Linking the K-PREP Reading with the Lexile® Framework

A Study to Link the Kentucky Performance Rating for Educational Progress Reading with The Lexile® Framework for Reading

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

1000 Park Forty Plaza Drive, Suite 120
Durham, North Carolina 27713
www.MetaMetricsInc.com
www.Lexile.com

Table of Contents

Introduction 1
The Lexile Framework for Reading 3
The Semantic Component 3
The Syntactic Component 4
Calibration of Text Difficulty 5
The Lexile Scale 5
Validity of The Lexile Framework for Reading 7
Text Measure Error Associated with the Lexile Framework 11
Lexile Item Bank 13
The K-PREP Reading - Lexile Framework Linking Process 17
Description of the Assessments 17
Study Design 20
Description of the Sample 20
Linking the K-PREP Reading Scale Scores with the Lexile Scale 26
Validity of the K-PREP Reading–Lexile Link 29
Principal Components Analysis 44
The Lexile Framework and Forecasted Comprehension Rates 47
Conclusions, Caveats, and Recommendations 51
References 61
Appendix A-1

Introduction

Often it is desirable to convey more information about test performance than can be incorporated into a single primary score scale. Two examples arise in large-scale assessment. In one situation, one test can provide a unique type of information (such as national comparisons available from NAEP) but is not administered very often. At the same time another test is administered more often, but is not able to provide the breadth of information (such as a state assessment). An auxiliary score scale for a test can be established to provide this additional information through assessment scale linkages. Once linkages are established between the two assessments, then the results of the more-frequently-administered assessment can be translated in terms of the scale for the other assessment.

In another situation, the linkage between two score scales can be used to provide a context for understanding the results of one of the assessments. For example, sometimes it is hard to explain what a student can read based on the results of a reading comprehension test. Parents typically ask the questions "If my child is in the fourth grade and scores 210 on the Kentucky Performance Rating for Educational Progress (K-PREP) Reading Test, what does this mean?" or "Based on my child's test results, what can he or she read and how well?" or "Is my child well prepared to meet the reading demands of grade level materials?" Once a linkage is established with an assessment that is related to specific book or text titles, then the results of the assessment can be explained and interpreted in the context of the specific titles that a student can read.

Auxiliary score scales can be used to "convey additional normative information, test-content information, and information that is jointly normative and content based. For many test uses, an auxiliary scale conveys information that is more crucial than the information conveyed by the primary score scale. In such instances, the auxiliary score is the one that is focused on, and the primary scale can be viewed more as a vehicle for maintaining interpretability over time" (Petersen, Kolen, and Hoover, 1989, p. 222). One such auxiliary scale is The Lexile® Framework for Reading, which was developed to appropriately match readers with text at a level that provides challenge but not frustration.

Linking assessment results with the Lexile Framework provides a mechanism for matching each student's reading ability with text on a common scale. It serves as an anchor to which texts and assessments can be connected allowing parents, teachers, and administrators to speak the same language. In addition, the Lexile Framework provides a common way to monitor if students are "on track" for the reading demands of various postsecondary endeavors. By using the Lexile Framework, the same metric is applied to

the books students read, the tests they take, and the results that are reported. Parents often ask questions like the following:

- How can I help my child become a better reader?
- How do I challenge my child to read so that she is ready for various college and career options?

Questions like these can be challenging for parents and educators. By linking the K-PREP Reading Test with The Lexile Framework for Reading, educators and parents will be able to answer these questions and will be better able to use the results from the test to improve instruction and to develop each student's level of reading comprehension.

This research study was designed to determine a mechanism to provide reading levels that can be matched to text based on the K-PREP Reading test scores. 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:

- present a solution for matching readers with text;
- provide Kentucky with Lexile measures on the K-PREP Reading Test;
- develop tables for converting K-PREP Reading Test scores to Lexile measures; and
- produce a report that describes the linking analysis procedures.

The Lexile Framework for Reading

All symbol systems share two features: a semantic component and a syntactic component. In language, the semantic units are words. Words are organized according to rules of syntax into thought units and sentences (Carver, 1974). In all cases, the semantic units vary in familiarity and the syntactic structures vary in complexity. The comprehensibility or difficulty of a message is dominated by the familiarity of the semantic units and by the complexity of the syntactic structures used in constructing the message.

The Semantic Component

As far as the semantic component is concerned, it is clear that most operationalizations are proxies for the probability that an individual will encounter a word in a familiar context and thus be able to infer its meaning (Bormuth, 1966). This is the basis of exposure theory, which explains the way receptive or hearing vocabulary develops (Miller and Gildea, 1987; Stenner, Smith, and Burdick, 1983). Klare (1963) hypothesized that the semantic component varied along a familiarity-to-rarity continuum. This concept was further developed by Carroll, Davies, and Richman (1971), whose word-frequency study examined the reoccurrence of words in a five-million-word corpus of running text. Knowing the frequency of words as they are used in written and oral communication provided the best means of inferring the likelihood that a word would be encountered by a reader and thus become a part of that individual's receptive vocabulary.

Variables such as the average number of letters or syllables per word have been observed to be proxies for word frequency. There is a strong negative correlation between the length of words and the frequency of word usage. Polysyllabic words are used less frequently than monosyllabic words, making word length a good proxy for the likelihood that an individual will be exposed to a word.

In a study examining receptive vocabulary, Stenner, Smith, and Burdick (1983) analyzed more than 50 semantic variables in order to identify those elements that contributed to the difficulty of the 350 vocabulary items on Forms L and M of the *Peabody Picture Vocabulary Test*—Revised (Dunn and Dunn, 1981). Variables included part of speech, number of letters, number of syllables, the modal grade at which the word appeared in school materials, content classification of the word, the frequency of the word from two different word counts, and various algebraic transformations of these measures.

The first word frequency measure used was the raw count of how often a given word appeared in a corpus of 5,088,721 words sampled from a broad range of school materials (Carroll, Davies, and Richman, 1971). For example, the word "accident"

appears 176 times in the 5,088,721-word corpus. The second word frequency measure used was the frequency of the "word family." A word family included: (1) the stimulus word; (2) all plurals (adding "-s" or "-es" or changing "-y" to "-ies"); (3) adverbial forms; (4) comparatives and superlatives; (5) verb forms ("-s," "-d," "-ed," and "-ing"); (6) past participles; and (7) adjective forms. For example, the word family for "accident" would include "accidental," "accidentally," "accidentals," and "accidents," and they would all have the same word frequency of 334. The frequency of a word family was based on the sum of the individual word frequencies from each of the types listed.

Correlations were computed between algebraic transformations of these means (mean frequency of the words in the test item and mean frequency of the word families in the test item) and the rank order of the test items. Since the items were ordered according to increasing difficulty, the rank order was used as the observed item difficulty. The log of the mean word frequency provided the strongest correlation with item rank order (r = -0.779) for the items on the combined form.

The Lexile Framework currently employs a 300-million-word corpus when examining the semantic component of text. This corpus was assembled from the more than 15,000 texts that were measured by MetaMetrics for publishers from 1998 through January 2000. When text is analyzed by MetaMetrics, all electronic files are initially edited according to established guidelines used with the Lexile Analyzer software. These guidelines include the removal of all incomplete sentences, chapter titles, and paragraph headings; running of a spell check; and re-punctuating where necessary to correspond to how the book would be read by a child (for example, at the end of a page). The text is then submitted to the Lexile Analyzer that examines the lengths of the sentences and the frequencies of the words and reports a Lexile measure for the book. When enough additional texts have been analyzed to make an adjustment to the corpus necessary and desirable, a linking study will be conducted to adjust the calibration equation such that the Lexile measure of a text based on the current corpus will be equivalent to the Lexile measure based on the new corpus.

The Syntactic Component

Klare (1963) provides a possible interpretation for how sentence length works in predicting passage difficulty. He speculated that the syntactic component varied with the load placed on short-term memory. Crain and Shankweiler (1988), Shankweiler and Crain (1986), and Liberman, Mann, Shankweiler, and Westelman (1982) have also supported this explanation. The work of these individuals has provided evidence that sentence length is a good proxy for the demand that structural complexity places upon verbal short-term memory.

While sentence length has been shown to be a powerful proxy for the syntactic complexity of a passage, an important caveat is that sentence length is not the underlying causal influence (Chall, 1988). Researchers sometimes incorrectly assume that manipulation of sentence length will have a predictable effect on passage difficulty. Davidson and Kantor (1982), for example, illustrated rather clearly that sentence length can be reduced and difficulty increased and vice versa.

Based on previous research, it was decided to use sentence length as a proxy for the syntactic component of reading difficulty in the Lexile Framework.

Calibration of Text Difficulty

The research study on semantic units (Stenner, Smith, and Burdick, 1983) was extended to examine the relationship of word frequency and sentence length to reading comprehension. In 1987(a), Stenner, Smith, Horabin, and Smith performed exploratory regression analyses to test the explanatory power of these variables. This analysis involved calculating the mean word frequency and the log of the mean sentence length for each of the 66 reading comprehension passages on the *Peabody Individual Achievement Test* (Dunn and Markwardt, 1970). The observed difficulty of each passage was the mean difficulty of the items associated with the passage (provided by the publisher) converted to the logit scale. A regression analysis based on the word-frequency and sentence-length measures produced a regression equation that explained most of the variance found in the set of reading comprehension tasks. The resulting correlation between the observed logit difficulties and the theoretical calibrations was 0.97 after correction for range restriction and measurement error. The regression equation was further refined based on its use in predicting the observed difficulty of the reading comprehension passages on 8 other standardized tests. The resulting correlation between the observed logit difficulties and the theoretical calibrations across the 9 tests was 0.93 after correction for range restriction and measurement error.

Once a regression equation is established linking the syntactic and semantic features of text to the difficulty of text, the equation can be used to calibrate test items and text.

The Lexile Scale

In developing the Lexile Scale, the Rasch model (Wright and Stone, 1979) was used to estimate the difficulties of the items and the abilities of the 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 group 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 group. If two different items are administered to the second group, 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 a scale with a fixed zero was to identify two anchor points for the scale. The following criteria were used to select the two anchor points: they should be intuitive, easily reproduced, and widely recognized. For example, with most thermometers the anchor points are the freezing and boiling points of water. For the Lexile Scale, the anchor points are text from seven basal primers for the low end and text from *The Electronic Encyclopedia* (Grolier, Inc., 1986) for the high end. These points correspond to the middle of first grade text and the midpoint of workplace text.

The next step was to determine the unit size for the scale. For the Celsius thermometer, the unit size (a degree) is $1/100^{th}$ of the difference between freezing (0 degrees) and boiling (100 degrees) water. For the Lexile Scale the unit size (a Lexile) was defined as $1/1000^{th}$ of the difference between the mean difficulty of the primer material and the mean difficulty of the encyclopedia samples. Therefore, a Lexile by definition equals $1/1000^{th}$ of the difference between the difficulty of the primers and the difficulty of the encyclopedia.

The third step was to assign a value to the lower anchor point. The low-end anchor on the Lexile Scale was assigned a value of 200.

Finally, a linear equation of the form

$$[(\text{Logit} + \text{constant}) \times \text{CF}] + 200 = \text{Lexile text measure} \qquad \text{Equation (1)}$$

was developed to convert logit difficulties to Lexile calibrations. The values of the conversion factor (CF) and the constant were determined by substituting in the low-end anchor point and then solving the system of equations.

The Lexile Scale ranges from below zero to above 2000L. There is a not an explicit bottom or top to the scale, but rather two anchor points on the scale (described above) that describe different levels of reading comprehension. The Lexile Map, a graphic representation of the Lexile Scale from 200L to 1600L, provides a context for understanding reading comprehension.

Validity of The Lexile Framework for Reading

Validity refers to the "degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests" (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 1999). In other words, does the test measure what it is supposed to measure? For the Lexile Framework, which measures a skill, the most important aspect of validity that should be examined is construct validity. The validity of the Lexile Framework can be evaluated by examining how well Lexile measures relate to other measures of reading comprehension and text difficulty.

Lexile Framework and other Measures of Reading Comprehension. *Table 1* presents the results from studies where students were administered a Lexile assessment and another assessment of reading comprehension. There is a strong relationship between reading comprehension ability as measured by the Lexile Framework and reading comprehension ability as measured by other assessments.

Table 1. Results from linking studies conducted with The Lexile Framework for Reading.

