.13 - 0.52	0.88
6	1,330	44	58 (19 - 94)	0.05 - 0.45	0.84
7	1,478	43	58 (31 - 84)	0.08 - 0.51	0.87
8	944	47	55 (29 - 85)	0.17 - 0.53	0.88
<b>Total</b>	6,859				

\* N size reflects the removal of 54 students for missing, unusable, or duplicate student IDs.

\*\* One item removed from Grade 5 and two items from Grade 8.

Coefficient Alphas for each of the six Quantile Linking Tests, one for each grade, ranged from 0.84 to 0.88. These values indicate strong internal consistency reliability for each of the six tests and high consistency across the six tests.

## Study Design

A single-group/common person design was chosen for this study (Kolen and Brennen, 2004). This design is most useful “when (1) administering two forms to examinees is operationally possible, (2) differential order effects are not expected to occur, and (3) it is difficult to obtain participation of a sufficient number of examinees in an equating study that uses the random groups design” (pp. 16–17). The Quantile Linking Tests were administered between May 1, 2012 and June 6, 2012, within two weeks of the administration of the K-PREP Math Tests.

## Analysis of the K-PREP Math Test/Quantile Linking Test Sample

The sample of students for the study was recruited by the Kentucky Department of Education. The participating schools were located from across Kentucky with a total of 76 schools from 53 districts participating in the linking study.

Table 17 presents the number of students tested in the linking study and the percentage of students with complete data (both a K-PREP Math score and a Quantile Linking Test

Quantile measure). A total of 6,859 students (Grades 3 through 8), or 99.1%, had both test scores. This sample will be referred to as the calibration sample.

*Table 17.* Number of students sampled and number of students in the calibration sample.

Grade	State <i>N</i> Received	Quantile Linking Test <i>N</i>	Matched <i>N</i>	Matched Percent
3	50,572	912	908	99.6
4	49,411	1,195	1,170	97.9
5	50,672	1,029	1,029	100.0
6	50,446	1,332	1,330	99.8
7	49,452	1,490	1,478	99.2
8	49,267	955	944	98.8
<b>Total</b>	299,820	6,922	6,859	99.1

*Table 18* shows, for each grade level, the number of students (*N*) in the state and the proportion the *N*-count represents in the final sample. The table further summarizes the number of student test scores (by grade level) that were removed from analysis, and the reason for their removal. This sample will be referred to as the final sample. Of the 299,820 students in the state sample, 298,984 (99.7%) remain in the final sample.

*Table 18.* Comparison of state sample and final sample and the reason for student removal.

Grade	State <i>N</i> Received	Accommodated Students	Exempt Students	Missing Scale Scores	Final <i>N</i>	Final Percent
3	50,572	63	78	0	50,431	99.7
4	49,411	60	95	0	49,256	99.7
5	50,672	56	81	0	50,535	99.7
6	50,446	59	81	0	50,306	99.7
7	49,452	44	73	0	49,335	99.8
8	49,267	45	101	0	49,121	99.7
<b>Total</b>	299,820	327	509	0	298,984	99.7

*Table 19* presents the demographic characteristics of all students in the K-PREP Math state sample, the calibration sample, and the final sample of students included in this study. Across the samples, the final sample is virtually identical to the state sample with an n-count only differing by 836 students. Therefore, it is not necessary to report the state sample in any further tables.

*Table 19.* Percentage of students in the K-PREP Math state sample, the calibration sample, and the final sample for selected demographic characteristics.

Student Characteristic	Category	State Sample ( <i>N</i> = 299,820)	Calibration Sample ( <i>N</i> = 6,859)	Final Sample ( <i>N</i> = 298,984)
Gender	Female	48.9	49.4	48.9
	Male	51.1	50.6	51.1
Ethnicity*	African American	12.4	5.0	12.4
	American Indian	0.9	0.7	0.9
	Asian	1.5	0.6	1.5
	Hispanic	3.6	1.9	3.6
	Pacific Islander	0.2	0.1	0.2
	White	83.1	91.7	83.2

Student Characteristic	Category	State Sample (N = 299,820)	Calibration Sample (N = 6,859)	Final Sample (N = 298,984)
IEP	Yes	0.5	10.7	0.5
	No	99.5	89.3	99.5
LEP	Yes	2.1	0.6	2.0
	No	97.9	99.4	98.0
Migrant	Yes	0.2	0.1	0.2
	No	99.8	99.9	99.8
Homeless	Yes	3.7	3.0	3.7
	No	95.0	95.1	95.0
	Unknown	1.3	1.8	1.3
Lunch Status	Free	51.0	51.5	51.0
	Reduced	7.3	7.7	7.3
	No Lunch/Not Indicated	41.7	40.8	41.7
Accommodation	Yes	0.1	0.1	.
	No	99.9	99.1	100.0
Accommodation-Audio	Yes	0.5	0.2	0.5
	No	99.5	99.8	99.5
Accommodation-Braille	No	100.0	100.0	100.0
Accommodation-Large Print	Yes	0.1	0.1	0.1
	No	99.9	99.9	99.9
Exempt	Yes	0.2	0.0	0.0
	No	99.8	100.0	100.0
Grade	3	16.9	13.2	16.9
	4	16.5	17.0	16.5
	5	16.9	15.0	16.9
	6	16.8	19.4	16.8
	7	16.5	21.6	16.5
	8	16.4	13.8	16.4

*\* Does not add to 100% because more than one category can apply.*

Two steps were performed prior to the linking analysis. First, a concurrent calibration of the K-PREP Math Tests with Quantile items anchored to their theoretical values was conducted on the calibration sample to place the K-PREP Math items onto the Quantile scale. Students and items were submitted to a Winsteps analysis using a logit convergence criterion of 0.0001 and a residual convergence criterion of 0.003 (Linacre, 2011).

Second, a scoring run using only the K-PREP Math items on the Quantile scale was conducted to obtain calibrated Quantile measures for the students. These calibrated Quantile measures were used in the subsequent linking process.

Table 20 presents the descriptive statistics for the K-PREP Math scale score calibration sample as well as the calibrated sample Quantile Linking Test Quantile measure. Evaluating the Quantile measures on the K-PREP Math Tests and the Quantile Linking Tests show very comparable results. The correlations between the calibration sample K-PREP Math scale scores and the calibration sample Quantile measures range between 0.811 and 0.851. Based upon the correlations between the K-PREP Math scale scores and the Quantile Linking Test Quantile measures presented in Table 20, they are within the typical range of alternate-form reliability coefficients; therefore, the Quantile Linking Tests can be considered a T-parallel form of the K-PREP Math Tests (See Note 1).

Table 20. Descriptive statistics for the calibration sample K-PREP Math scale scores and Quantile measures and the calibration sample Quantile Linking Test Quantile measures.

Grade	<i>N</i>	Calibration Sample K-PREP Quantile Scale Score Mean (SD)	Calibration Sample K-PREP Math Quantile Measure Mean (SD)	Calibration Sample Quantile Linking Test Quantile Measure Mean (SD)	<i>r</i>
3	908	208.02 (18.7)	652.17 (215.2)	656.72 (203.3)	0.812
4	1,170	206.34 (17.3)	666.25 (207.6)	677.63 (185.5)	0.828
5	1,029	205.81 (16.9)	756.59 (188.1)	762.79 (184.9)	0.811
6	1,330	208.94 (16.5)	858.05 (161.9)	860.90 (180.0)	0.816
7	1,478	206.48 (17.0)	894.42 (175.9)	895.52 (183.2)	0.826
8	944	207.60 (16.7)	1009.16 (185.0)	1013.75 (183.7)	0.851
<b>Total</b>	6,859				

Based upon the correlations between the scale scores on K-PREP Math tests and the Quantile Linking Tests Quantile measures presented in *Table 20*, it can be concluded that the two tests are measuring similar mathematics constructs.

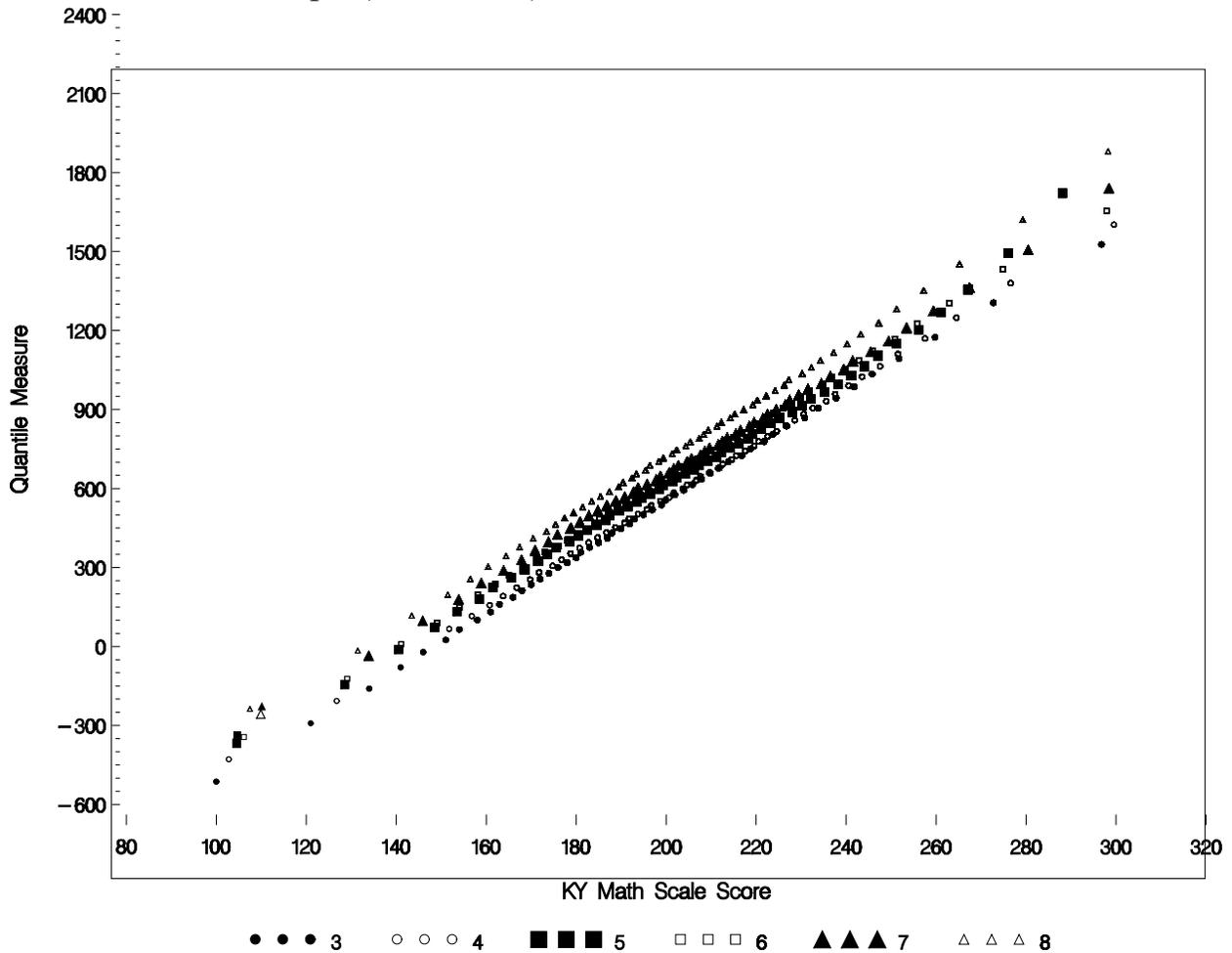
*Table 21* presents the descriptive statistics of the final sample K-PREP Math Test scale scores as well as the final sample Quantile Linking Test Quantile measures. The correlations between the two scores are perfect since the correlations are based on the same students taking a single set of items which are calibrated to two scales.

