1220L | Lexile: Matching readers to text

The Lexile Framework® for Reading Technical Report

MetaMetrics, Inc.

1000 Park Forty Plaza Drive, Suite 120
Durham, NC 27713
www.Lexile.com
April 2007

The Lexile Framework® for Reading
Theoretical Framework and Development

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.

Readability Formulas and Reading Levels

Readability Formulas. Traditional readability formulas have been used for more than 60 years. These formulas are generally based on a theory about written language and use mathematical equations to calculate text difficulty. While each has discrete features, nearly all attempt to assign difficulty based on a combination of semantic (vocabulary) features and syntactic (sentence length) features. Traditional readability formulas are all based on a simple theory about written language and a simple equation to calculate text difficulty.

Unless a user is interested in doing research, there is little to be gained from choosing a highly complex formula. A simple two-variable formula is sufficient, especially if one of the variables is a word or semantic variable and the other is a sentence of syntactic variable. Beyond these two variables, further additions add relatively little predictive validity compared to the added application time involved and a formula with very many variables is likely to be unreliably applied by hand.

The earliest formulas of readability appeared in the 1920s. Some of them were esoteric and primarily intended for chemistry and physics textbooks, or for shorthand dictation materials. The first milestone that provided an objective way of estimating word difficulty was Thorndike's *The Teacher Word Book* published in 1921. The concepts discussed in Thorndike's book led Lively and Pressey in 1923 to develop the first readability formula based on the tabulations of the frequency of which words appear. In 1928, Vogel and Washburne developed a formula that took the form of a regression equation involving more than one language variable. This format became the prototype for most of the formulas that followed. The work of Washbourne and Morphett in 1938 provided a formula, which yielded scores on a grade-placement scale. The trend to make the formulas easy to apply resulted in the most widely used of all readability formulas—Flesch's Reading Ease Formula (1948). Dale and Chall (1948) published another two-variable formula that became very popular in educational circles. Spache designed his renowned formula using a word-list approach in 1953. This design was useful for grades 1 through 3 at a time when most formulas were designed for the upper grade levels. This same year, Taylor proposed the cloze procedure for measuring readability. Twelve years later, Coleman used this procedure for the creation of his fill-in-the-blank method as a criterion for his formula. Danielson and Bryan developed the first computer-generated formulas in 1963. Also, in 1963, Fry simplified the process of interpreting readability formulas by developing a readability graph. Later, in 1977, he extended his readability graph and his method is the most widely used of all current methods (Klare, 1984; Zakaluk and Samuels, 1988).

Two often-used formulas—the Fog Index and the Flesch-Kincaid Readability Formula—can be calculated by hand for short passages. First, select a passage that contains 100 words. For a lengthy piece of text, select several different 100-word passages.

For the *Fog Index*, first determine the average number of words per sentence. If the passage does not end at a sentence break, calculate the percentage of the final sentence in the passage and add to the count of the number of sentences. Determine the percentage of "long" words (ones with three of more syllables). Add the two measures and multiple by 0.4. This number indicates the approximate Reading Grade Level (RGL) of the passage.

For the *Flesch-Kincaid Readability Formula* (found in Microsoft Word), use the following equation:

RGL = 0.39(average number of words per sentence) +
11.8(average number of syllables per word) – 15.59

For a lengthy piece of text, using either formula, average the RGLs for the several different 100-word passages.

Another readability formula commonly used is ATOS™ for Books developed by Advantage Learning Systems. ATOS is based on the following variables related to the reading demands of text: words per sentence, characters per word, and average grade level of the words. ATOS uses whole-book scans instead of text samples and results are reported on a grade-level scale.

Guided Reading Levels. Within the Guided Reading framework (Fountas & Pinnell, 1996), books are assigned to levels by teachers according to specific characteristics. These characteristics include the level of support provided by the text (e.g., the use and role of illustrations, the size and layout of the print) and the predictability and pattern of language (e.g., oral language compared to written language). An initial list of leveled books is provided so teachers can have a place to start when leveling a book.

For students in kindergarten through grade 3, there are 18 Guided Reading Levels, A through R (kindergarten—Levels A through C; 1st Grade—Levels A through I; 2nd Grade—Levels C through P; and 3rd Grade—Levels J through R). The books include a variety of genres: informational texts on a variety of topics, "How to" books, mysteries, realistic fiction, historical fiction, biography, fantasy, traditional folk and fairy tales, science fiction, and humor.

How do readability formulas and reading levels relate to readers? The previous section described how to level books in terms of grade levels and reading levels based on the characteristics of the text. But, how do we connect these levels to the reader? Do we say that a reader in grade 6 should only read books that have a readability level between 6.0 and 6.9? How do we know that a student is reading at Guided Reading Level “G” and when is he or she ready to move on to Level “H”? What we need is some way to put readers on these scales.

To match students with readability levels, we need to determine their “reading“ or “social studies” grade level, which is often not the same as their “nominal” grade level (the grade level of the class they are in). On a test, a grade equivalent (GE) is a score that represents the typical (mean or median) performance of students tested in a given month of the school year. For example, if Alicia, a fourth-grade student, obtained a GE of 4.9 on a fourth-grade reading test, her score is like the score a student at the end of the ninth month of fourth grade would likely score on that same reading test. There are two main problems with grade equivalents:

1. *How grade equivalents are derived determine the appropriate conclusions that may be drawn from the scores*. For example, if Stephanie scores 5.9 on a fourth-grade mathematics test it is not appropriate to conclude that Stephanie has mastered the mathematics content

of the 5th grade (in fact, it may be unknown how 5th grade students would perform on the 4th grade test). It certainly cannot be assumed that Stephanie has the prerequisites for 6th grade mathematics. All that is known for sure is that Stephanie is well above average in mathematics.