Standardized Test	Grades in Study	N	Correlation Between Test Score and Lexile Measure
Stanford Achievement Tests (Ninth Edition)	4, 6, 8, 10	1,167	0.92
Stanford Diagnostic Reading Test (Version 4)	4, 6, 8, 10	1,169	0.91
North Carolina End-of-Grade Test of Reading Comprehension (NCEOG)	3, 4, 5, 8	956	0.90
TerraNova Assessment Series (CTBS/5)	2, 4, 6, 8	2,713	0.92
Texas Assessment of Academic Skills (TAAS)	3 through 8	3,623	0.73 to 0.78*
Gates-MacGinitie Reading Test	2, 4, 6, 8, 10	4,644	0.90
Metropolitan Achievement Test (Eighth Edition)	2, 4, 6, 8, 10	2,382	0.93
Texas Assessment of Knowledge and Skills (TAKS)	3, 5, 8	1,960	0.60 to 0.73*
The Iowa Tests (Iowa Tests of Basic Skills and Iowa Tests of Educational Development)	3, 5, 7, 9, and 11	4,666	0.88
Stanford Achievement Test (Tenth Edition)	2, 4, 6, 8, and 10	3,064	0.93
Oregon Reading/Literature Knowledge and Skills Test	3, 5, 8, and 10	3,180	0.89
Mississippi Curriculum Test	2, 4, 6, and 8	7,045	0.90
Georgia Criterion Referenced Competency Test (CRCT and GHSGT)	1 – 8, and 11	16,363	0.72 to 0.88*
Wyoming Performance Assessment for Wyoming Students (PAWS)	3, 5, 7, and 11	3,871	0.91
Arizona Instrument to Measure Progress (AIMS)	3, 5, 7, and 10	7,735	0.89
South Carolina Palmetto Achievement Challenge Tests (PACT)	3 – 8	15,559	0.87 to 0.88*

Notes: Results are based on final samples used with each linking study.
*TAAS, TAKS, CRCT/GHSGT, and PACT were not vertically equated; separate linking equations were derived for each grade.

Lexile Framework and the Difficulty of Basal Readers. In a study conducted by Stenner, Smith, Horabin, and Smith (1987b) Lexile calibrations were obtained for units in 11 basal series. It was presumed that each basal series was sequenced by difficulty. So, for example, the latter portion of a third-grade reader is presumably more difficult than the first portion of the same book. Likewise, a fourth-grade reader is presumed to be more difficult than a third-grade reader. Observed difficulties for each unit in a basal series were estimated by the rank order of the unit in the series. Thus, the first unit in the first book of the first grade was assigned a rank order of one and the last unit of the eighth-grade reader was assigned the highest rank order number.

Correlations were computed between the rank order and the Lexile calibration of each unit in each series. After correction for range restriction and measurement error, the average disattenuated correlation between the Lexile calibration of text comprehensibility and the rank order of the basal units was 0.995 (see *Table 2*).

Table 2. Correlations between theory-based calibrations produced by the Lexile equation and rank order of unit in basal readers.

Basal Series	Number of Units	r_{OT}	R_{OT}	R'_{OT}
Ginn Rainbow Series (1985)	53	.93	.98	1.00
HBJ Eagle Series (1983)	70	.93	.98	1.00
Scott Foresman Focus Series (1985)	92	.84	.99	1.00
Riverside Reading Series (1986)	67	.87	.97	1.00
Houghton-Mifflin Reading Series (1983)	33	.88	.96	.99
Economy Reading Series (1986)	67	.86	.96	.99
Scott Foresman American Tradition (1987)	88	.85	.97	.99
HBJ Odyssey Series (1986)	38	.79	.97	.99
Holt Basic Reading Series (1986)	54	.87	.96	.98
Houghton-Mifflin Reading Series (1986)	46	.81	.95	.98
Open Court Headway Program (1985)	52	.54	.94	.97
Total/Means*	660	.839	.965	.995

r_{OT} = raw correlation between observed difficulties (O) and theory-based calibrations (*T*).

R_{OT} = correlation between observed difficulties (O) and theory-based calibrations (*T*) corrected for range restriction.

R'_{OT} = correlation between observed difficulties (O) and theory-based calibrations (*T*) corrected for range restriction and measurement error.

*Mean correlations are the weighted averages of the respective correlations.

Based on the consistency of the results in *Table 2*, the Lexile theory was able to account for the unit rank ordering of the 11 basal series even with numerous differences in the series—prose selections, developmental range addressed, types of prose introduced (i.e., narrative versus expository), and purported skills and objectives emphasized.

Lexile Framework and the Difficulty of Reading Test Items. In a study conducted by Stenner, Smith, Horabin, and Smith (1987a), 1,780 reading comprehension test items appearing on nine nationally-normed tests were analyzed. The study correlated empirical item difficulties provided by the publishers with the Lexile calibrations specified by the computer analysis of the text of each item. The empirical difficulties were obtained in one of three ways. Three of the tests included observed logit difficulties from either a Rasch or three-parameter analysis (e.g., NAEP). For four of the tests, logit difficulties were estimated from item *p*-values and raw score means and standard deviations (Poznanski, 1990; Wright, and Linacre, 1994). Two of the tests provided no item parameters, but in each case items were ordered on the test in terms of difficulty (e.g., PIAT). For these two tests, the empirical difficulties were approximated by the difficulty rank order of the items. In those cases where multiple questions were asked about a single passage, empirical item difficulties were averaged to yield a single observed difficulty for the passage.

Once theory-specified calibrations and empirical item difficulties were computed, the two arrays were correlated and plotted separately for each test. The plots were checked for unusual residual distributions and curvature, and it was discovered that the Lexile equation did not fit poetry items or noncontinuous prose items (e.g., recipes, menus, or shopping lists). This indicated that the universe to which the Lexile equation could be generalized was limited to continuous prose. The poetry and noncontinuous prose items were removed and correlations were recalculated. *Table 3* contains the results of this analysis.

Table 3. Correlations between theory-based calibrations produced by the Lexile equation and empirical item difficulties.

Test	Number of Questions	Number of Passages	Mean	SD	Range	Min	Max	r_{OT}	R_{OT}	R'_{OT}
SRA	235	46	644	353	1303	33	1336	.95	.97	1.00
CAT-E	418	74	789	258	1339	212	1551	.91	.95	.98
Lexile	262	262	771	463	1910	–304	1606	.93	.95	.97
PIAT	66	66	939	451	1515	242	1757	.93	.94	.97
CAT-C	253	43	744	238	810	314	1124	.83	.93	.96
CTBS	246	50	703	271	1133	173	1306	.74	.92	.95
NAEP	189	70	833	263	1162	169	1331	.65	.92	.94
Battery	26	26	491	560	2186	–702	1484	.88	.84	.87
Mastery	85	85	593	488	2135	–586	1549	.74	.75	.77
Total/ Mean	1780	722	767	343	1441	50	1491	.84	.91	.93

r_{OT} = raw correlation between observed difficulties (O) and theory-based calibrations (*T*).
R_{OT} = correlation between observed difficulties (O) and theory-based calibrations (*T*) corrected for range restriction.
R'_{OT} = correlation between observed difficulties (O) and theory-based calibrations (*T*) corrected for range restriction and measurement error.
*Means are computed on Fisher Ƶ transformed correlations.

The last three columns in *Table 3* show the raw correlation between observed (O) item difficulties and theoretical (T) item calibrations, with the correlations corrected for restriction in range and measurement error. The Fisher Ƶ mean of the raw correlations (r_{OT}) is 0.84. When corrections are made for range restriction and measurement error, the Fisher Ƶ mean disattenuated correlation between theory-based calibration and empirical difficulty in an unrestricted group of reading comprehension items (R'_{OT}) is 0.93. These results show that most attempts to measure reading comprehension, no matter what the item form, type of skill objectives assessed, or response requirement used, measure a common comprehension factor specified by the Lexile theory.

Text Measure Error Associated with the Lexile Framework

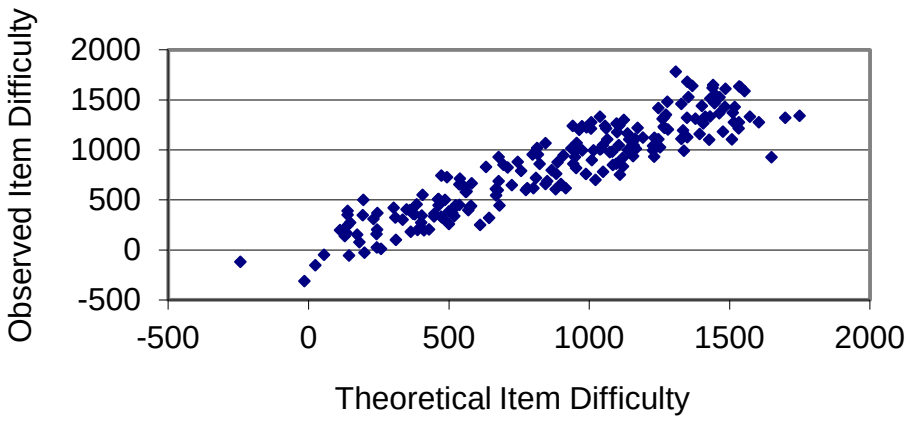
To determine a Lexile measure for a text, the standard procedure is to process the entire text. All pages in the work are concatenated into an electronic file that is processed by a software package called the Lexile Analyzer (developed by MetaMetrics, Inc.). The analyzer "slices" the text file into as many 125-word passages as possible, analyzes the set of slices, and then calibrates each slice in terms of the logit metric. That set of calibrations is then processed to determine the Lexile measure corresponding to a 75% comprehension rate. The analyzer uses the slice calibrations as test item calibrations and

then solves for the measure corresponding to a raw score of 75% (e.g., 30 out of 40 correct, as if the slices were test items). The Lexile Analyzer automates this process, but what "certainty" can be attached to each text measure?

Using the bootstrap procedure to examine error due to the text samples, the above analysis could be repeated (Efron, 1981; Sitter, 1992). The result would be an identical text measure to the first because there is no sampling error when a complete text is calibrated.

There is, however, another source of error that increases the uncertainty about where a text is located on the Lexile Map. The Lexile Theory is imperfect in its calibration of the difficulty of individual text slices. To examine this source of error, 200 items that had been previously calibrated and shown to fit the model were administered to 3,026 students in Grades 2 through 12 in a large urban school district. For each item the observed item difficulty calibrated from the Rasch model was compared with the theoretical item difficulty calibrated from the regression equation used to calibrate texts. A scatter plot of the data is presented in *Figure 1*.

Figure 1. Scatter plot between observed item difficulty and theoretical item difficulty.



The correlation between the observed and the theoretical calibrations for the 200 items was 0.92 and the root mean square error was 178L. Therefore, for an individual slice of text the measurement error is 178L.

The standard error of measurement associated with a text is a function of the error associated with one slice of text (178L) and the number of slices that are calibrated from a text. Very short books have larger uncertainties than longer books. A book with only

four slices would have an uncertainty of 89L whereas a longer book such as *War and Peace* (4,082 slices of text) would only have an uncertainty of 3L (*Table 4*).

Table 4. Standard errors for selected values of the length of the text.

Title	Number of Slices	Text Measure	Standard Error of Text
The Stories Julian Tells	46	520	26
Bunnicula	102	710	18
The Pizza Mystery	137	620	15
Meditations of First Philosophy	206	1720	12
Metaphysics of Morals	209	1620	12
Adventures of Pinocchio	294	780	10
Red Badge of Courage	348	900	10
Scarlet Letter	597	1420	7
Pride and Prejudice	904	1100	6
Decameron	2431	1510	4
War and Peace	4082	1200	3

A typical Grade 3 reading test has approximately 2,000 words in the passages. To calibrate this text, it would be sliced into 16 125-word passages. The error associated with this text measure would be 45L. A typical Grade 7 reading test has approximately 3,000 words in the passages and the error associated with the text measure would be 36L. A typical Grade 10 reading test has approximately 4,000 words in the passages and the error associated with the text measure would be 30L.

The Find A Book (www.Lexile.com) contains information about each book analyzed: author, Lexile measure and Lexile Code, awards, ISBN, and developmental level as determined by the publisher. Information concerning the length of a book and the extent of illustrations—factors that affect a reader's perception of the difficultly of a book—can be obtained from MetaMetrics.

Lexile Item Bank

The Lexile Item Bank contains over 10,000 items that have been developed since 1986 for research purposes with the Lexile Framework.