*Table 21.* Descriptive statistics for the final sample K-PREP Math scale scores and the final sample Quantile Linking Test Quantile measures.

Grade	<i>N</i>	Final Sample K-PREP Math Scale Score Mean (SD)	Final Sample Quantile Measure Mean (SD)	<i>r</i>
3	50,431	205.93 (19.0)	634.02 (206.7)	1.000
4	49,256	206.51 (16.7)	679.46 (179.0)	1.000
5	50,535	205.95 (17.2)	764.34 (188.6)	1.000
6	50,306	206.27 (17.7)	831.77 (193.2)	1.000
7	49,335	205.82 (18.2)	888.35 (196.8)	1.000
8	49,121	206.67 (17.6)	1003.52 (193.4)	1.000
<b>Total</b>	298,984			

*Figure 9* shows the relationship between the K-PREP Math scale scores and the Quantile Linking Test Quantile measures for the final sample for each grade. In each grade it can be seen that there is a linear relationship between the K-PREP Math scale score and the final sample Quantile measure reinforcing the use of linear equating.

Figure 9. Scatter plot of the K-PREP Math scale scores and the Quantile measures for the final sample ( $N = 298,984$ )



### Linking the K-PREP Math Scale with the Quantile Scale

Linking in general means “putting the scores from two or more tests on the same scale” (National Research Council, 1999, p.15). This study was designed to provide information that could be used to match students’ mathematical achievement with instructional resources—to identify the materials, concepts, and skills a student should be matched with for successful mathematical instruction, given their performance on the K-PREP Math Test.

*Linking Analyses.* Two score scales (e.g., the Quantile Scale and the K-PREP Math Test scale) can be linked using linear equating when the underlying item response models used to develop assessments are different. The linear equating method is most appropriate when (1) sample sizes are small; (2) test forms have similar difficulties; and

(3) simplicity in conversion tables or equations, in conducting analyses, and in describing procedures are desired (Kolen and Brennan, 2004).

In linear equating, a transformation is chosen such that scores on two tests are considered to be equated if they correspond to the same number of standard deviations above (or below) the mean in some group of examinees (Angoff, 1984, cited in Petersen, Kohen, and Hoover, 1989; Kolen and Brennan, 2004). Given scores  $x$  and  $y$  on tests  $X$  and  $Y$ , the linear relationship is

$$\frac{(x - \mu_x)}{\sigma_x} = \frac{(y - \mu_y)}{\sigma_y} \quad (\text{Equation 6})$$

and the linear transformation  $l_x$  (called the SD line in this report) used to transform scores on test  $Y$  to scores on text  $X$  is

$$x = l_x(y) = \left( \frac{\sigma_x}{\sigma_y} \right) y + \left( \mu_x - \frac{\mu_y \sigma_x}{\sigma_y} \right) \quad (\text{Equation 7})$$

Linear equating using an SD-line approach is preferable to linear regression because the tests are not perfectly correlated. With less than perfectly reliable tests, linear regression is dependent on which way the regression is conducted: predicting scores on test  $X$  from scores on test  $Y$  or predicting scores on test  $Y$  from scores on test  $X$ . The SD line provides the symmetric linking function that is desired.

The final linking equation between the K-PREP Math scale scores and the Quantile scale can be written as:

$$\text{Quantile measure} = \text{Slope(K-PREP Math scale score)} + \text{Intercept} \quad (\text{Equation 8})$$

where the slope is the ratio of the standard deviations of the K-PREP Math scale scores and Quantile Linking Test Quantile measures. These values can be found in *Table 21*.

Using the final sample data described in *Table 21*, the linear linking functions relating the K-PREP Math scale scores and Quantile measures for all students in the sample are presented in *Table 22*. Separate linking functions were developed for each grade of the K-PREP Math Test. Conversion tables were developed for each grade in order to express the K-PREP Math scores in the Quantile metric and were delivered to the Kentucky Department of Education in electronic format (also see Appendix B).

Table 22. Linear linking equation coefficients used to predict Quantile measures from the K-PREP Math scale scores.

Grade	Slope	Intercept
3	██████████	██████████
4	██████████	██████████
5	██████████	██████████
6	██████████	██████████
7	██████████	██████████
8	██████████	██████████

Table 23 contains the capped Quantile measures by grade. The measures that are reported for an individual student should reflect the purpose for which they will be used. If the purpose is instructional, then the scores should be capped at the upper bound of measurement error (e.g., at the 95<sup>th</sup> percentile point). In an instructional environment, all scores at or below 0Q should be reported as “EM” (Emerging Mathematician); no student should receive a negative Quantile measure.

Table 23. Capped values of the Quantile measure by grade.

Grade	Capped Quantile Measure
3	██████
4	██████
5	██████
6	██████
7	██████
8	██████

### Validity of the K-PREP Math Test - Quantile Link

Table 24 contains the percentile ranks of the Quantile Linking Test Quantile measures and the K-PREP Math Test Quantile measures (based on the final sample). The criterion

of a half standard deviation (100L) on the Quantile scale was used to determine the size of the difference. In examining the values, there were some differences in the tails of the distributions. Since few students receive extreme scores, this does not cause reason for concern. This supports the use of Quantile measures on the K-PREP Math Test.

*Table 24.* Comparison of the Quantile measures for selected percentile ranks on the Quantile Linking Test and the final sample K-PREP Math Test.

Grade 3			Grade 4		
Percentile Rank	Linking Test Quantile Measure	K-PREP Math Sample Quantile Measure	Percentile Rank	Linking Test Quantile Measure	K-PREP Math Sample Quantile Measure
1	194	199	1	268	341
5	319	308	5	353	405
10	385	363	10	403	470
25	520	482	25	520	545
50	646	635	50	635	663
75	815	776	75	810	781
90	981	907	90	940	921
95	1069	983	95	1024	996
99	1210	1135	99	1229	1168

Table 24 (continued). Comparison of the Quantile measures for selected percentile ranks on the Quantile Linking Test and the final sample K-PREP Math Test.

Grade 5		
Percentile Rank	Linking Test Quantile Measure	K-PREP Math Sample Quantile Measure
1	398	392
5	502	491
10	549	535
25	634	633
50	734	754
75	877	886
90	1022	1017
95	1120	1094
99	1269	1269

Grade 6		
Percentile Rank	Linking Test Quantile Measure	K-PREP Math Sample Quantile Measure
1	502	458
5	607	534
10	652	600
25	756	687
50	859	818
75	968	960
90	1073	1080
95	1138	1156
99	1265	1330

Grade 7		
Percentile Rank	Linking Test Quantile Measure	K-PREP Math Sample Quantile Measure
1	529	523
5	632	610
10	675	653
25	774	739
50	886	858
75	1008	1009
90	1111	1138
95	1215	1236
99	1319	1419

Grade 8		
Percentile Rank	Linking Test Quantile Measure	K-PREP Math Sample Quantile Measure
1	657	622
5	733	721
10	776	765
25	875	864
50	1001	996
75	1114	1117
90	1282	1249
95	1346	1326
99	1492	1513

Performance levels provide a common meaning of test scores throughout a state concerning what is expected at various levels of competence. In Kentucky, the K-PREP Math Tests are designed to use the Kentucky Core Academic Standards (KCAS) as the foundation which are the Common Core State Standards for Mathematics. They are designed to identify and define what a student knows and can do at a specific grade and to help parents, educators, and students understand the performance-level scores a student receives on the K-PREP Math Tests. *Table 25* presents the performance level cut scores on the K-PREP Math Test and the associated Quantile measures. The Kentucky Department of Education recognizes four categories called performance levels: Novice, Apprentice, Proficient, and Distinguished. The values in the table are the cut scores associated with the bottom score for each category.