Passage Selection. Passages selected for use are selected from "real world" reading materials that students may encounter both in and out of the classroom. Sources include textbooks, literature, and periodicals from a variety of interest areas and material written by authors of different backgrounds. The following criteria are used to select passages:

- the passage must develop one main idea or contain one complete piece of information;
- understanding of the passage is independent of the information that comes before or after the passage in the source text; and
- understanding of the passage is independent of prior knowledge not contained in the passage.

With the aid of a computer program, item writers examine blocks of text (minimum of three sentences) that are calibrated to be within 100L of the source text. From these blocks of text item writers are asked to select four to five that could be developed as items. If it is necessary to shorten or lengthen the passage in order to meet the criteria for passage selection, the item writer can immediately recalibrate the text to ensure that it is still targeted within 100L of the complete text (source targeting).

Item Format. The native Lexile item format is embedded completion. The embedded completion format is similar to the fill-in-the-blank format. When properly written, this format directly assesses the reader's ability to draw inferences and establish logical connections between the ideas in the passage (Haladyna, 1994). The reader is presented with a passage of approximately 30 to 150 words in length. The passages are shorter for beginning readers and longer for more advanced readers. The passage is then response illustrated (a statement is added at the end of the passage with a missing word or phrase followed by four options). From the four presented options, the reader is asked to select the "best" option that completes the statement. With this format, all options are semantically and syntactically appropriate completions of the sentence, but one option is unambiguously the "best" option when considered in the context of the passage.

The statement portion of the embedded completion item can assess a variety of skills related to reading comprehension: paraphrase information in the passage, draw a logical conclusion based on the information in the passage, make an inference, identify a supporting detail, or make a generalization based on the information in the passage. The statement is written to ensure that by reading and comprehending the passage the reader is able to select the correct option. When the embedded completion statement is read by itself, each of the four options is plausible.

Item Writer Training. Item writers are classroom teachers and other educators who have had experience with the everyday reading ability of students at various levels. The use of individuals with these types of experiences helps to ensure that the items are valid measures of reading comprehension. Item writers are provided with training materials concerning the embedded completion item format and guidelines for selecting passages, developing statements, and selecting options. The item writing materials also contain incorrect items that illustrate the criteria used to evaluate items and corrections based on those criteria. The final phase of item writer training is a short practice session with three items.

Item writers are provided vocabulary lists to use during statement and option development. The vocabulary lists were compiled from spelling books one grade level below the level the item would typically be used with. The rationale was that these words should be part of a reader's "working" vocabulary since they had been learned the previous year.

Item writers are also given extensive training related to "sensitivity" issues. Part of the item writing materials address these issues and identify areas to avoid when selecting passages and developing items. The following areas are covered: violence and crime, depressing situations/death, offensive language, drugs/alcohol/tobacco, sex/attraction, race/ethnicity, class, gender, religion, supernatural/magic, parent/family, politics, animals/environment, and brand names/junk food. These materials were developed based on material published by McGraw-Hill (*Guidelines for Bias-Free Publishing,* 1983). This publication discusses the equal treatment of the sexes, fair representation of minority groups, and the fair representation of disabled individuals.

Item Review. All items are subjected to a two-stage review process. First, items are reviewed and edited by an editor according to the 19 criteria identified in the item writing materials and for sensitivity issues. Approximately 25% of the items developed are deleted for various reasons. Where possible items are edited and maintained in the item bank.

Items are then reviewed and edited by a group of specialists that represent various perspectives—test developers, editors, and curriculum specialists. These individuals examine each item for sensitivity issues and for the quality of the response options. During the second stage of the item review process, items are either "approved as presented," "approved with edits," or "deleted." Approximately 10% of the items written are "approved with edits" or "deleted" at this stage. When necessary, item writers receive additional on-going feedback and training.

Item Analyses. As part of the linking studies and research studies conducted by MetaMetrics, items in the Lexile Item Bank are evaluated in terms of difficulty (relationship between logit [observed Lexile measure] and theoretical Lexile measure), internal consistency (point-biserial correlation), and bias (ethnicity and gender where possible). Where necessary, items are deleted from the item bank or revised and recalibrated.

During the spring of 1999, 8 levels of a Lexile assessment were administered in a large urban school district to students in grades 1 through 12. The 8 test levels were administered in grades 1, 2, 3, 4, 5, 6, 7-8, and 9-12 and ranged from 40 to 70 items depending on the grade level. A total of 427 items were administered across the 8 test levels. Each item was answered by at least 9,000 students (the number of students per

level ranged from 9,286 in grade 2 to 19,056 in grades 9-12). The item responses were submitted to a Winsteps IRT analysis. The resulting item difficulties (in logits) were assigned Lexile measures by multiplying by 180 and anchoring each set of items to the mean theoretical difficulty of the items on the form.

The Kentucky Performance Rating for Educational Progress Reading - Lexile 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 Reading Tests are designed to assess students' understanding of the Kentucky Core Academic Standards (KCAS) which are the Common Core State Standards for English Language Arts (ELA). The K-PREP Reading 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.

All parts of the K-PREP Reading Tests consist of reading passages and associated items. The Grades 3 through 5 test items are distributed approximately equally (50%/50%) between literary and informative passages. The Grades 6 through 8 tests have a few more items related to informative passages (55%) than to literary passages (45%). All tests contain approximately equal proportions of items in each of four domains: Key Ideas, Craft and Structure, Integration of Ideas, and Vocabulary and Acquisition.

The first part of each K-PREP Reading Test is the NRT portion which contains 30 items. The other two parts of each test are the CRT portion and contain 33 to 46 items, with up to three short-answer and two extended-response items. One reading passage and its associated items in the third part of each test is a field-test set.

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 Reading 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 Lexile Framework for Reading. The Lexile Framework is a tool that can help teachers, parents, and students locate challenging reading materials. Text complexity (difficulty) and reader ability are measured in the same unit—the Lexile. Text complexity is

determined by examining such characteristics as word frequency and sentence length. Items and text are calibrated using the Rasch model. The typical range of the Lexile Scale is from 200L to 1600L, although actual Lexile measures can range from below zero (BR) to above 2000L (see the discussion on pages 5-6 for more information).

Using multiple-choice items, the Lexile Framework measures reading ability by focusing on skills readers use when studying written materials sampled from various content areas. Each test item consists of a passage that is response-illustrated (a statement is added at the end of the passage with a missing word or phrase followed by four options, or distractors). The skills measured by these items include referring to details in the passage, drawing conclusions, and making comparisons and generalizations. Lexile items do not require prior knowledge of ideas outside of the passage, vocabulary taken out of context, or formal logic.

Linking tests for the K-PREP Reading Tests were developed for administration to students in Grades 3 through 8. Characteristics of the Lexile Linking Tests were as similar to the K-PREP Reading Tests as possible, including the number of operational items per test and difficulty of the items. All items on the Lexile Linking Tests were four-choice multiple-choice items and were previously field-tested items with known statistics.

The K-PREP Reading Test is administered in three parts. Part A is a 30-item norm-referenced test. Approximately 20% of the items from Part A have been linked to the Kentucky state standards. However, the NRT portion of the test will use a different version in the future. To maintain the consistency of the link, no items from Part A were used to create the specifications for the Lexile Linking Tests.

The Lexile Linking Tests contained 40 items on each test for Grades 3 through 8, except the Grade 5 test, which contained 43 items. Parts B and C of the K-PREP Reading Tests contain primarily multiple-choice items. Each test contained 2-4 polytomous open-response items. Because the Lexile Linking Test includes only dichotomous items, the total possible score for operational items on Parts B and C of each grade of the K-PREP Reading Test was computed by using the number of points possible. This process yielded a range of score points from 28 (Grade 3) to 43 (Grade 5). After review, a 40 item (40 point) minimum was established to ensure a consistently reliable and robust Lexile measure (scale score).

The items for the Lexile Linking Test were chosen to optimize the match to the K-PREP Reading Test. The *p*-values associated with the operational items were converted to Lexile measures using a computer program developed by MetaMetrics, Inc. (no date). Each Lexile Linking Test had a mean Lexile measure established through analysis of passage difficulties of the K-PREP Reading Test, normative grade-level means, and the item difficulties for the K-PREP Reading Test for 2012. The following mean targets were

set: Grade 3, 732L; Grade 4, 855L; Grade 5, 954L; Grade 6, 1049L; Grade 7, 1141L; and Grade 8, 1186L.

Evaluation of T-parallel Lexile Linking Tests. After administration, the Lexile Linking Test items were reviewed. Based on the item examination, two items from Grade 5 and one item from each of the Grades 6 and 8 Lexile Linking Tests were removed from further analyses. These items exhibited low point-biserial correlations. While a few 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). The raw score descriptive statistics for the Lexile Linking Tests are presented in *Table 5*.

Table 5. Descriptive statistics from the development of the Lexile Linking Test raw scores.

Grade	*N**	Raw Score Mean (SD)	Minimum Score		Maximum Score	
			Observed	Possible	Observed	Possible
3	1,415	23.31 (8.7)	4	0	40	40
4	1,022	24.50 (8.5)	0	0	40	40
5	1,109	26.44 (8.5)	0	0	41	41
6	953	24.80 (8.3)	0	0	39	39
7	1,023	25.48 (8.1)	6	0	40	40
8	958	25.30 (8.1)	5	0	39	39
Total	6,480					

* N size reflects the removal of 48 students for missing, unusable, or duplicate student IDs.

Selected item statistics for the Lexile Linking Tests are presented in *Table 6*.

Table 6. Item statistics from the administration of the Lexile Linking Test.

Grade	*N* (Persons)	*N* (Items)	Percent Correct Mean (Range)	Point-Biserial Range	Coefficient Alpha
3	1,415	40	58 (31 - 87)	0.14 - 0.54	0.91
4	1,022	40	61 (36 - 85)	0.21 - 0.52	0.97
5	1,109	41	64 (37 - 89)	0.24 - 0.53	0.90
6	953	39	63 (11 - 88)	0.28 - 0.54	0.90
7	1,023	40	64 (35 - 87)	0.08 - 0.51	0.89
8	958	39	65 (38 - 89)	0.23 - 0.52	0.90
Total	6,480				

The Coefficient Alpha correlations for each of the six Lexile Linking Tests, one for each grade, ranged from 0.89 to 0.97. This indicates strong internal consistency reliability for each of the six tests and high consistency across these 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 sets of items to examinees is operationally possible, and (2) differential order effects are not expected to occur" (pp. 16–17). The Lexile Linking Tests were administered between May 1, 2012 and June 6, 2012, within two weeks of the administration of the K-PREP Reading tests.

Description of the 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 83 schools from 56 districts participating in the linking study.

Table 7 presents the number of students tested in the linking study and the percentage of students with complete data (both a K-PREP Reading score and a Lexile Linking Test Lexile measure). A total of 6,480 students (Grades 3 through 8), or 99%, had both test scores. This sample will be referred to as the calibration sample.

Table 7. Number of students sampled and number of students in the calibration sample.

Grade	State *N* Received	Lexile Linking Test *N*	Matched *N*	Matched Percent
3	50,512	1,425	1,415	99.3
4	49,339	1,027	1,022	99.5
5	50,730	1,118	1,109	99.2
6	50,400	970	953	98.2
7	49,405	1,026	1,023	99.7
8	49,222	962	958	99.6
Total	299,608	6,528	6,480	99.3

Table 8 shows, for each grade level, the number of students (*N*) in the state and the proportion the *N*-count represents in the final sample. Of the 299,608 students in the state sample, 299,072 (99.8%) remain 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.

Table 8. Comparison of state sample and final sample and the reason for student removal.

Grade	State *N* Received	Accommodated Students	Exempt Students	Missing Scale Scores	Final *N*	Final Percent
3	50,512	61	26	1	50,424	99.8
4	49,339	53	32	7	49,247	99.8
5	50,730	51	32	4	50,643	99.8
6	50,400	55	37	4	50,304	99.8
7	49,405	37	33	6	49,329	99.8
8	49,222	41	52	4	49,125	99.8
Total	299,608	298	212	26	299,072	99.8

Table 9 presents the demographic characteristics of all students in the K-PREP Reading state sample, the calibration sample, and the final sample of students included in this study. Across the samples, the final sample is virtually identical with the state sample and the n count only differs by 536 students. Therefore, it is not necessary to report the state sample in any further tables.

Table 9. Percentage of students in the K-PREP Reading state matched sample, calibration sample, and final sample for selected demographic characteristics.