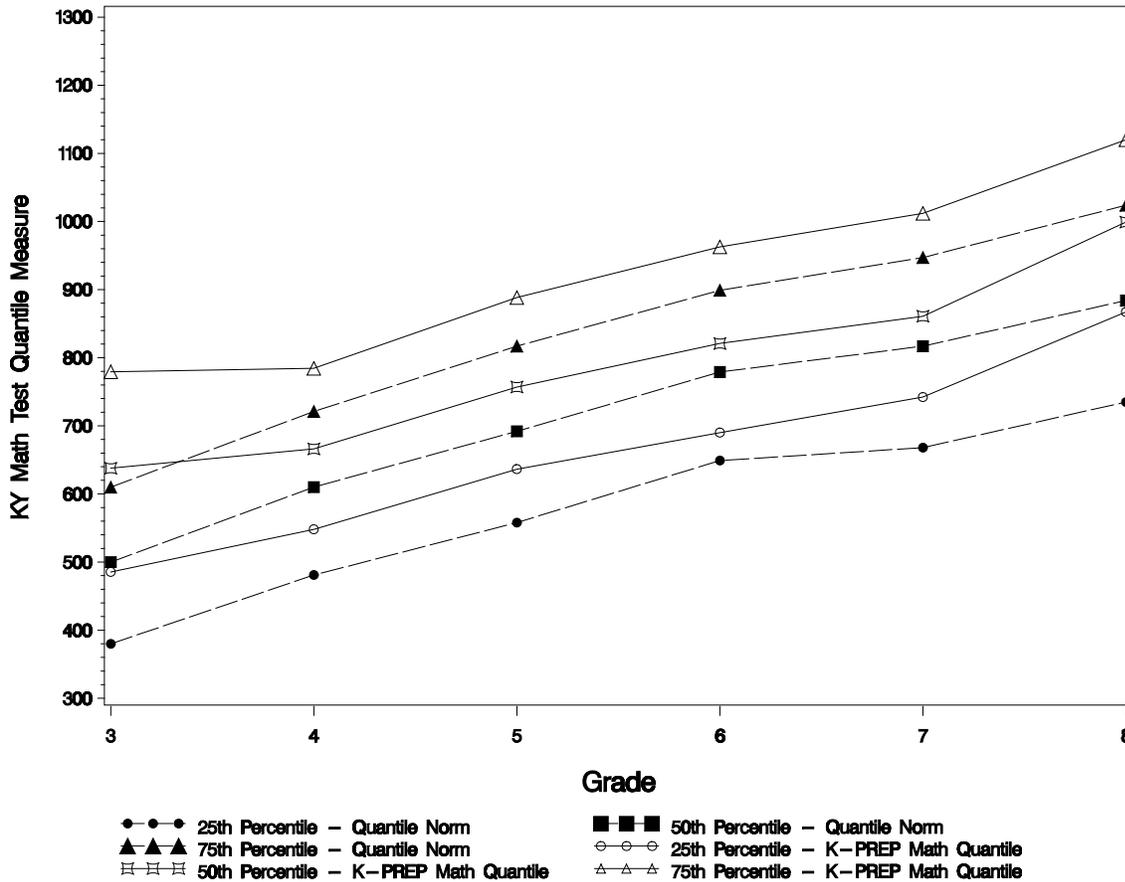
*Table 25.* Performance level cut scores on the K-PREP Math test and the associated Quantile measures.

Grade	Apprentice		Proficient		Distinguished	
	K-PREP Math Scale Score	Quantile Measure	K-PREP Math Scale Score	Quantile Measure	K-PREP Math Scale Score	Quantile Measure
3	192	██████	210	██████	234	██████
4	194	██████		██████		229
5	192	██████	210	██████	229	██████
6	191	██████	210	██████	231	██████
7	192	██████	210	██████	231	██████
8	192	██████	210	██████	232	██████

The next graph shows the Quantile measures for the K-PREP Math tests final sample Quantile measures and the Quantile norms. These norms were created based on linking studies conducted with the Quantile Framework. The sample’s distribution of scores from this study was similar to the distribution of scores on norm-referenced assessments and other standardized measures of mathematics achievement. The results compared favorably with other mathematics measures which reinforced MetaMetrics’ confidence in the Quantile norms.

As can be seen in *Figure 10*, the Quantile measures for the K-PREP Math Tests are higher than the Quantile measure norms. This indicates that the state sample in this study is more able than the Quantile norms. This is especially true in Grade 3.

Figure 10. Selected Percentiles (25<sup>th</sup>, 50<sup>th</sup>, and 75<sup>th</sup>) plotted for the K-PREP Math Quantile measures for the final sample ( $N = N=298,984$ ) against the Quantile measure norms.



The following box and whisker plots (Figures 11 and 12) show the progression of scores (the  $y$ -axis) from grade to grade (the  $x$ -axis). For each grade, the box refers to the interquartile range. The line within the box indicates the median and the  $\bullet$  indicates the mean. The end of each whisker shows the minimum and maximum values of the Quantile Linking Tests Quantile measures and the K-PREP Math tests Quantile measures for each grade (the  $y$ -axis). The Quantile measures are on a vertical scale and Figures 11 and 12 demonstrate this by showing that as the grade increases so do the Quantile scores on the K-PREP Math tests. The pattern of Quantile measures is the same for each figure.

Figure 11. Box and whisker plot of the Quantile Linking Tests Quantile measures by grade, calibration sample (N = 8,959).

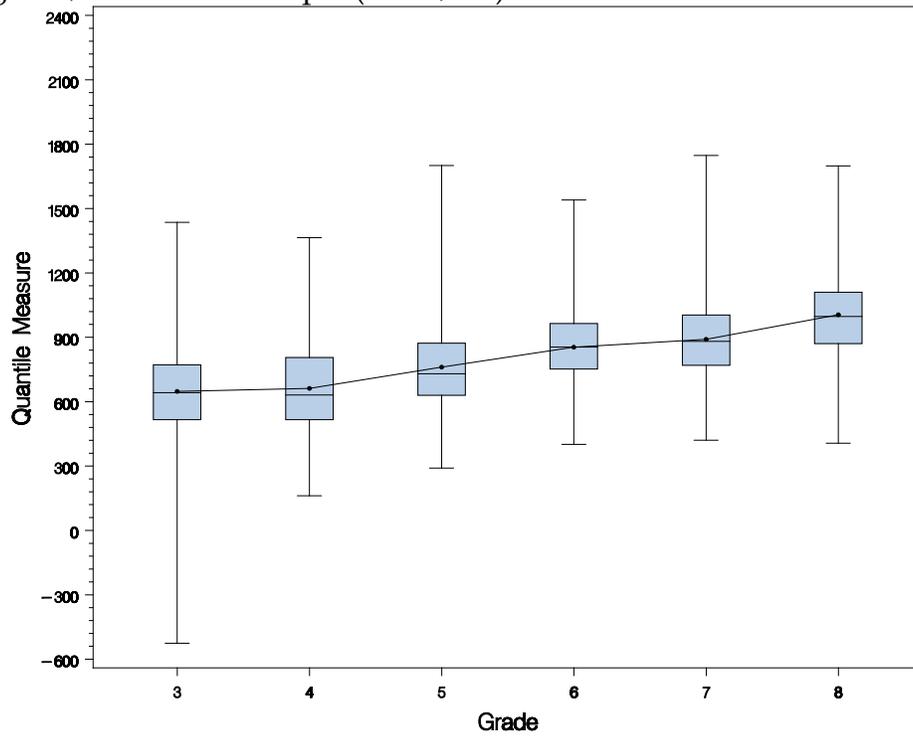
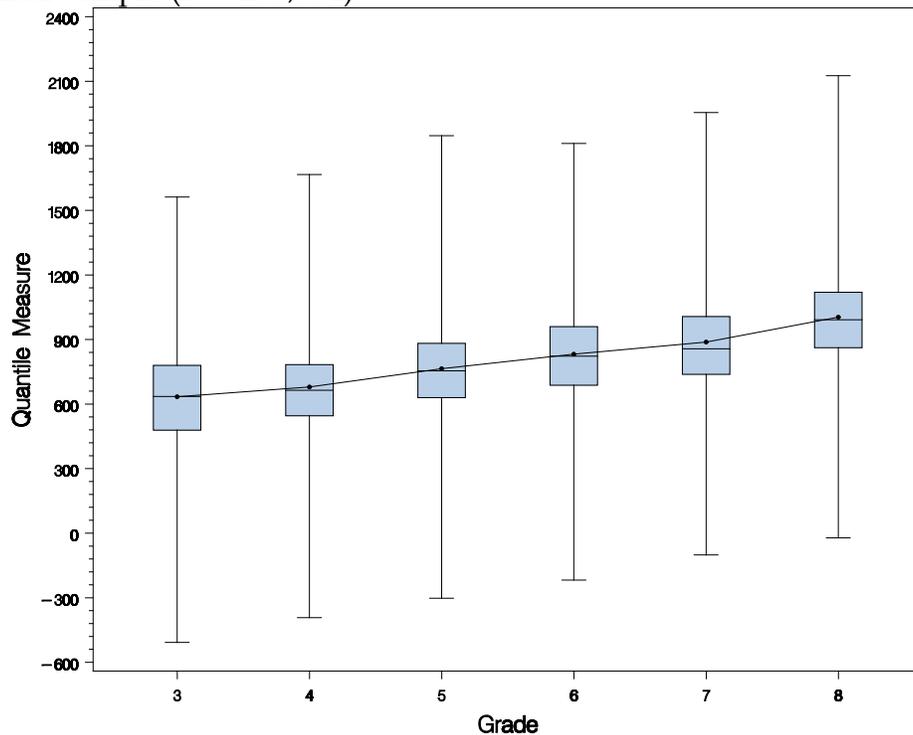


Figure 12. Box and whisker plot of the K-PREP Math test Quantile measures by grade, final sample (N = 298,984).



## Principal Components Analysis

In order to further examine the construct being measured by the K-PREP Math Test and the Quantile Linking Test, principal components analyses were performed. For each grade, the items from the K-PREP Math Test and Quantile Linking Test were included in that grade's principal components analysis. This investigation was undertaken to determine if the two tests measure the same construct and the overall results could therefore be considered unidimensional.

Table 26 shows the first five principal components based on the largest eigenvalues for each grade. The first component is large and the subsequent components show small eigenvalues. This lends itself to the conclusion that there is one primary component for the items from the two tests.

Table 26. Eigenvalues associated with the first five components in PCA of K-PREP Math and LT items.

Grade	Principal Components				
	1	2	3	4	5
3	24.152106	3.462631	2.411448	2.252719	2.011075
4	23.251613	3.636585	2.256565	1.894274	1.829318
5	22.769664	3.222487	2.407136	2.238558	1.935168
6	21.399884	3.058488	2.296837	2.185316	1.866030
7	23.918976	3.921680	2.442217	1.744154	1.648346
8	24.572150	2.654300	2.152032	2.076297	1.914398

The proportion of variance explained by the first five components is presented in Table 27. Again, the first component explains the most variance. If the first component is at least 20% of the variance, this is an indication of unidimensionality (Reckase, 1979).

Table 27. Proportion of variance explained by the first five components in PCA of K-PREP Math and LT items.

Grade	Proportion Explained by Each Component				
	1	2	3	4	5
3	0.277610	0.039800	0.027718	0.025893	0.023116
4	0.258351	0.040407	0.025073	0.021047	0.020326
5	0.242230	0.034282	0.025608	0.023814	0.020587
6	0.232607	0.033244	0.024966	0.023753	0.020283
7	0.262846	0.043095	0.026838	0.019167	0.018114
8	0.258654	0.027940	0.022653	0.021856	0.020152

Table 28 presents the results to assess the reasonableness of an assumption of unidimensionality. In general, the assumption of unidimensionality is considered tenable if the first eigenvalue is large and the second eigenvalue is not much larger than the remaining eigenvalues (Lord, 1980). The criterion of what is “large,” however, has not been formally operationalized (Hattie, 1985). This method calculates the ratio of the first to second eigenvalues. The ratios in Table 28 are large. Based on Gorsuch’s (1983) suggestion that if the ratios are greater than or equal to three, then these ratios are consistent with the assumption of unidimensionality.

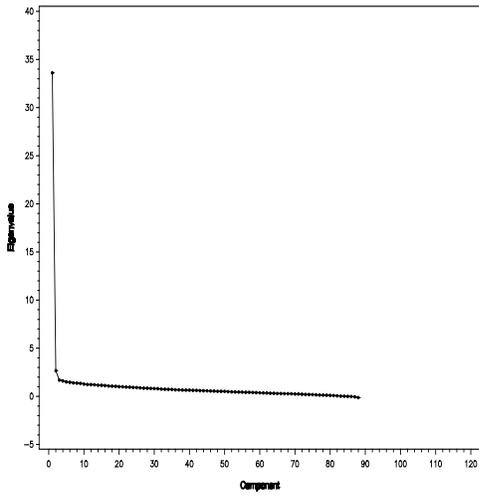
Table 28. Ratio that assess unidimensionality for principal components analysis.

Grade	Ratio
3	6.975074
4	6.393804
5	7.065867
6	6.996884
7	6.099166
8	9.257489

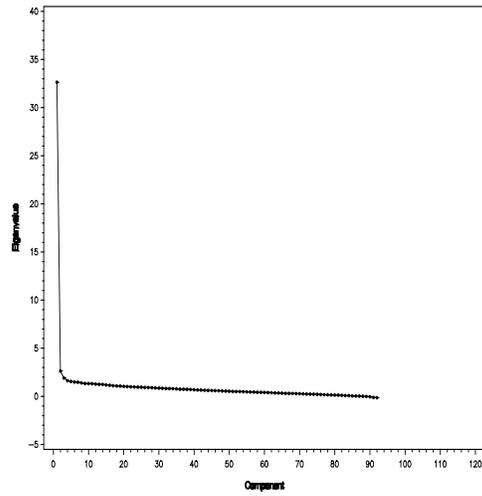
The scree plots for each grade in Figure 13 show a strong first factor based on the steep decrease from the first component to the second and subsequent components.

Figure 13. Scree plots for the principal components analysis by grade.

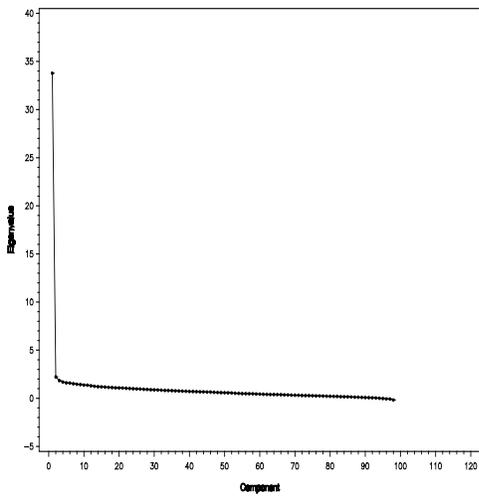
Grade 3



Grade 4



Grade 5



Grade 6

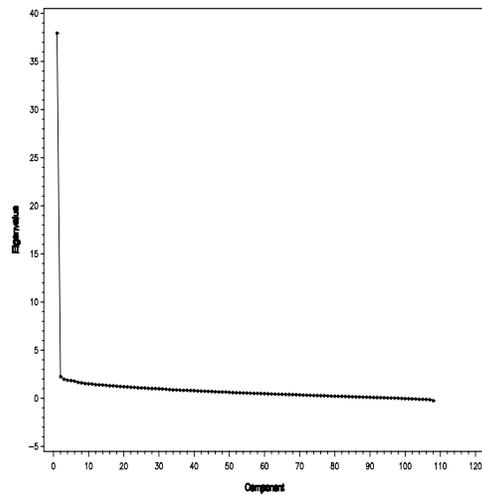
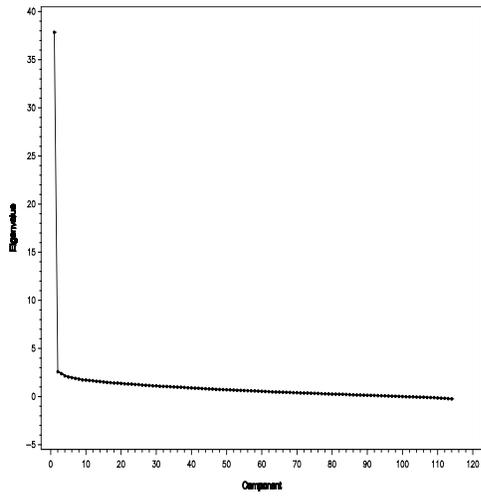
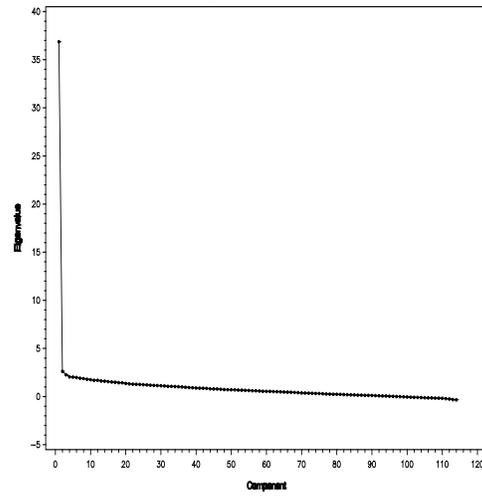


Figure 13 (continued). Scree plots for the principal components analysis by grade.  
Grade 7



Grade 8



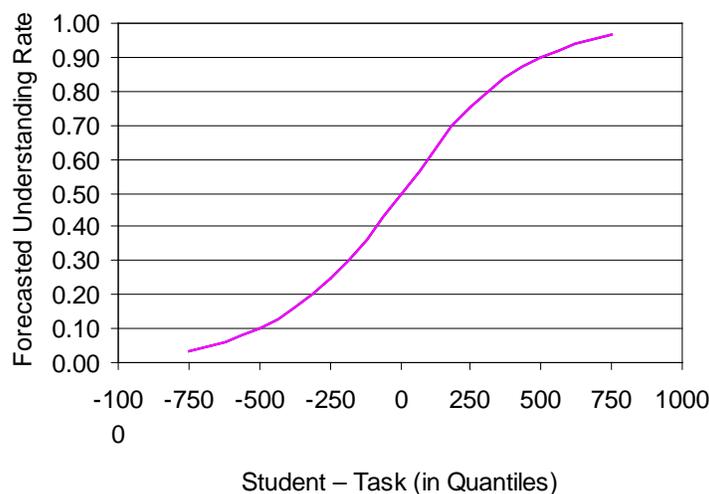
## Quantile Framework and Instruction

Quantile measures are available from many norm-referenced and criterion-referenced assessments, in addition to state tests and instructional products. Students who take a mathematics achievement test that is linked with the Quantile Framework or one that reports directly in the Quantile metric will receive a Quantile measure. Educators can use these Quantile measures to match students, by *readiness level*, to level-appropriate instructional materials and forecast understanding. For example, a student with a Quantile measure of 500Q should be ready for instruction of mathematics problems at a demand level of 500Q.

*Differentiated Instruction.* A Quantile measure for materials is a number indicating the mathematical demand of the material in terms of the concept/application solvability. The Quantile measure for an individual student is the level at which he or she is ready for instruction (50% competency with the material) and has knowledge of the prerequisite mathematical concepts and skills necessary to succeed. The Quantile scale ranges from Emerging Mathematician (0Q and below) to above 1600Q. The Quantile measure does not relate to a specific grade, *per se*, so the score is developmental as it spans the mathematics continuum from kindergarten mathematics through the content typically taught in Algebra II, Geometry, Trigonometry, and Pre-calculus. The measure can be used by a teacher to determine what mathematical instruction the student is likely to be ready for next.

*Figure 14* shows the general relationship between the student-task discrepancy and forecasted understanding. When the student measure and the task mathematical demand are the same (difference of 0Q), then the forecasted understanding, or success rate, is modeled as 50% and the student is likely ready for instruction on the skill or concept.

Figure 14. Relationship between student mathematical demand discrepancy and forecasted understanding (success rate).



An appropriate instructional range for the Quantile measure of a student is 50Q above and 50Q below the Quantile measure of the student (44% - 56% competency). This range identifies the “learning frontier” of mathematics skills in which a student has the prerequisite knowledge and skills needed to understand the instruction and will likely have success with tasks related to the skill/concept after this introductory instruction.