Student Characteristic	Category	State Sample (*N* = 299,608)	Calibration Sample (*N* = 6,480)	Final Sample (*N* = 299,072)
Gender	Female	48.9	49.6	48.9
	Male	51.1	50.4	51.1
Ethnicity*	African American	12.4	8.1	12.4
	American Indian	0.9	1.1	0.9
	Asian	1.5	0.8	1.5
	Hispanic	3.6	3.4	3.6
	Pacific Islander	0.2	0.2	0.2
	White	83.2	90.9	83.2
IEP	Yes	11.2	10.8	11.2
	No	88.8	89.2	88.8
LEP	Yes	2.0	2.3	2.0
	No	98.0	97.7	98.0
Migrant	Yes	0.2	0.3	0.2
	No	99.8	99.7	99.8
Homeless	Yes	3.7	2.8	3.7
	No	95.0	97.1	95.0
	Unknown	1.3	0.1	1.3
Lunch Status	Free	51.0	53.2	51.0
	Reduced	7.3	8.1	7.3
	No Lunch/Not Indicated	41.7	38.7	41.7

Student Characteristic	Category	State Sample (*N* = 299,608)	Calibration Sample (*N* = 6,480)	Final Sample (*N* = 299,072)
Accommodation	Yes	0.1	0.2	.
	No	99.9	99.8	100.0
Accommodation-Audio	Yes	0.5	0.3	0.5
	No	99.5	99.7	99.5
Accommodation-Braille	No	100.0	100.0	100.0
Exempt	Yes	0.1	.	.
	No	99.9	100.0	100.0
Grade	3	16.9	21.7	16.9
	4	16.5	15.8	16.5
	5	16.9	17.1	16.9
	6	16.8	14.8	16.8
	7	16.5	15.8	16.5
	8	16.4	14.8	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 Reading Tests with Lexile items anchored to their theoretical values was conducted on the calibration sample to place the K-PREP Reading items onto the Lexile 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 Reading items on the Lexile scale was conducted to obtain calibrated Lexile measures for the students. These calibrated Lexile measures were used in the subsequent linking process.

Table 10 presents the descriptive statistics for the K-PREP Reading scale score calibration sample as well as the calibrated sample Lexile Linking Test Lexile measure. Evaluating the Lexile measures on the K-PREP Reading Tests and the Lexile Linking Tests show comparable results. The correlations between the calibration sample K-PREP Reading scale scores and the calibration sample Lexile Linking Test measures range from 0.712 and 0.787. Based upon the correlations between the K-PREP Reading

scale scores and the Lexile Linking Test Lexile measures presented in *Table 10*, they are within the typical range of alternate-form reliability coefficients; therefore, the Lexile Linking Tests can be considered a T-parallel form of the K-PREP Reading tests (See Note 1).

Table 10. Descriptive statistics for the calibration sample K-PREP Reading scale scores and Lexile measures and the calibration sample Lexile Linking Test Lexile measures.

Grade	*N*	Calibration Sample K-PREP Reading Scale Score Mean (SD)	Calibration Sample K-PREP Reading Lexile Measure Mean (SD)	Calibration Sample Lexile Linking Test Lexile Measure Mean (SD)	*r*
3	1,415	208.27 (17.3)	620.93 (186.5)	625.80 (228.6)	0.768
4	1,022	207.93 (18.5)	765.48 (198.1)	777.21 (237.9)	0.712
5	1,109	206.50 (17.5)	909.12 (189.4)	916.79 (245.0)	0.787
6	953	206.18 (17.5)	980.30 (187.6)	990.75 (236.8)	0.770
7	1,023	209.01 (16.1)	1075.19 (176.1)	1086.00 (223.2)	0.774
8	958	207.64 (15.8)	1125.08 (170.3)	1137.00 (220.9)	0.782
Total	6,480				

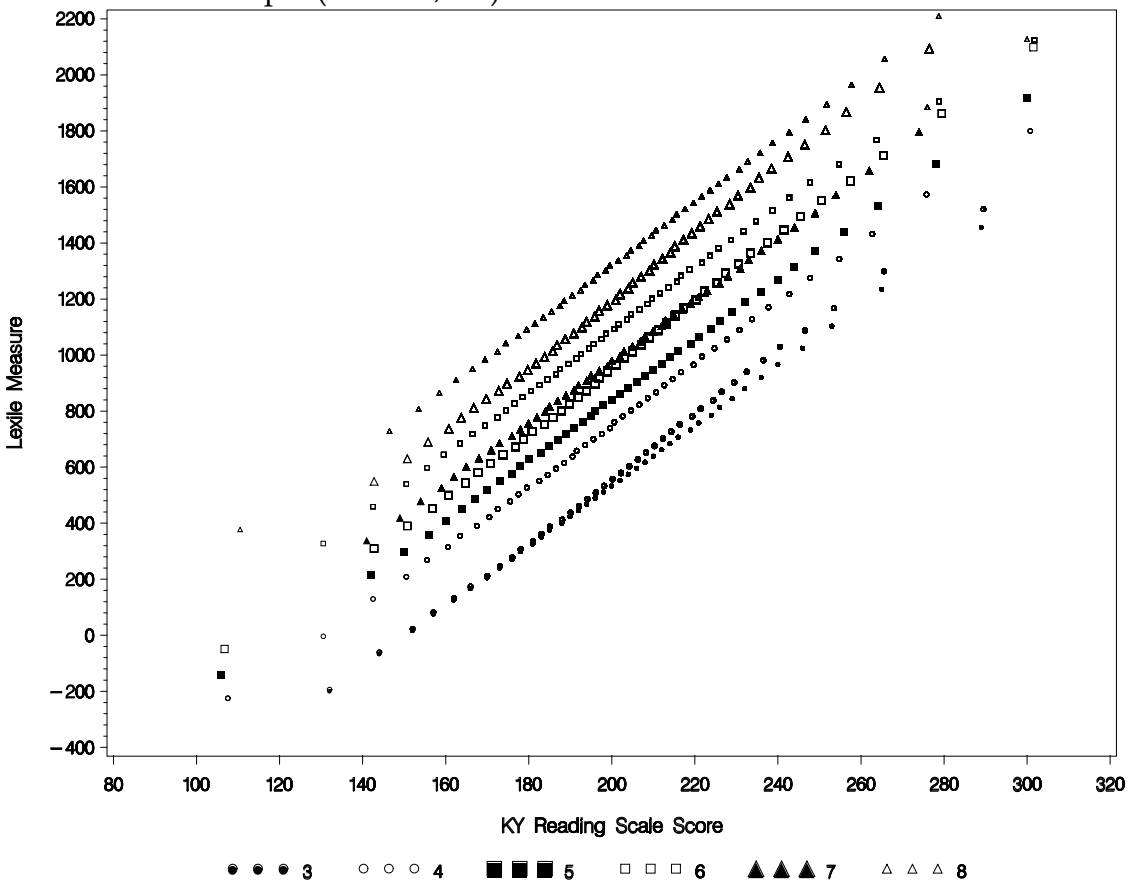
Table 11 presents the descriptive statistics of the final sample K-PREP Reading Test scale scores as well as the final sample Lexile Linking Test Lexile measures. The correlations between the two scores are perfect since the correlations are based on the same students taking a single set of items which are calibrated to two scales.

Table 11. Descriptive statistics for the final sample K-PREP Reading scale scores and the final sample Lexile Linking Test Lexile measures.

Grade	*N*	Final Sample K-PREP Reading Scale Score Mean (SD)	Final Sample Lexile Measure Mean (SD)	*r*
3	50,424	208.96 (17.3)	628.38 (186.2)	1.000
4	49,247	208.12 (17.4)	767.52 (185.5)	1.000
5	50,643	207.57 (17.5)	920.69 (188.4)	1.000
6	50,304	207.98 (18.1)	999.60 (193.8)	1.000
7	49,329	208.64 (16.7)	1071.20 (182.0)	1.000
8	49,125	208.21 (16.4)	1131.23 (177.1)	1.000
Total	299,072			

Figure 2 shows the relationship between the K-PREP Reading scale scores and the Lexile Linking Test Lexile 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 Reading scale score and the final sample Lexile measure reinforcing the use of linear equating.

Figure 2. Scatter plot of the K-PREP Reading scale scores and the Lexile measures for the final sample (*N* = 299,072)



Linking the K-PREP Reading Scale Scores with the Lexile Scale

Linking in general means "putting the scores from two or more tests on the same scale" (National Research Council, 1999, p.15). MetaMetrics and the Kentucky Department of Education conducted this linking study for the purpose of matching students with books and texts—to predict the books and texts a student should be matched with for successful reading experiences, given their performance on the K-PREP Reading.

Linking Analyses. Two score scales (e.g., the K-PREP Reading scale and the Lexile Scale) can be linked using linear equating when (1) test forms have similar difficulties; and (2) 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 sets of items 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,

Kolen, and Hoover, 1989; Kolen and Brennan, 2004). Given scores x and y on tests X and Y, the linear relationship is

$$\frac{(x-\mu_X)}{\sigma_X} = \frac{(y-\mu_y)}{\sigma_y} \qquad \text{(Equation 2)}$$

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

$$x = l_x(y) = \left(\frac{\sigma_X}{\sigma_y}\right)y + \left(\mu_x - \frac{\mu_y \sigma_X}{\sigma_y}\right) \qquad \text{(Equation 3)}$$

Linear equating by definition has the same mean and standard deviation for the overall equation when the scale is vertically aligned. The means and standard deviations are the same for the Linking test and the Target test when calculated across grades. The values are somewhat different when the formula is developed by grade. 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 K-PREP Reading scale scores and Lexile measures can be written as:

Lexile measure = Slope (K-PREP Reading scale score) + constant (Equation 4)

where the slope is the ratio of the standard deviations of the K-PREP Reading scale scores and Lexile Linking Test Lexile measures. These values can be found in *Table 11*.

Using the final sample data described in *Table 11* the linear linking functions relating the K-PREP Reading Tests scale scores and Lexile measures for students in the final sample are presented in *Table 12*. Separate linking functions were developed for each grade of the K-PREP Reading Test because a vertical scale is not employed.

Table 12. Linear linking equation coefficients used to predict Lexile measures from the K-PREP Reading Test scale scores.

Grade	Slope	Intercept
3	[redacted]	[redacted]
4	[redacted]	[redacted]
5	[redacted]	[redacted]
6	[redacted]	[redacted]
7	[redacted]	[redacted]
8	[redacted]	[redacted]

Conversion tables were developed for all grade levels in order to express the K-PREP Reading scale scores in the Lexile metric and were delivered to the Kentucky Department of Education in electronic format (see Appendix). *Table 13* contains the capped Lexile 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 accountability (at the student, school, or district level), then actual measures should be reported at all score points. If the purpose is instructional, then the scores should be capped at the upper bound of measurement error (e.g., at the 95th percentile point). In an instructional environment where the purpose of the Lexile measure is to appropriately match readers with books, no student should receive a negative Lexile measure. Measures of 0L or below are reported as "BR" for "Beginning Reader."

Table 13. Capped values of the Lexile measure by grade.

Grade	Capped Lexile Measure
3	[redacted]
4	[redacted]
5	[redacted]
6	[redacted]
7	[redacted]
8	[redacted]

Validity of the K-PREP Reading–Lexile Link

Table 14 contains the percentile ranks of the Lexile Linking Test Lexile measures and the K-PREP Reading Test Lexile measures based on the final sample. The criterion of a half standard deviation (100L) on the Lexile scale was used to determine the size of the difference. In examining the values, the measures are very similar except at the upper tail of the distributions where few students score. This supports the use of Lexile measures on the K-PREP Reading Test.

Table 14. Comparison of the Lexile measures for selected percentile ranks on the Lexile Linking Test and the final sample K-PREP Reading Test.

Grade 3		
Percentile Rank	Linking Test Lexile Measure	K-PREP Reading Sample Lexile Measure
1	223	209
5	311	328
10	362	403
25	452	510
50	596	618
75	756	758
90	920	876
95	1026	919
99	1238	1102

Grade 4		
Percentile Rank	Linking Test Lexile Measure	K-PREP Reading Sample Lexile Measure
1	337	360
5	452	456
10	500	531
25	589	649
50	758	766
75	907	884
90	1100	1001
95	1242	1076
99	1377	1183

Table 14 (continued). Comparison of the Lexile measures for selected percentile ranks on the Lexile Linking Test and the final sample K-PREP Reading Test.

Grade 5		
Percentile Rank	Linking Test Lexile Measure	K-PREP Reading Sample Lexile Measure
1	446	515
5	533	601
10	607	677
25	742	785
50	923	925
75	1064	1066
90	1231	1152
95	1294	1228
99	1514	1314

Grade 6		
Percentile Rank	Linking Test Lexile Measure	K-PREP Reading Sample Lexile Measure
1	547	603
5	635	689
10	707	753
25	835	850
50	959	989
75	1141	1128
90	1323	1268
95	1404	1311
99	1757	1482

Grade 7		
Percentile Rank	Linking Test Lexile Measure	K-PREP Reading Sample Lexile Measure
1	695	682
5	747	780
10	815	835
25	920	944
50	1064	1064
75	1217	1184
90	1377	1315
95	1519	1370
99	1654	1512

Grade 8		
Percentile Rank	Linking Test Lexile Measure	K-PREP Reading Sample Lexile Measure
1	724	773
5	803	859
10	871	892
25	972	999
50	1118	1129
75	1278	1248
90	1462	1366
95	1543	1431
99	1676	1572

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 Reading Test is aligned to the Common Core State Standards for ELA. This test is

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 Reading Test. *Table 15* presents the performance level cut scores on the K-PREP Reading Test and the associated Lexile measures. There are four performance 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.

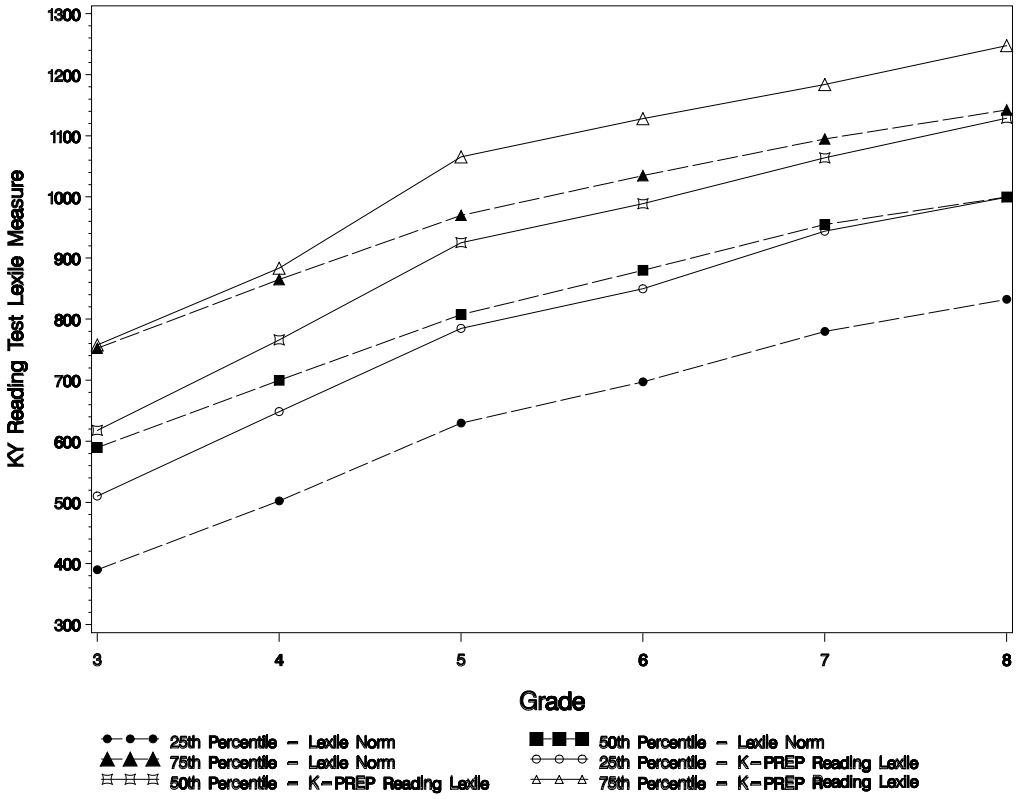
Table 15. K-PREP Reading performance level cut scores on the K-PREP scale and the associated Lexile measures.

Grade	Apprentice		Proficient		Distinguished	
	K-PREP Reading Scale Score	Lexile Measure	K-PREP Reading Scale Score	Lexile Measure	K-PREP Reading Scale Score	Lexile Measure
3	198	[redacted]	210	[redacted]	226	[redacted]
4	197	[redacted]	210	[redacted]	227	[redacted]
5	198	[redacted]	210	[redacted]	226	[redacted]
6	199	[redacted]	210	[redacted]	227	[redacted]
7	199	[redacted]	210	[redacted]	226	[redacted]
8	199	[redacted]	210	[redacted]	225	[redacted]

Figure 3 shows the Lexile measures for the K-PREP Reading Test as compared to the norms that have been developed for use with The Lexile Framework for Reading. These norms were created based on linking studies conducted with the Lexile Framework.

Overall, it can be seen that the K-PREP Reading Lexile measures are higher across the grades at each percentile. The 25th percentile for the K-PREP Reading Lexile measures is closer to the 50th percentile Lexile measures. The 50th percentile for the K-PREP Reading Lexile measures is closer to the 75th percentile Lexile measures except for Grade 3. It appears that the state K-PREP Reading scores were higher than the Lexile norms. The students in Kentucky were more able than the Lexile norms.

Figure 3. Selected Percentiles (25th, 50th, and 75th) plotted for the K-PREP Reading Lexile measure for the final sample (*N* =299,072) against the Lexile measure norms.



The following box and whisker plots (*Figures 4* and *5*) 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 • represents the mean. The end of each whisker represents the minimum and maximum values of the scores (the *y*-axis).

The Lexile measures are on a vertical scale and *Figures 4* and *5* demonstrate this by showing that as the grade increases so do the K-PREP Reading Lexile measures.

Figure 4. Box and whisker plot of the K-PREP Reading Tests Lexile measures by grade, calibration sample (*N* = 6,480).

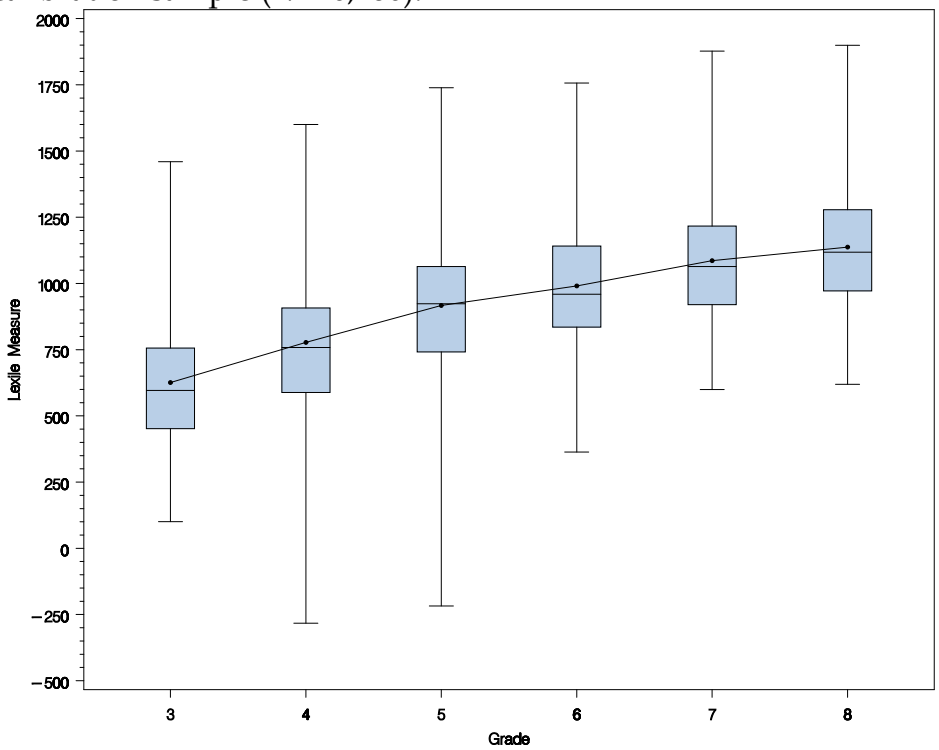
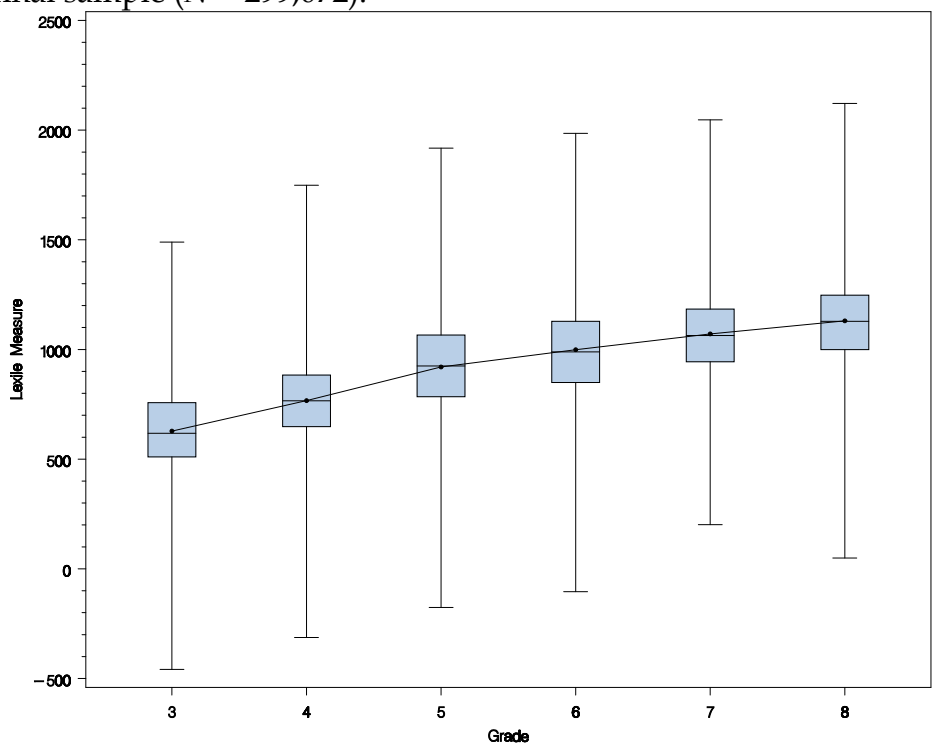


Figure 5. Box and whisker plot of the K-PREP Reading Tests Lexile Measure by grade, final sample (*N* = 299,072).



Principal Components Analysis

In order to further examine the construct being measured by the K-PREP Reading Test and the Lexile Linking Test, principal components analyses were performed. For each grade, the items from the K-PREP Reading Test and Lexile 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 16 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 16. Eigenvalues associated with the first five components in principal components analysis of K-PREP Reading and Lexile Linking items.

Grade	Principal Components				
	1	2	3	4	5
3	23.069387	3.120140	1.967544	1.501931	1.450718
4	21.881496	2.659025	2.189826	1.725157	1.609526
5	24.362863	2.277490	2.032068	1.612609	1.512813
6	23.506638	2.604609	2.149846	1.759853	1.611373
7	22.803816	2.671953	2.254952	1.761257	1.680019
8	22.704378	2.458105	2.243040	1.762081	1.729783

The proportion of variance explained by the first five components is presented in *Table 17*. 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 17. Proportion of variance explained by the first five components in principal components analysis of K-PREP Reading and Lexile Linking items.

Grade	Proportion Explained by Each Component				
	1	2	3	4	5
3	0.295761	0.040002	0.025225	0.019256	0.018599
4	0.276981	0.033659	0.027719	0.021837	0.020374
5	0.304536	0.028469	0.025401	0.020158	0.018910
6	0.283213	0.031381	0.025902	0.021203	0.019414
7	0.268280	0.031435	0.026529	0.020721	0.019765
8	0.270290	0.029263	0.026703	0.020977	0.020593

Table 18 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 18* 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 18. Ratios that assess unidimensionality for principal components analysis.

Grade	Ratio
3	7.393703
4	8.229142
5	10.697241
6	9.025015
7	8.534511
8	9.236539

The scree plots for each grade in *Figure 6* show a strong first factor based on the steep decrease from the first component to the second and subsequent components.

Figure 6. Scree plots for the principal components analysis by grade.