Quantile measures provide reliable, actionable results because instruction and assessment are described using the same metric. When instruction is measured at a unique mathematical level of understanding and any form of assessment can be reported using the same scale, equal levels of achievement are observed.

By understanding the interaction between student measures and resource measures (e.g., textbook lessons, instructional materials), any level of understanding can be used as a benchmark. An individual can modulate his or her own likely success rate by lowering the difficulty of the task (i.e., increase to 90% understanding) or increasing the difficulty of the task (i.e., lower to 40% understanding) depending on the situation (refer to Figure 14). This flexibility allows the teacher, parent, or student the ultimate control to modulate the fit between person and task.

The primary utility of the Quantile Framework is its ability to forecast what will likely happen when students confront resources and instruction on specific mathematical skills and concepts. With every application by teacher, student, or parent there is a test of the Framework’s accuracy. The Framework makes a point prediction every time a resource or lesson is chosen for a student. Anecdotal evidence suggests that the Quantile Framework predicts as intended. That is not to say that there is an absence of error in forecasted understanding. There is error in resource measures based on QTaxon

(mathematical skills and concepts) measures, student measures, and their difference modeled as forecasted understanding. However, the error is sufficiently small that the judgments about students, resources, and understanding rates are useful.

The subjective experience of 25%, 50%, and 75% understanding/success as reported by students varies greatly. A 1000Q student being instructed on 1000Q QTaxons (50% understanding) has a successful instructional experience – he has the background knowledge needed to learn and apply the new information. Teachers working with such a student report that the student can engage with the skills and concepts that are the focus of the instruction and, as a result of the instruction, are able to solve problems utilizing those skills. In short, such students appear to understand what they are learning. A 1000Q student being instructed on 1200Q QTaxons (25% understanding) encounters so many unfamiliar skills and difficult concepts that the learning is frequently lost. Such students report frustration and seldom engage in instruction at this level of understanding. Finally, a 1000Q student being instructed on 800Q QTaxons (75% understanding) reports that he is able to engage with the skills and concepts with minimal instruction, is able to solve complex problems related to the skills and concepts, is able to connect the skills and concepts with skills and concepts from other strands, and experiences fluency and automaticity of skills.

*Quantile Framework and the CCSS.* Although states have developed their own individual curriculum standards for years, recently there has been an unprecedented focus on developing common curriculum standards for use throughout the United States of America. Guided and supported by the Council of Chief State School Officers (CCSSO) and the National Governors Association (NGA), departments of education in the states, the United States territories and the District of Columbia have collaborated to identify common standards in English/language arts, mathematics and other content areas. Educators, researchers and educational policy makers were involved extensively in the effort to identify, catalog, review and adopt standards that would lead to students being “college and career ready” by the end of high school. The Common Core State Standards (CCSS) are the culmination of this work. They were released in June 2010 by the CCSSO and the NGA Center for Best Practices. As of March 2011 forty states, the District of Columbia and the U.S. Virgin Islands had adopted the standards. Currently, forty-five states have adopted the CCSS for Mathematics. The standards may be viewed at <http://www.corestandards.org/> (NGA Center & CCSSO, 2010a, 2010b). Additional information about the development of the CCSS may be found at the CCSSO website (<http://www.ccsso.org/>) and the website of the NGA (<http://www.nga.org/>).

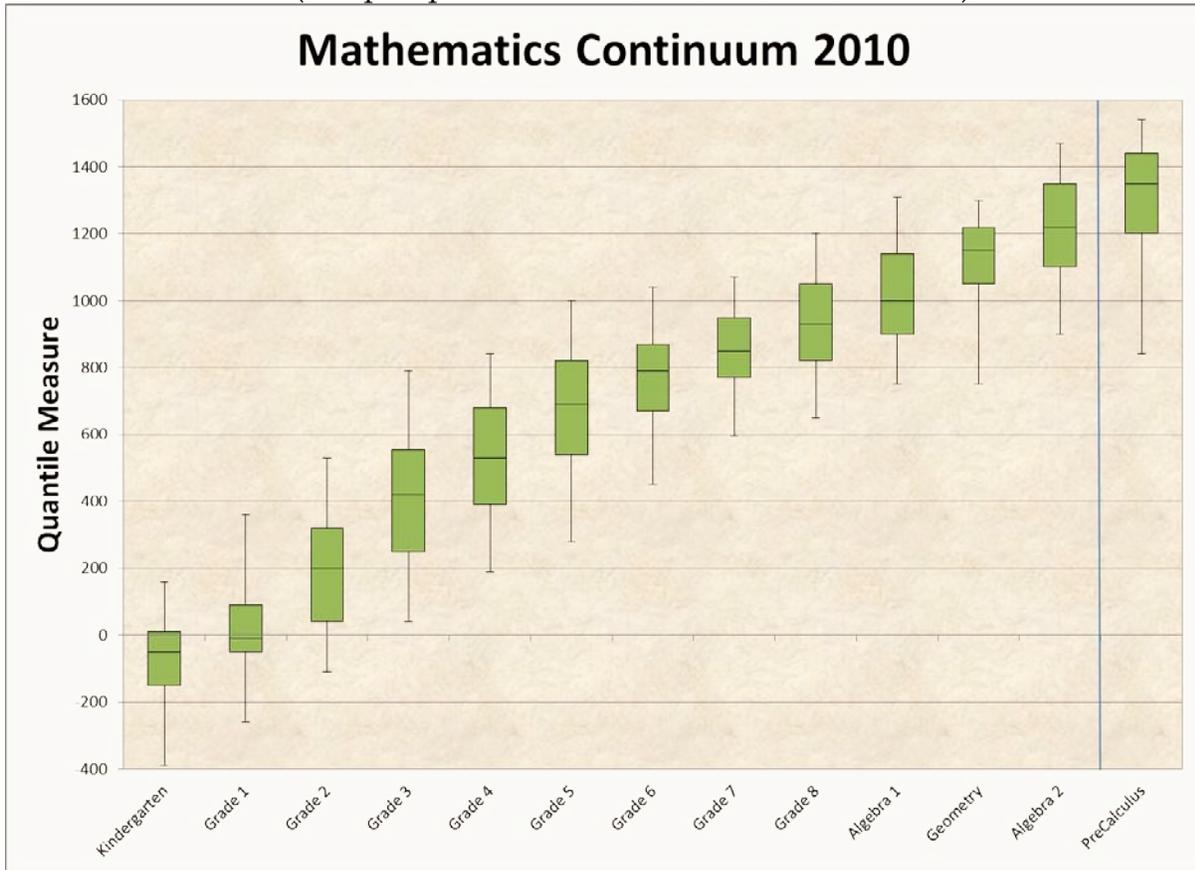
There is increasing recognition of the importance of bridging the gap that exists between K-12 and higher education and other postsecondary endeavors. Many state and policy leaders have formed task forces and policy committees such as P-20 councils. The Common Core State Standards (CCSS) for Mathematics were designed to enable all students to become college and career ready by the end of high school while

acknowledging that students are on many different pathways to this goal: “One of the hallmarks of the Common Core State Standards for Mathematics is the specification of content that all students must study in order to be college and career ready. This ‘college and career ready line’ is a minimum for all students” (NGA Center & CCSSO, 2010b, p. 4). The CCSS for Mathematics suggest that “college and career ready” means completing a sequence that covers Algebra I, Geometry, and Algebra II (or equivalently, Integrated mathematics 1, 2 and 3) during the middle school and high school years; and, leads to a student’s promotion into more advanced mathematics by their senior year. This has led some policy makers to generally equate the successful completion of Algebra II as a working definition of college and career ready. Exactly how and when this content must be covered is left to the states to designate in their implementations of the CCSS for Mathematics throughout K-12.

The *mathematical demand* of a mathematical textbook (in Quantile measures) quantitatively defines the level of mathematical achievement that a student needs in order to be ready for instruction on the mathematical content of the textbook. Assigning QTaxon(s) and Quantile measures to a textbook is done through a calibration process. Textbooks were analyzed at the lesson level and the calibrations were completed by subject matter experts (SMEs) experienced with the Quantile Framework and with the mathematics taught in mathematics classrooms. The intent of the calibration process is to determine the mathematical demand presented in the materials. Textbooks contain a variety of activities and lessons. In addition, some textbook lessons may include a variety of skills. Only one Quantile measure is calculated per lesson and is obtained through analyzing the Quantile measures of the QTaxons that have been mapped to the lesson. This Quantile measure represents the composite task demand of the lesson.

MetaMetrics has calibrated more than 41,000 instructional materials (e.g., textbook lessons, instructional resources) across the K-12 mathematics curriculum. *Figure 15* shows the continuum of calibrated textbook lessons from Kindergarten through Pre-calculus where the median of the distribution for Pre-calculus is 1350Q. The range between the first quartile and the median of the first three chapters of Pre-calculus textbooks is from 1200Q to 1350Q. This range describes an initial estimate of the mathematical achievement level needed to be ready for mathematical instruction corresponding to the “college and career readiness” standard in the Common Core State Standards for Mathematics.

Figure 15. A continuum of mathematical demand for Kindergarten through Pre-calculus textbooks (box plot percentiles: 5<sup>th</sup>, 25<sup>th</sup>, 50<sup>th</sup>, 75<sup>th</sup>, and 95<sup>th</sup>).



This information describing college and career reading in mathematics can be used to interpret the K-PREP Math Tests performance standards. For each grade the “Proficient” range of Quantile measures as defined by the K-PREP Math Tests is compared to the mathematical demands in the next grade. As can be seen in Figure 16, almost all students scoring at the “proficient” level should be prepared for the mathematical demands of the next grade.

Figure 16. K-PREP Math “proficient” ranges (expressed as Quantile measures) compared with the mathematical demands of the next grade, by grade.