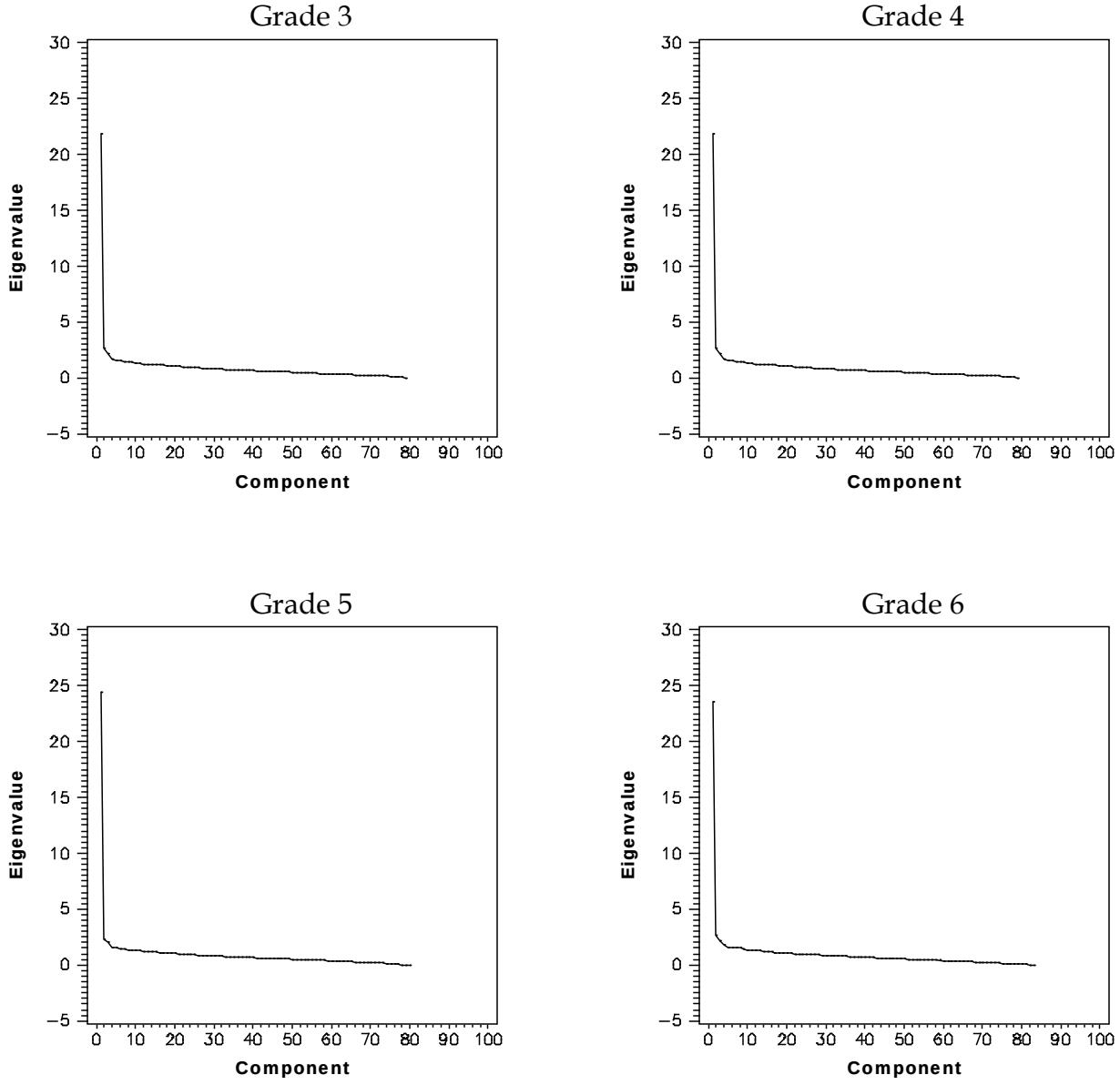
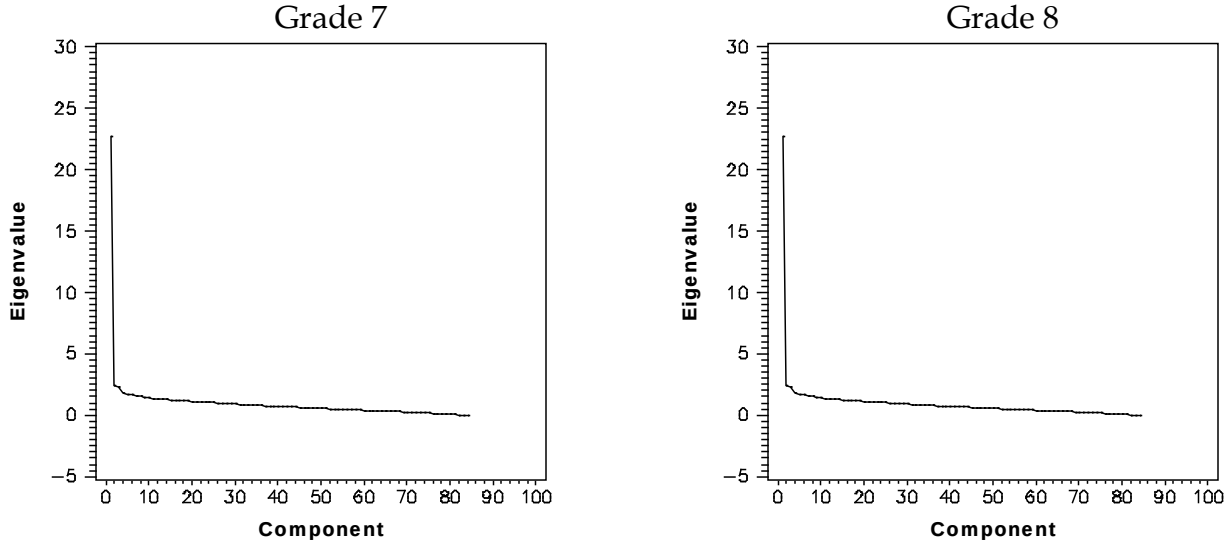


Figure 6 (continued). Scree plots for the principal components analysis by grade.



The Lexile Framework and Forecasted Comprehension Rates

A reader with a measure of 600L who is given a text measured at 600L is expected to have a 75-percent comprehension rate. This 75-percent comprehension rate is the basis for selecting text that is targeted to a reader's reading ability, but what exactly does it mean? And what would the comprehension rate be if this same reader were given a text measured at 350L or one at 850L?

The 75-percent comprehension rate for a reader-text pairing can be given an operational meaning by imagining the text is carved into item-sized slices of approximately 125-140 words with a question embedded in each slice. A reader who answers three-fourths of the questions correctly has a 75-percent comprehension rate.

Suppose instead that the text and reader measures are not the same. It is the difference in Lexile measures between reader and text that governs comprehension. If the text measure is less than the reader measure, the comprehension rate will exceed 75 percent. If not, it will be less. The question is "By how much?" What is the expected comprehension rate when a 600L reader reads a 350L text?

If all the item-sized slices in the 350L text had the same calibration, the 250L difference between the 600L reader and the 350L text could be determined using the Rasch model equation. This equation describes the relationship between the measure of a student's level of reading comprehension and the calibration of the items. Unfortunately, comprehension rates calculated by this procedure would be biased because the calibrations of the slices in ordinary prose are not all the same. The average difficulty level of the slices *and* their variability both affect the comprehension rate.

Although the exact relationship between comprehension rate and the pattern of slice calibrations is complicated, Equation 5 is an unbiased approximation:

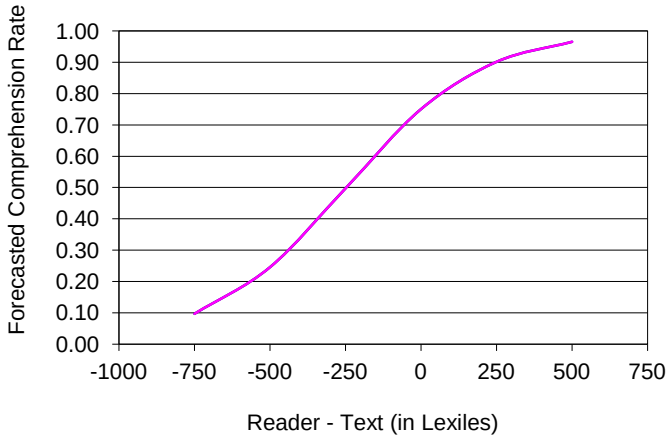
$$Rate = \frac{e^{ELD+1.1}}{1+e^{ELD+1.1}} \qquad \text{(Equation 5)}$$

where ELD is the "effective logit difference" given by

ELD = (Reader Lexile measure – Text Lexile measure) ÷ 225. (Equation 6)

Figure 7 that follows shows the general relationship between reader-text discrepancy and forecasted comprehension rate. When the reader measure and the text calibration are the same (difference of 0L) then the forecasted comprehension rate is 75 percent. In the example in the preceding paragraph, the difference between the reader measure of 600L and the text calibration of 350L is 250L. Referring to *Figure 7* and using +250L (reader minus text), the forecasted comprehension rate for this reader-text combination would be 90 percent.

Figure 7. Relationship between reader-text discrepancy and forecasted comprehension rate.



Tables 19 and *20* show comprehension rates calculated for various combinations of reader measures and text calibrations.

Table 19. Comprehension rates for the same individual with materials of varying comprehension difficulty.

Person Measure	Text Calibration	Sample Titles	Forecast Comprehension
1000	500	*Tornado* (Byars)	96%
1000	750	*The Martian Chronicles* (Bradbury)	90%
1000	1000	*Reader's Digest*	75%
1000	1250	*The Call of the Wild* (London)	50%
1000	1500	*On the Equality Among Mankind* (Rousseau)	25%

Table 20. Comprehension rates of different person abilities with the same material.

Person Measure	Calibration for a Grade 10 Biology Textbook	Forecast Comprehension Rate
500	1000	25%
750	1000	50%
1000	1000	75%
1250	1000	90%
1500	1000	96%

The subjective experience of 50-percent, 75-percent, and 90-percent comprehension as reported by readers varies greatly. A 1000L reader reading 1000L text (75-percent comprehension) reports confidence and competence. Teachers listening to such a reader report that the reader can sustain the meaning thread of the text and can read with motivation and appropriate emotion and emphasis. In short, such readers appear to comprehend what they are reading. A 1000L reader reading 1250L text (50-percent comprehension) encounters so much unfamiliar vocabulary and difficult syntactic structures that the meaning thread is frequently lost. Such readers report frustration and seldom choose to read independently at this level of comprehension. Finally, a 1000L reader reading 750L text (90-percent comprehension) reports total control of the text, reads with speed, and experiences automaticity during the reading process.

The primary utility of the Lexile Framework is its ability to forecast what happens when readers confront text. With every application by teacher, student, librarian, or parent there is a test of the Framework's accuracy. The Framework makes a point prediction every time a text is chosen for a reader. Anecdotal evidence suggests that the Lexile Framework predicts as intended. That is not to say that there is an absence of error in forecasted comprehension. There is error in text measures, reader measures, and their difference modeled as forecasted comprehension. However, the error is sufficiently small that the judgments about readers, texts, and comprehension rates are useful.

Relationship between Linking Error and Forecasted Comprehension Rate. Using Equation 5 with different combinations of reader measure and text difficulty, the effect of linking error on forecasted comprehension rate can be examined. *Table 21* below shows the changes in the forecasted comprehension rate for different combinations of reader and text interactions. When the linking error is small, 5–10L, then the effect on forecasted comprehension rate is a minimal difference (1 to 2 percent) increase or decrease in comprehension.

Table 21. Effect of reader-text discrepancy on forecasted comprehension rate.

Reader Lexile Measure	Text Lexile Measure	Difference	Forecasted Comprehension Rate
1000L	970L	30L	77.4%
1000L	975L	25L	77.0%
1000L	980L	20L	76.7%
1000L	985L	15L	76.3%
1000L	990L	10L	75.8%
1000L	995L	5L	75.4%
1000L	1000L	0L	75.0%
1000L	1005L	¯5L	74.6%
1000L	1010L	¯10L	74.2%
1000L	1015L	¯15L	73.8%
1000L	1020L	¯20L	73.3%
1000L	1025L	¯25L	72.9%
1000L	1030L	¯30L	72.4%

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 can my child actually read?"),
- 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 Reading scale scores and the Lexile measures permits readers to be matched with books and texts that provide an appropriate level of challenge while avoiding frustration. The result of this purposeful match may be that students will read more, and, thereby read better. The real power of the Lexile Framework is in examining the growth of readers—wherever the reader may be in the development of his or her reading skills. Readers can be matched with texts that they are forecasted to read with 75-percent comprehension. As a reader grows, he or she can be matched with more demanding texts. And, as the texts become more demanding, then the reader grows.

Recommendations about reporting Lexile measures for readers. Lexile measures are reported as a number followed by a capital "L" for "Lexile." There is no space between the measure and the "L," and measures of 1,000 or greater are reported without a comma (e.g., 1050L). All Lexile measures should be rounded to the nearest 5L to avoid over interpretation of the measures. As with any test score, uncertainty in the form of measurement error is present.

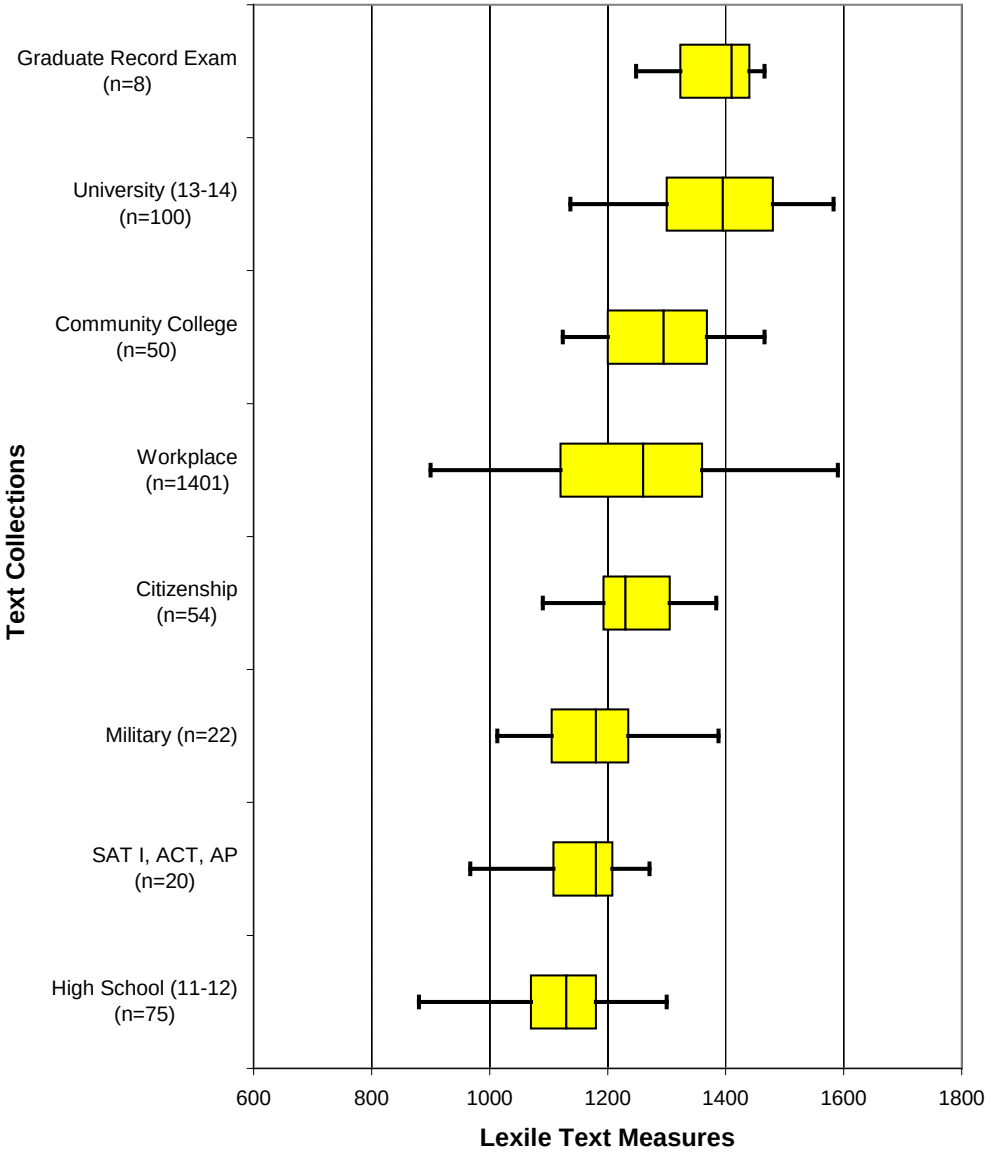
Lexile measures that are reported for an individual student should reflect the purpose for which they will be used. If the purpose is research (e.g., to measure growth at the student, grade, school, district, or state level), then actual measures should be used at all score points, rounded to the nearest integer. A computed Lexile measure of 772.51 would be reflected as 773L. If the purpose is instructional, then the Lexile measures should be capped at the upper bound of measurement error (e.g., at the 95th percentile) to ensure developmental appropriateness of the material. MetaMetrics expresses these as "Reported Lexile Measures" and recommends that these measures be reflected on individual score reports. In instructional environments where the purpose of the Lexile measure is to appropriately match readers with books, all scores at or below 0L should be reported as "BR" (Beginning Reader); no student should receive a negative Lexile measure.

Some assessments report a Lexile range for each student, which is 50L above and 100L below the student's actual Lexile measure. This range represents the boundaries between the easiest kind of reading material for the student and the level at which the student will be more challenged, yet can still read successfully.

Text Complexity. 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.

In the *Journal of Advanced Academics* (Summer 2008), Williamson investigated the gap between high school textbooks and various reading materials across several postsecondary domains. As can be seen in *Figure 8*, the resources Williamson used were organized into four domains that correspond to the three major postsecondary endeavors that students can choose—further education, the workplace, or the military—and the broad area of citizenship, which cuts across all postsecondary endeavors. Williamson discovered a substantial increase in reading expectations and text complexity from high school to postsecondary domains— a gap large enough to help account for high remediation rates and disheartening graduation statistics (Smith, 2011).

Figure 8. A continuum of text difficulty for the transition from high school to postsecondary experiences (box plot percentiles: 5th, 25th, 50th, 75th, and 95th).[1]



Expanding on Williamson's work, Stenner, Sanford-Moore, and Williamson (2012) aggregated the readability information across the various postsecondary options available to a high school graduate to arrive at a standard of reading needed by individuals to be considered "college and career ready." In their study, they included additional citizenship materials beyond those examined by Williamson (e.g., national and international newspapers and other adult reading materials such as Wikipedia articles). Using a weighted mean of the medians for each of the postsecondary options

[1] Reprinted from Williamson, G. L. (2008). A text readability continuum for postsecondary readiness. *Journal of Advanced Academics, 19*(4), 602-632.

(education, military, work place, and citizenship), a measure of 1300L was defined as the general reading demand for postsecondary options and could be used to judge a student's "college and career readiness."

In Texas, two studies were conducted to examine the reading demands in various postsecondary options – technical college, community college, and 4-year university programs. Under Commissioner Raymond Paredes, THECB conducted a research study in 2007 (and extended in 2008) which addressed the focal question of "how well does a student need to read to be successful in community colleges, technical colleges, and universities in Texas?" THECB staff collected a sample of books that first year students in Texas would be required to read in each setting. These books were measured in terms of their text complexity using The Lexile Framework for Reading. Since the TAKS had already been linked with Lexile measures for several years, the THECB study was able to overlay the TAKS cut scores onto the post high school reading requirements. (For a complete description of this report, please visit www.thecb.state.tx.us/index.cfm?objectid=31BFFF6B-BB41-8A43-C76A99EDA0F38B7D.)

Since the THECB study was completed, other states have followed the Texas example and used the same approach in examining the gap from high school to the postsecondary world. In 2009, a similar study was conducted for the Georgia Department of Education; and in 2010, a study was conducted for the Tennessee Department of Education. In terms of mean text demand, the results across the three states produced similar estimates of the reading ability needed in higher-education institutions: Texas, 1230L; Georgia, 1220L; and Tennessee, 1260L. When these results are incorporated with the reading demands of other postsecondary endeavors (military, citizenship, workplace, and adult reading materials [national and international newspapers] and Wikipedia articles) used by Stenner, Koons, and Swartz (2010), the college and career readiness standard for reading is 1293L. These results are based on more than 105,000,000 words from approximately 3,100 sources from the adult text space.

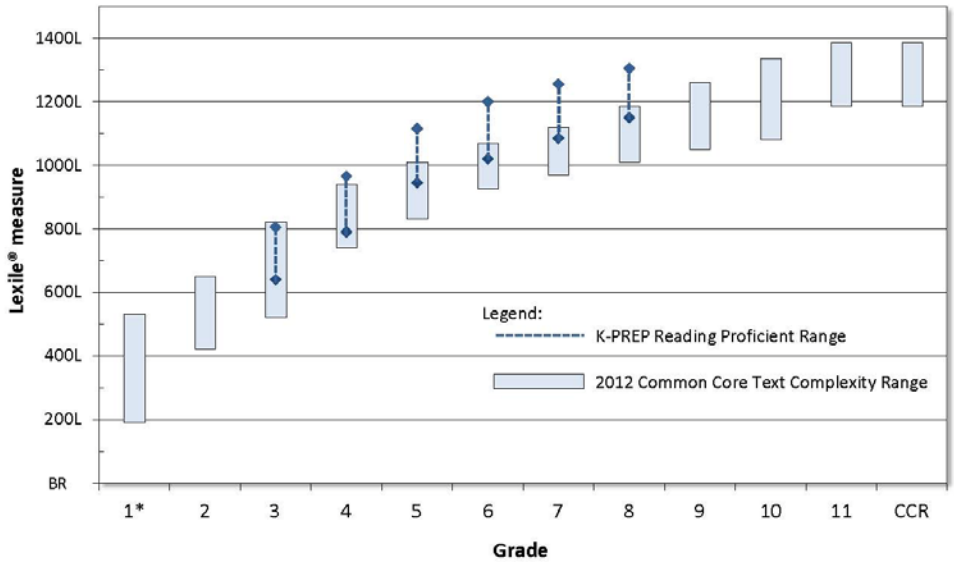
The question for educators becomes how to determine if a student is "on track" for college and career as previously defined in the Common Core State Standards and described above. "As state departments of education, and the districts and schools within those respective states, transition from *adopting* the new Common Core State Standards to the more difficult task of *implementing* them, the challenge now becomes how to translate these higher standards into tangible, practical and cost-effective curricula" (Smith, 2012). Implementing the Common Core will require districts and schools to develop new instructional strategies and complementary resources that are not only aligned with these national college- and career-readiness standards, but also utilize and incorporate proven and cost-effective tools that are universally accessible to all stakeholders.

The Standards for English Language Arts focus on the importance of text complexity. As stated in Standard 10, students must be able to "read and comprehend complex literary and informational texts independently and proficiently" (Common Core State Standards for English Language Arts, College and Career Readiness Anchor Standards for Reading, NGA Center and CCSSO, 2010, p.10).

The Common Core State Standards recommends a three-part model for evaluating the complexity of a text that takes into account its qualitative dimensions, quantitative measure, and reader, and task considerations. It describes text complexity as "the inherent difficulty of reading and comprehending a text combined with consideration of reader and task variables…a three-part assessment of text [complexity] that pairs qualitative and quantitative measures with reader-task considerations" (NGA Center and CCSSO, 2010, p. 43). In simpler terms, *text complexity is a transaction between text, reader, and task*. The quantitative aspect of defining text complexity consists of a stair-step progression of increasingly difficult text by grade levels (see Figure 1) (Common Core State Standards for English Language Arts, Appendix A, NGA Center and CCSSO, 2010, p. 8).

MetaMetrics (Williamson, Koons, Sandvik, and Sanford-Moore, 2012) has conducted research to describe the typical reading demands and develop a text continuum of reading materials across Grades 1-12. This continuum can be "stretched" to describe the reading demands expected of students in Grades 1-12 who are "on track" for college and career (Sanford-Moore and Williamson, 2012). MetaMetrics' research on the typical reading demands of college and careers contributed to the Common Core State Standards as a whole and, more specifically, to the Lexile-based grade bands in *Figure 9*. *Figure 9* shows the relationship between the "Proficient" performance standard for each grade level established on the K-PREP Reading test and the "stretch" reading demands. This shows that the K-PREP Reading performance standards for "Proficient" at each grade level is set at a level that is consistent with being "on track" for college and career readiness at the end of Grade 12.