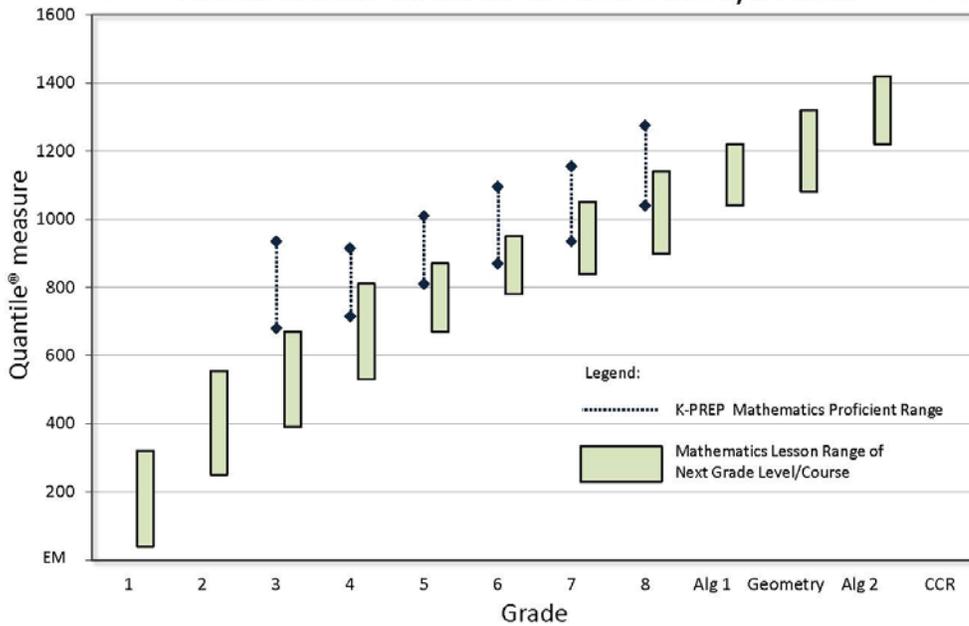
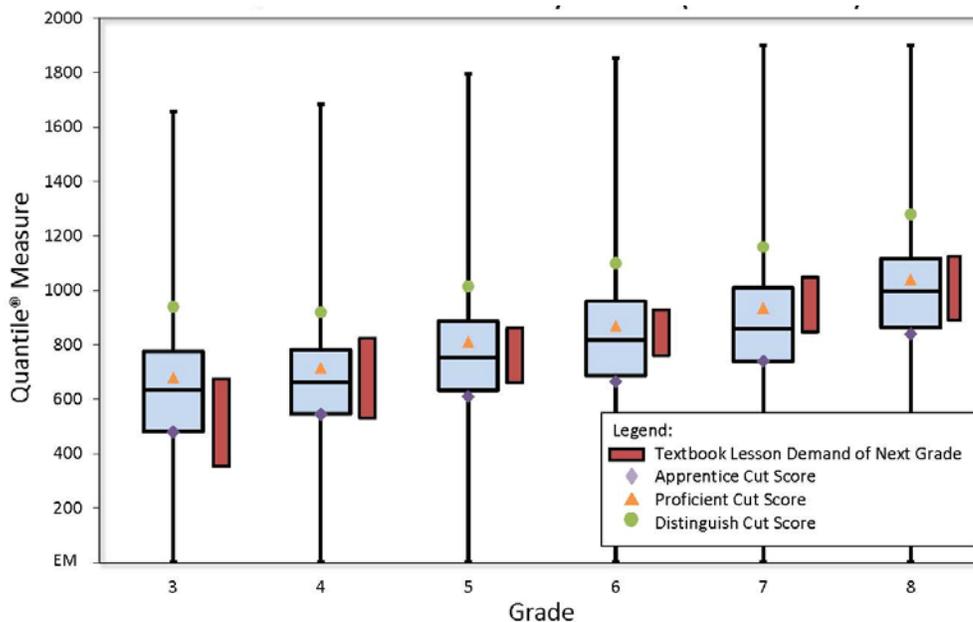


Figure 17 shows that the spring 2012 student performance on the K-PREP Math Test at each grade level is “on track” for college and career readiness in Grades 3 and 4. In comparing the performance of students in Grades 5 through 8, some students will need encouragement with supplemental materials at the next grade. Students can be matched with mathematics materials that are at or above the recommendations in the Common Core State Standards for each grade.

Figure 17. K-PREP Math 2011-2012 student performance expressed as Quantile measures.



In 2010, MetaMetrics and the Kentucky Department of Education (KDE) conducted a study to link the Kentucky Core Content Test in mathematics with the Quantile scale (MetaMetrics, 2011). The minimum score considered “proficient” at each grade level on the KCCT in mathematics is presented in *Table 29*. In 2012, KDE transitioned their assessment program to the K-PREP Math Test to align with the Common Core State Standards in Mathematics and to describe student mathematics performance in relation to college and career readiness. One outcome of this change was to set the performance standards for K-PREP Math Test at a higher level such that a smaller proportion of students would likely reach the “proficient” level (Ujifusa, 2012). For comparison purposes, the minimum “proficient” score for the K-PREP Math Test is also repeated from *Table 25*. The Quantile scale can be used as an external “yardstick” to evaluate this change in mathematics demand on the Kentucky math assessment. The information in *Table 29* shows the K-PREP Math Test standards are demanding more of students in terms of mathematics ability.

Table 29. Minimum “proficient” score on KCCT in Mathematics and K-PREP Math.

Grade	KCCT Proficient Cut Score (2010)	K-PREP Proficient Cut Score (2012)
3	435Q	680Q
4	585Q	715Q
5	685Q	810Q
6	760Q	870Q
7	830Q	935Q
8	920Q	1040Q

## Conclusions, Caveats, and Recommendations

Forging a link between scales is a way to add value to one scale without having to administer an additional test. Value can be in the form of any or all of the following:

- increased *interpretability* (e.g., “Based on this test score, what mathematical skills and concepts does my child actually know?”),
- increased *diagnostic capability* (e.g., “Based on this test score, what are the student’s weaknesses?”), or
- increased *instructional use* (e.g., “Based on these test scores, I need to modify my instruction to include these skills.”).

The link that has been established between the K-PREP Math test and the Quantile Framework permits students to be matched with resources and materials that provide an appropriate level of challenge while avoiding frustration. The result of this purposeful match may be that students will be less fearful of mathematics, and, thereby become better mathematical thinkers. The real power of the Quantile Framework is in examining the growth in mathematical achievement of students – wherever the student may be in the development of his or her mathematical skills and concepts. Students can be matched with resources and materials for which they are forecasted to experience 50% understanding, therefore, they are ready for instruction on the topic. As a student’s mathematical achievement grows, he or she can be matched with more demanding skills and concepts. And, as the skills and concepts become more demanding, then the student grows.

The development of the link between the scores on the K-PREP Math test and the Quantile scale has been described and evaluated in this study. There are many factors that can affect the linking process. In this study two of the factors include:

- sample characteristics (e.g., gender, ethnicity), and
- relationship of sample distribution characteristics to the distribution characteristics of the state.

*Conventions for Reporting.* Quantile measures are reported as a number followed by a capital “Q” for “Quantile.” There is no space between the measure and the “Q” and measures of 1,000 or greater are reported without a comma (e.g., 1050Q). All Quantile person measures should be rounded to the nearest 5Q to avoid over interpretation of the measures. As with any test score, uncertainty in the form of measurement error is present.

*Next Steps.* To utilize the results from this study, Quantile measures need to be incorporated into the K-PREP Math test results processing and interpretation

frameworks. Suggested resources need to be developed for ranges of students. Care must be taken to ensure that the resources and materials on the lists are also developmentally appropriate for the students. The Quantile measure is one factor related to understanding and is a good starting point in the selection process of materials and resources for a specific student. Other factors such as student developmental level, motivation, and interest; amount of background knowledge possessed by the student; and characteristics of the resources and skills also need to be considered when matching resources and instruction with a student.

In this era of student-level accountability and high-stakes assessment, differentiated instruction—the attempt “on the part of classroom teachers to meet students where they are in the learning process and move them along as quickly and as far as possible in the context of a mixed-ability classroom” (Tomlinson, 1999)—is a means for all educators to help students succeed. Differentiated instruction promotes high-level and powerful curriculum for all students, but varies the level of teacher support, task complexity, pacing, and avenues to learning based on student readiness, interest, and learning profile. One strategy for managing a differentiated classroom suggested by Tomlinson is the use of multiple resources and supplementary materials that can be identified with the aid of the Quantile Framework. Equipped with a student’s Quantile measure, teachers can connect him or her to textbook lessons, worksheets, games, websites, and trade books that have appropriate Quantile measures. By incorporating Quantile measures into the planning of mathematics instruction, it becomes possible to forecast with greater probability how successfully students are likely to understand the material presented to them. Teachers can provide instruction on QTaxons with Quantile measures below the targeted instruction when students are not ready for that instruction by focusing on prerequisite QTaxons. On the other hand, teachers can focus enrichment activities on the impending QTaxons.

Two resources are available on the Quantile Framework website – Quantile Teacher Assistant and Math@Home. In order to support instruction with the many resources connected with the Quantile Framework, the Quantile Teacher Assistant (QTA) was developed to simplify and gather all relevant information. When using the QTA (<http://qta.quantiles.com/>), teachers can identify a specific state objective and determine the knowledge base. In addition, teachers can differentiate instruction by indicating the range of Quantile measures for their students in their classrooms. Math@Home (<http://mah.quantiles.com/>) activities reinforce mathematical skills covered in the previous school year and lay the groundwork for what will be taught when students return to class in the fall. By incorporating fun family games into everyday activities, students can practice mathematical skills year-round and parents can feel more confident about helping their children with mathematics.

The following is a list of suggestions that can be used to leverage a student's Quantile measure in the classroom:

- Start class with warm-up problems and activities related to the prerequisite skills from a knowledge cluster.
- Enhance major themes of mathematics by building a bank of skills at varying levels that not only support a theme but also provide a way for all students to participate in the theme successfully. For example, consider how addition progresses from single numbers to multi-digit numbers, and then moves to decimals and fractions.
- Sequence mathematical skills according to their difficulty as much as possible.
- Develop a mathematics folder that goes home with students and returns weekly for review. The folder can contain examples of practice skills within a student's range, applications of topics outside the classroom, reports of recent assessments, and a parent form to record the amount of time spent working mathematics problems at home.
- Choose skills lower in a student's Quantile range when factors make the student view mathematics as more challenging, threatening, or unfamiliar. Select skills at or above a student's range to stimulate growth, when a topic holds high interest for a student, or when additional support such as background teaching or peer tutoring is provided.
- Develop individualized lists of skills that are tailored to provide appropriately challenging and curriculum suitable for all students.

Below are some suggestions related to leveraging a student's Quantile measure at home:

- Ensure that each child gets plenty of mathematical practice, concentrating on skills within his or her Quantile range. Parents can ask their child's teacher to print a list of appropriate skills or search the mathematics skill database on the Quantile website.
- Communicate with the child's teachers about the child's mathematical needs and accomplishments. They can use the Quantile scale to describe their assessment of the child's mathematical achievement.
- When a new topic proves too challenging for a child, use activities or other materials from the Web site to help. Review the prerequisite QTaxons to ensure that gaps or misconceptions are not interfering with the current topic.
- Celebrate a child's mathematical accomplishments. The Quantile Framework provides an easy way for students to track their own growth. Parents and children can set goals for mathematics—spending so much time daily working on mathematical problems, discussing situational topics such as statistics from a newspaper or discounts at the store, reading a book about a

mathematical topic, trying new kinds of Web sites and games, or working a certain number of mathematics problems per week. When children reach the goal, make it an occasion!

## Notes

1. A T-parallel test is a test that is designed to be “theoretically parallel” to another test in that it has the same number of items/ points, the same overall level of difficulty in terms of raw score means and standard deviations, and assesses the same construct domain (MetaMetrics, Inc., 1998).

## References

- American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.
- Anastasi, A. (1982). *Psychological Testing* (5<sup>th</sup> ed.). New York: MacMillan Publishing Company, Inc.
- Bond, T.G. & Fox, C.M. (2001). *Applying the Rasch model: Fundamental measurement in the human sciences*. Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.
- Burg, S.S. (2007). *An investigation of dimensionality across grade levels and effects on vertical linking for elementary grade mathematics achievement tests*. University of North Carolina, Chapel Hill, NC.
- Camilli, G. & Shepard, L.A. (1994). *Methods for identifying biased test items*. Thousand Oaks, CA: Sage Publications, Inc.
- Cochran, W.G. & Cox, G.M. (1957). *Experimental Designs*. New York: John Wiley & Sons.
- Davis, F. (1944). Fundamental factors of comprehension in reading, *Psychometrika*, 9, 185-197.
- 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.
- Emenogu, B.C. & Childs, R.A. (2005). Curriculum, translation, and differential functioning of measurement and geometry items. *Canadian Journal of Education*, 28(1 & 2), 128-146.
- Geary, D.C., & Hamson, C.O. (2002). Improving the mathematics and science of achievement of American children: Psychology's role. *American Psychological Association Online*. Retrieved May 23, 2002, from <http://www.apa.org/ed/geary.html>.
- Goldstein, H. (2003). International comparisons of student attainment: Some issues arising from the PISA study. Retrieved December 14, 2004, from [www.ioe.ac.uk/hgpersonal](http://www.ioe.ac.uk/hgpersonal).

- Gorsuch, R.L. (1983). *Factor analysis*. Mahwah, NJ: Lawrence Erlbaum Associates Inc.
- Haladyna, T.M. (1994). *Developing and validating multiple-choice test items*. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Hambleton, R.K. & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Kluwer-Nijhoff Publishing.
- Hambleton, R.K., Swaminathan, H., & Rogers, H.J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage Publications, Inc.
- Hattie, J. (1985). Methodology review: Assessing unidimensionality of tests and items. *Applied Psychological Bulletin*, 9, 139-164.
- Holland, P.W. & Wainer, H. (1993). *Differential item functioning*. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.
- Jaeger, R. M. (1973). The national test equating study in reading: The anchor test study. *Measurement in Education*, 4, 1-8.
- Kentucky Department of Education (2012). Unbridled learning will move Kentucky forward. Retrieved from [http://www.maysville-online.com/news/opinion/editorial/unbridled-learning-will-move-kentucky-forward/article\\_a6b40cc1-5906-5af4-8d74-7e7f65a7613d.html](http://www.maysville-online.com/news/opinion/editorial/unbridled-learning-will-move-kentucky-forward/article_a6b40cc1-5906-5af4-8d74-7e7f65a7613d.html)
- Kirsch, I. S., Jungeblut, A., Jenkins, L., & Kolstad, A. (1993). *Adult literacy in America: A first look at the results of the National Adult Literacy Survey*. Washington, DC: National Center for Education Statistics, U.S. Department of Education.
- Kirsch, I. S., Jungeblut, A., & Mosenthal, P.B. (1994). *Moving toward the measurement of adult literacy*. Paper presented at the March NCES Meeting, Washington, DC.
- Kolen, M.J. & Brennan, R.L. (2004). *Test equating, scaling, and linking: Methods and practices*. (2<sup>nd</sup> ed.) New York: Springer Science + Business Media, LLC.
- Linacre, J.M. (2011). WINSTEPS (Version 3.73) [Computer Program]. Chicago: Author.
- Lord, F.M. (1980). *Applications of item response theory to practical testing problems*. New York: Erlbaum Associates.
- Loret, P. G., Seder, A., Bianchini, J. C., & Vale, C. A. (1974). *Anchor test study final report: Project report* (Vols. 1-30). Berkeley, CA: Educational Testing Service. (ERIC Document Nos. ED 092 601-ED 092 631).

Messick, S. (1993). Validity. In R.L. Linn (Ed.), *Educational Measurement* (3rd ed., pp. 13–104). New York: American Council on Education and Macmillan Publishing Company.

MetaMetrics, Inc. (October 29, 1998). *Linking*. Unpublished manuscript. Durham, NC: Author.

MetaMetrics, Inc. (2011). *Linking the KCCT in Mathematics with the Quantile Framework*. Durham, NC: Author.

MetaMetrics, Inc. (2005). *PA Series mathematics technical manual*. Durham, NC: Author.

MetaMetrics, Inc. (2008). *The Quantile® Framework for Mathematics norms* [Unpublished normative data]. Durham, NC: Author.

National Council of Teachers of Mathematics (NCTM). (2000). *Principles and standards for school mathematics*. Reston, VA: Author.

National Governors Association Center for Best Practices (NGA Center) & the Council of Chief State School Officers (CCSSO). (2010a). *Common core state standards for mathematics*. Retrieved from [http://www.corestandards.org/assets/CCSSI\\_Math%20Standards.pdf](http://www.corestandards.org/assets/CCSSI_Math%20Standards.pdf).

National Governors Association Center for Best Practices (NGA Center) & the Council of Chief State School Officers (CCSSO). (2010b). *Common core state standards for mathematics: Appendix A*. Retrieved from [http://www.corestandards.org/assets/CCSSI\\_Mathematics\\_Appendix\\_A.pdf](http://www.corestandards.org/assets/CCSSI_Mathematics_Appendix_A.pdf).

National Research Council. (1999). *Uncommon measures: Equivalence and linkage among educational tests*. Washington, D.C.: National Academy Press.

National Research Council. (2001). *Adding it up: Helping children learn mathematics*. J. Kilpatrick, J. Swafford, and B. Findell (Eds.). Mathematics Learning Study Committee, Center for Education, Division of Behavioral and Social Sciences and Education. Washington, DC: National Academy Press.

National Research Council. (2002). *Helping children learn mathematics*. Mathematics Learning Study Committee, J. Kilpatrick and J. Swafford, Editors. Center for Education, Division of Behavioral and Social Sciences and Education. Washington, DC: National Academy Press.

- Petersen, N.S., Kolen, M.J., & Hoover, H.D. (1989). Scaling, Norming, and Equating. In R.L. Linn (Ed.), *Educational Measurement* (3<sup>rd</sup> ed. pp. 221-262). New York: American Council on Education and Macmillan Publishing Company.
- Poznanski, J.B. (1990). A meta-analytic approach to the estimation of item difficulties. Unpublished doctoral dissertation, Duke University, Durham, NC.
- Price, L. R., Lurie, A., & Wilkins, C. (2001). EQUIPERCENT: A SAS program for calculating equivalent scores using the equipercentile method. *Applied Psychological Measurement*, 25, 332-341.
- Rasch, G. (1980). *Probabilistic models for some intelligence and attachment tests*. Chicago: The University of Chicago Press. (First published in 1960).
- Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal of Educational Statistics*, 4, 207-230.
- Reder, S. (1998). Dimensionality and construct validity of the NALS assessment. In M.C. Smith (Ed.), *Literacy for the twenty-first century: Research, policy, practices and the National Adult Literacy Survey*. Westport, CT: Praeger Publishing.
- Rentz, R. R., & Bashaw, W. L. (1975). *Equating reading tests with the Rasch model* (Vol. 1-2). Athens, GA: University of Georgia, Educational Research Laboratory.
- Rentz, R. R., & Bashaw, W. L. (1977). The National Reference Scale for Reading: An application of the Rasch model. *Journal of Educational Measurement*, 14, 161-179.
- Roussos, L., Schnipke, D., & Pashley, P. (1999). A generalized formula for the Mantel-Haenszel differential item functioning parameter. *Journal of Behavioral and Educational Statistics*, 24, 293-322.
- Salganik, L. H., & Tal, J. (1989). A Review and Reanalysis of the ETS/NAEP Young Adult Literacy Survey. Washington, DC: Pelavin Associates.
- Salvia, J. & Ysseldyke, J.E. (1998). *Assessment* (7<sup>th</sup> ed.). Boston: Houghton Mifflin Company.
- SAS Institute, Inc. (1985). The FREQ procedure. In *SAS Users Guide: Statistics, Version 5 Edition*. Cary, NC: Author.
- Sitter, R.R. (1992). Comparing three bootstrap methods for survey data. *The Canadian Journal of Statistics*, 20(2), 135-154.