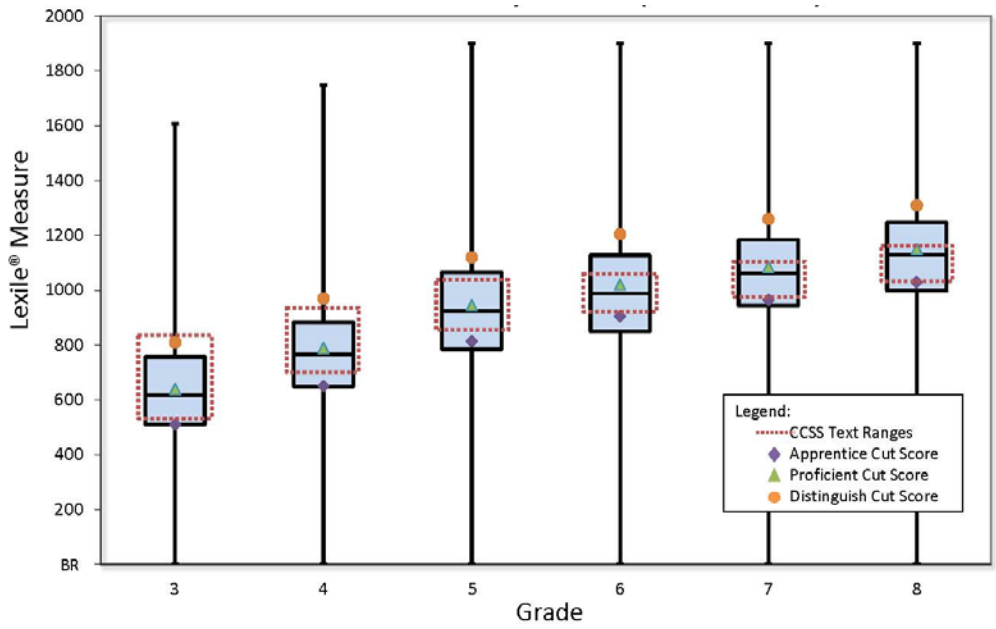
Figure 9. Comparison of K-PREP Reading proficient standards for college and career reading levels described by the CCSS.



*Note: Grade 1 Stretch bands are not published in Appendix A.

Figure 10 shows that the spring 2012 student performance on the K-PREP Reading Test at each grade level is "on track" for college and career readiness. Students can be matched with reading materials that are at or above the recommendations in Appendix A of the CCSS for ELA for each grade level.

Figure 10. K-PREP Reading 2011-2012 student performance expressed as Lexile measures.



In 2010, MetaMetrics and the Kentucky Department of Education (KDE) conducted a study to link the Kentucky Core Content Test in Reading with the Lexile scale (MetaMetrics, 2011). The minimum score considered "proficient" at each grade level on the KCCT in Reading is presented in *Table 22*. In 2012, KDE transitioned their assessment program to the K-PREP Reading Test to align with the Common Core State Standards in English/Language Arts and to describe student reading performance in relation to college and career readiness. One outcome of this change was to set the performance standards for K-PREP 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 Reading Test is also repeated from *Table 15*. The Lexile scale can be used as an external "yardstick" to evaluate this change in reading demand on the Kentucky reading assessment. The information in *Table 22* shows that the K-PREP Reading Test standards are demanding more of students in terms of reading ability.

Table 22. Minimum "proficient" score on KCCT and K-PREP.

Grade	KCCT Proficient Cut Score (2010)	K-PREP Proficient Cut Score (2012)
3	535L	640L
4	630L	790L
5	795L	945L
6	895L	1020L
7	960L	1085L
8	1010L	1150L

Next Steps. To utilize the results from this study, Lexile measures need to be incorporated into the K-PREP Reading results processing and interpretation frameworks. This information can then be used in a variety of areas within the educational system—instruction, assessment, communication to name a few.

Within the *instructional area*, suggested book lists can be developed for ranges of readers. Care must be taken to ensure that the books on the lists are also developmentally appropriate for the readers. The Lexile measure is one factor related to comprehension and is a good starting point in the selection process of a book for a specific reader. Other factors such as student developmental level, motivation, and interest; amount of background knowledge possessed by the reader; and characteristics of the text such as illustrations and formatting also need to be considered when matching a book with a reader.

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 texts and supplementary materials.

The Lexile Framework is an objective tool that can be used to determine a student's readiness for a reading experience; the Lexile Framework "targets" text (books, newspapers, periodicals) for readers at a 75 percent comprehension level—a level that is challenging, but not frustrating (Schnick and Knickelbine, 2000).

Within the *communication* area, Lexile measures can be used to communicate with students, parents, teachers, educators, and the community by providing a common language to use to talk about reading growth and development. By aligning all areas of the educational system, parents can be included in the instructional process. With a variety of data related to a student's reading level a more complete picture can be formed and more informed decisions can be made concerning reading-group placement, amount of extra instruction needed, and promotion/retention decisions.

It is much easier to understand what a national percentile rank of 50 means when it is tied to the reading demands of book titles that are familiar to adults. Parents are encouraged to help their children achieve high standards by expecting their children to succeed at school, communicating with their children's teachers and the school, and helping their children keep pace and do homework.

Through the customized reading lists and electronic database of titles, parents can assist their children in the selection of reading materials that are at the appropriate level of challenge and monitor the reading process at home. A link can be provided to the "Find a Book" website. This site provides a quick, free resource to battle "summer slide" – the learning losses that students often experience during the summer months when they are not in school. Lexile measures make it easy to help students read and learn all summer long and during the school year. This website can help build a reading list of books at a young person's reading level that are about subjects that interest him or her. This website can be viewed at http://www.lexile.com/findabook/.

In one large school district, the end-of-year testing results are sent home to parents in a folder. The folder consists of a Lexile Map on one side and a letter from the superintendent on the other side. The school district considers this type of material as "refrigerator-friendly." They encourage parents to put the Lexile Map on the

refrigerator and use it to monitor and track the reading progress of their child throughout the school year.

The community-at-large (business leaders, citizens, politicians, and visitors) sees the educational system as a reflection of the community. Through the reporting of assessment results (after all, that is what the community is most interested in—results), people can understand what the community values and see the return for its investment in the schools and its children.

One way to involve the community is to work with the public libraries and local bookstores when developing reading lists. The organizations should be contacted early enough so that they can be sure that the books will be available. Often books can be displayed with their Lexile measures for easy access.

Many school districts make presentations to civic groups to educate the community as to their reading initiatives and how the Lexile Framework is being utilized in the school. Conversely, many civic groups are looking for an activity to sponsor, and it could be as simple as "donate-a-book" or "sponsor-a-reader" campaigns.

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, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Bormuth, J.R. (1966). Readability: New approach. *Reading Research Quarterly*, *7*, 79-132.

Carroll, J.B., Davies, P., & Richman, B. (1971). *Word frequency book*. Boston: Houghton Mifflin.

Carver, R.P. (1974). Measuring the primary effect of reading: Reading storage technique, understanding judgments and cloze. *Journal of Reading Behavior*, *6*, 249-274.

Chall, J.S. (1988). "The beginning years." In B.L. Zakaluk and S.J. Samuels (Eds.), *Readability: Its past, present, and future*. Newark, DE: International Reading Association.

Crain, S. & Shankweiler, D. (1988). "Syntactic complexity and reading acquisition." In A. Davidson and G.M. Green (Eds.), *Linguistic complexity and text comprehension: Readability issues reconsidered*. Hillsdale, NJ: Erlbaum Associates.

Davidson, A. & Kantor, R.N. (1982). On the failure of readability formulas to define readable text: A case study from adaptations. *Reading Research Quarterly*, *17*, 187-209.

Dunn, L.M. & Dunn, L.M. (1981). *Manual for Forms L and M of the Peabody Picture Vocabulary Test – Revised*. Circle Pines, MN: American Guidance Service.

Dunn, L.M. & Markwardt, F.C. (1970). *Peabody Individual Achievement Test*. Circle Pines, MN: American Guidance Service.

Efron, B. (1981). Nonparametric estimates of the standard error: The Jackknife, the Bootstrap, and other resampling techniques. *Biometrika*. *68*, 589-599.

Gorsuch, R.L. (1983). *Factor analysis*. Mahwah, NJ: Lawrence Erlbaum Associates Inc.

Grolier, Inc. (1986). *The electronic encyclopedia*. Danbury, CT: Author.

Hattie, J. (1985). Methodology review: Assessing unidimensionality of tests and items. *Applied Psychological Bulletin*. *9*, 139-164.

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

Klare, G.R. (1963). *The measurement of readability*. Ames, IA: Iowa State University Press.

Kolen, M.J. & Brennan, R.L. (2004). *Test equating, scaling, and linking: Methods and practices*. 2nd edition. New York: Springer Science + Business Media, LLC.

Liberman, I.Y., Mann, V.A., Shankweiler, D., & Westelman, M. (1982). Children's memory for recurring linguistic and non-linguistic material in relation to reading ability. *Cortex*, *18*, 367-375.

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.

McGraw-Hill Book Company. (1983). *Guidelines for bias-free publishing*. Monterey, CA: Author.

MetaMetrics, Inc. (2008). The text continuum in 2008. Presented at the Lexile National Conference, San Antonio, TX.

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

Miller, G.A. & Gildea, P.M. (1987). How children learn words. *Scientific American*, *257*, 94-99.

National Governors Association Center for Best Practices & Council of Chief State School Officers. (2010). Common Core State Standards for English Language Arts & Literacy in History/Social Studies, Science, and Technical Subjects, Appendix A. Washington, DC: Author.

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

Petersen, N.S., Kolen, M.J., & Hoover, H.D. (1989). "Scaling, Norming, and Equating." In R.L. Linn (Ed.), *Educational Measurement* (Third Edition) (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.

Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal of Educational Statistics, 4*, 207-230.

Sanford-Moore, E., & Williamson, G. L. (2012). *Bending the text complexity curve to close the gap* (MetaMetrics Research Brief). Durham, NC: MetaMetrics, Inc.

Schnick, T. & Knickelbine, M. (2000). *The Lexile Framework: An introduction for educators*. Durham, NC: MetaMetrics, Inc.

Shankweiler, D. & Crain, S. (1986). Language mechanisms and reading disorder: A modular approach. *Cognition*, *14*, 139-168.

Sitter, R.R. (1992). Comparing three bootstrap methods for survey data. *The Canadian Journal of Statistics*, *20*(2), 135-154.

Smith, M. (2011, March 30). *Bending the reading growth trajectory: Instructional strategies to promote reading skills and close the readiness gap*. MetaMetrics Policy Brief. Durham, NC: MetaMetrics, Inc.

Smith, M. (2012, February). *Not so common: Comparing Lexile® measures with the standards' other text complexity tools*. MetaMetrics White Paper. Durham, NC: MetaMetrics, Inc.

Stenner, A.J. (1990). Objectivity: Specific and general. *Rasch Measurement Transactions*, *4*, 111.

Stenner, A. J., Sanford-Moore, E., & Williamson, G. L. (2012). *The Lexile® Framework for Reading quantifies the reading ability needed for "College & Career Readiness."* MetaMetrics Research Brief. Durham, NC: MetaMetrics, Inc.

Stenner, A.J., Smith, M., & Burdick, D.S. (1983). Toward a theory of construct definition. *Journal of Educational Measurement*, *20*(4), 305-315.

Stenner, A.J., Smith, D.R., Horabin, I., & Smith, M. (1987a). Fit of the Lexile Theory to item difficulties on fourteen standardized reading comprehension tests. Durham, NC: MetaMetrics, Inc.

Stenner, A.J., Smith, D.R., Horabin, I., & Smith, M. (1987b). Fit of the Lexile Theory to sequenced units from eleven basal series. Durham, NC: MetaMetrics, Inc.

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.

Williamson, G. L. (2008). A text readability continuum for postsecondary readiness. *Journal of Advanced Academics, 19*(4), 602-632.

Williamson, G. L. (2011, March 16-17). *Growth and growth norms: Standards for academic growth.* Presentation at the first semi-annual meeting of MetaMetrics' Technical Advisory Committee, held at MetaMetrics, Inc., Durham, NC.

Williamson, G. L., Koons, H., Sandvik, T., & Sanford-Moore, E. (2012). *The text complexity continuum in grades 1-12* (MetaMetrics Research Brief). Durham, NC: MetaMetrics, Inc.

Williamson, G. L., Thompson, C. L., and Baker, R. F. (2007, April). *North Carolina's growth in reading and mathematics.* AERA Distinguished Paper Presentation at the 2007 annual meeting of the American Educational Research Association (AERA), Chicago, IL.

Wright, B.D. & Linacre, J.M. (1994, August). *The Rasch model as a foundation for the Lexile Framework*. Unpublished manuscript.

Wright, B.D. & Stone, M.H. (1979). *Best Test Design*. Chicago: MESA Press.