- Starr, L. (2002). Math wars! *Education World*. Retrieved January 27, 2003, from [http://www.eduationworld.com/a\\_curr/curr071.shtml](http://www.eduationworld.com/a_curr/curr071.shtml).
- Stenner, A.J. (1990). Objectivity: Specific and general. *Rasch Measurement Transactions*, 4, 111.
- Thurstone, L. L. (1946). Note on a reanalysis of Davis' Reading Tests. *Psychometrika*, 11(2), 185.
- Tomlinson, C.A. (1999). *The differentiated classroom*. Alexandria, VA: Association for Supervision and Curriculum Development.
- Ujifusa, A. (2012, November 7). Ky. Road-Tests Common Core. *Ed Week*, 32(11), 1, 20.
- Wright, B.D. & Linacre, J.M. (1994, August). *The Rasch model as a foundation for the Quantile Framework*. Unpublished manuscript.
- Wright, B.D. & Stone, M.H. (1979). *Best Test Design*. Chicago: MESA Press.
- Zwick, R. (1987). Assessing the Dimensionality of the NAEP Reading Data. *Journal of Educational Measurement*, 24, 293–308.

# Appendix A

The Quantile Framework for Mathematics Map..... A-2

	EM*	0Q	100Q	200Q	300Q	400Q	500Q	600Q	700Q	800Q	900Q	1000Q	1100Q	1200Q	1300Q	1400Q
GEOMETRY	<b>EM</b> Use directional and positional words.	<b>40Q</b> Combine simple figures to create a given shape.	<b>160Q</b> Identify and name basic solid figures: rectangular prism, cylinder, pyramid, and cone; identify in the environment.	<b>200Q</b> Identify and name: hexagon, trapezoid, parallelogram and rhombus.	<b>400Q</b> Identify intersecting, parallel, skew, and perpendicular lines and line segments. Identify midpoints of line segments.	<b>500Q</b> Identify angles (acute, right, obtuse, and straight).	<b>680Q</b> Classify plane figures according to type of symmetry (line, rotational).	<b>770Q</b> Identify corresponding parts of similar and congruent figures.	<b>880Q</b> Describe cross-sectional views of three-dimensional figures.	<b>1020Q</b> Define and identify complementary and supplementary angles.	<b>1170Q</b> Use properties of triangles to solve problems related to isosceles and equilateral triangles.	<b>1280Q</b> Use trigonometric ratios to represent relationships in the coordinate plane.	<b>1340Q</b> Describe the transformations of solid figures in space.	<b>1400Q</b> Graph polar equations; identify transformations related to changes in constants and coefficients.		
MEASUREMENT	<b>EM</b> Measure length using nonstandard units.	<b>70Q</b> Determine the value of sets of coins.	<b>100Q</b> Measure lengths in inches/centimeters using appropriate tools and units.	<b>210Q</b> Tell time at the five-minute intervals.	<b>300Q</b> Make different sets of coins with equivalent values.	<b>400Q</b> Determine perimeter using concrete models, nonstandard units, and standard units.	<b>560Q</b> Use grids to develop the relationship between the total numbers of square units in a rectangle and the length and width of the rectangle ( $l \times w$ ).	<b>640Q</b> Calculate distances from scale drawings and maps.	<b>840Q</b> Use models to find volume for prisms and cylinders as the product of the area of the base (B) and the height. Calculate the volume of prisms.	<b>920Q</b> Use proportions to express relationships between corresponding parts of similar figures.	<b>1040Q</b> Use nets or formulas to find the surface area of prisms and cylinders.	<b>1140Q</b> Find the slope of a line given the graph of the line, an equation of the line, or two points on the line.	<b>1230Q</b> Find the ratio of perimeters, areas, and volumes of similar geometric figures using formulas to solve problems.	<b>1360Q</b> Determine the area and volume of figures using right triangle relationships, including trigonometric relationships.	<b>1400Q</b> Determine the magnitude and direction of a vector and solve problems.	
OPERATIONS	<b>EM</b> Read, write, and count using whole numbers; rote count forward to 30.	<b>30Q</b> Use place value with hundreds.	<b>160Q</b> Read and write word names for numbers from 1,000 to 9,999.	<b>210Q</b> Compare and order numbers less than 10,000.	<b>320Q</b> Identify combinations of fractions that make one whole.	<b>410Q</b> Round whole numbers to a given place value.	<b>560Q</b> Use the distributive property to simplify numerical expressions.	<b>600Q</b> Estimate products and quotients of decimals or of mixed numbers.	<b>720Q</b> Read, write, or model numbers in expanded form using exponents.	<b>830Q</b> Calculate unit rates to make comparisons.	<b>900Q</b> Determine the absolute value of a number.	<b>1000Q</b> Calculate using numbers expressed in scientific notation.	<b>1130Q</b> Factor polynomials.	<b>1210Q</b> Use rational exponents to simplify expressions.	<b>1310Q</b> Find sums, differences, products, and quotients of rational algebraic expressions.	<b>1400Q</b> Simplify complex fractions.
<b>The Quantile® Framework for Mathematics</b>																
NUMBERS AND	<b>EM</b> Use ordinal numbers beyond tenth to describe order.	<b>60Q</b> Identify odd and even numbers using objects.	<b>190Q</b> Represent fractions concretely and symbolically.	<b>250Q</b> Subtract 2- and 3-digit numbers with regrouping.	<b>330Q</b> Compare rational numbers in decimal form (tenths and hundredths) with and without models.	<b>450Q</b> Divide using single-digit divisors with and without remainders.	<b>580Q</b> Estimate and compute sums and differences with decimal numbers.	<b>650Q</b> Use powers of ten to multiply and divide whole numbers and decimals.	<b>780Q</b> Write numbers using prime factorization.	<b>890Q</b> Compute with rational numbers (positive and negative).	<b>970Q</b> Add, subtract, and multiply matrices (including scalar multiplication).	<b>1050Q</b> Add, subtract, and multiply polynomials.	<b>1170Q</b> Describe, compare, and simplify imaginary numbers.	<b>1210Q</b> Perform basic operations with complex numbers and graph complex numbers.	<b>1370Q</b> Locate points in a polar coordinate system. Convert between rectangular and polar systems.	<b>1400Q</b> Add and subtract vectors; multiply vectors by a scalar.
ALGEBRA/PATTERNS & FUNCTIONS	<b>EM</b> Describe likenesses and differences between and among objects.	<b>70Q</b> Identify a pattern and translate into another form (e.g., actions, words, objects).	<b>150Q</b> Find the value of an unknown in a number sentence.	<b>300Q</b> Write addition and subtraction sentences to represent a word problem.	<b>430Q</b> Describe the meaning of an unknown in the context of a word problem.	<b>530Q</b> Find the value of a variable in a number sentence.	<b>650Q</b> Use one-step equations and inequalities to model and solve problems.	<b>780Q</b> Identify situations or solve problems with varying rates of change.	<b>810Q</b> Determine the ratio or rate of change of a relation given a table or graph.	<b>990Q</b> Use systems of linear equations in two or more variables to solve problems.	<b>1090Q</b> Find and interpret the maximum, the minimum, and the intercepts of a quadratic function.	<b>1140Q</b> Describe the slope of a line given in the context of a problem situation.	<b>1200Q</b> Graph exponential functions of the form $f(x) = ab^x$ .	<b>1340Q</b> Rename logarithmic expressions using properties of logarithms.	<b>1400Q</b> Use the definition of an ellipse to identify characteristics, write an equation, and graph the relation.	
DATA ANALYSIS & PROBABILITY	<b>EM</b> Organize, display, and interpret information in concrete or picture graphs.	<b>40Q</b> Describe the probability of chance events as certain, impossible, more likely, less likely or equally likely to occur.	<b>200Q</b> Organize, display, and interpret information in line plots and tally charts.	<b>390Q</b> Organize, display, and interpret information in tables and graphs (frequency tables, pictographs, and line plots).	<b>440Q</b> Describe the probability of an event using a fraction or ratio.	<b>600Q</b> Organize, display, and interpret information in circle graphs.	<b>730Q</b> Determine odds given an event or a probability.	<b>850Q</b> Describe data using the mean.	<b>960Q</b> Organize, display, and interpret information in box-and-whisker plots.	<b>1050Q</b> Identify outliers and determine their effect on the mean, median, and range of a set of data.	<b>1100Q</b> Derive a linear equation that models a set of data using calculators. Use the model to make predictions.	<b>1230Q</b> Determine a simple probability using geometric figures.	<b>1460Q</b> Model periodic phenomena using trigonometric functions.			
	EM*	0Q	100Q	200Q	300Q	400Q	500Q	600Q	700Q	800Q	900Q	1000Q	1100Q	1200Q	1300Q	1400Q



The Quantile® Framework for Mathematics uses a common scale to measure both a student's mathematical ability and the difficulty of mathematical tasks. You can match a student's Quantile measure (e.g., 650Q) to the Quantile measure of a mathematical skill to see if the student is ready to learn that skill, needs to learn supporting skills first, or has already mastered it. The Quantile map depicts sample measures from the more than 500 skills taught from kindergarten through high school. The map shows that mathematics is

developmental—readiness to learn a specific skill depends on having learned more basic skills first. It also shows the connections between skills across the content strands (colored bars). Educators and parents use Quantile measures to monitor a student's development in mathematics and determine how best to teach a skill or concept. To get more information, use free resources and search the mathematical skills database, visit [www.Quantiles.com](http://www.Quantiles.com).

\*Emerging Mathematician (EM) represents a Quantile measure of 0Q and below.

The Quantile Framework for Mathematics was developed by MetaMetrics®, an educational measurement and research organization.

A student's Quantile measure is derived from a mathematics assessment, such as a state's year-end test, or classroom program that is linked with the Quantile Framework. For more information, visit [www.Quantiles.com](http://www.Quantiles.com).

MetaMetrics®, the MetaMetrics® logo and tagline, Quantile®, Quantile Framework® and the Quantile® logo are trademarks of MetaMetrics, Inc., and are registered in the United States and abroad. Copyright © MetaMetrics, Inc. All rights reserved.