2. *Grade equivalents represent unequal units*. The content of instruction varies somewhat from grade to grade (such as in high school where subjects may only be studied one or two years) and the emphasis placed on a subject may vary from grade to grade. Grade units are unequal and these inequalities occur irregularly in different subjects. A difference of one grade equivalent in reading in elementary school (2.6 to 3.6) is not the same as a difference of one grade equivalent in middle school (7.6 to 8.6).

To match students with Guided Reading Levels, the teacher makes decisions based on observations of what the child can or cannot do to construct meaning. Teachers also use ongoing assessments such as running records, individual conferences, and observations of students' reading to monitor and support student progress.

Both of these approaches to helping readers select books appropriate to their reading level—readability formulas and reading levels—are subjective and prone to misinterpretation. What is needed is one scale that can describe the reading demands of a piece of text and the readability of a child. The Lexile Framework for Reading is a powerful tool for determining the reading ability of children ***and*** finding texts that provide the appropriate level of challenge.

Jack Stenner, a leading psychometrician and one of the developers of the Lexile Framework, likens this situation to an experience he had several years ago with his son.

> Some time ago I went into a shoe store and asked for a fifth-grade shoe. The clerk looked at me suspiciously and asked if I knew how much shoe sizes varied among eleven-year-olds. Furthermore, he pointed out that shoe size was not nearly as important as purpose, style, color, and so on. But if I would specify the features I wanted and the size, he could walk to the back and quickly reappear with several options to my liking. The clerk further noted, somewhat condescendingly, that the store used the same metric to measure feet and shoes, and when there was a match between foot and shoe, the shoes got worn, there was no pain, and the customer was happy and became a repeat customer. I called home and got my son's shoe size and then asked the clerk for a "size 8-red-hightop-Penny Hardaway-basketball shoe." After a brief transaction, I had the shoes.
>
> I then walked next door to my favorite bookstore and asked for a fifth-grade fantasy novel. Without hesitation, the clerk led me to a shelf where she gave me three choices. I selected one and went home with *The Hobbit*, a classic that I had read three times myself as a youngster. I later learned my son had yet to achieve the reading fluency needed to enjoy *The Hobbit*. His understandable response to my gifts was to put the book down in favor of passionately practicing free throws in the driveway.

The next section of this technical manual describes the development and validation of The Lexile Framework for Reading.

The Lexile Framework for Reading

A reader's comprehension of text is dependent on many factors—the purpose for reading, the ability of the reader, and the text that is being read. The reader can be asked to read a text for entertainment (literary experience), to gain information, or to perform a task. The reader brings to the reading experience a variety of important factors: reading ability, prior knowledge, interest level, and developmental appropriateness. For any text, there are three factors associated with the readability of the text: difficulty, support, and quality. All of these factors are important considerations when evaluating the appropriateness of a text for a reader. The Lexile Framework focuses primarily on two: reader ability and text difficulty.

Within the Lexile Framework, text difficulty is determined by examining the characteristics of word frequency and sentence length. Text measures typically range from 0L to 1800L, but they can go below zero (reported as "Beginning Reader") and above 2000L. Within any one classroom there will be a range of reading materials.

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. It is clear that most operationalizations of semantic complexity 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 high 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 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). A "word family" included: (1) the stimulus word; (2) all plurals (adding "-s"

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. Correlations were computed between algebraic transformations of these means 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 mean log word frequency provided the highest correlation with item rank order (r = –0.779) for the items on the combined form.

The Lexile Framework currently employs a 600-million-word corpus when examining the semantic component of text. This corpus was assembled from the thousands of texts publishers have measured.

The Syntactic Component. Klare (1963) provided 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.

Calibration of Text Difficulty. A research study on semantic units conducted by 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, Horiban, 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*. 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 when the 9 tests were combined into one was 0.93 after correction for range restriction and measurement error.

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

The Lexile Scale. In developing the Lexile scale, the Rasch item response theory model (Wright and Stone, 1979) was used to estimate the difficulties of items and the abilities of persons on the logit scale.

The calibrations of the items from the Rasch model are objective in the sense that the relative difficulties of the items will remain the same across different samples of persons (specific objectivity). When two items are administered to the same person it can be determined which

item is harder and which one is easier. This ordering is likely to hold when the same two items are administered to a second person. If two different items are administered to the second person, there is no way to know which set of items is harder and which set is easier. The problem is that the location of the scale is not known. General objectivity requires that scores obtained from different test administrations be tied to a common zero—absolute location must be sample independent (Stenner, 1990). To achieve general objectivity, the theoretical logit difficulties must be transformed to a scale where the ambiguity regarding the location of zero is resolved.

The first step in developing 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 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 comprehensibility of the primers and the comprehensibility 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 anchor points and then solving the system of equations.

Validity of The Lexile Framework for Reading

Validity is the "extent to which a test measures what its authors or users claim it measures; specifically, test validity concerns the appropriateness of inferences that can be made on the basis of test results" (Salvia and Ysseldyke, 1998). The 1999 *Standards for Educational and Psychological Testing* (America Educational Research Association, American Psychological Association, and National Council on Measurement in Education) state that “validity refers to the degree to which evidence and theory support the interpretations of test scores entailed in the uses of tests” (p. 9). 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 construct validity of The Lexile Framework for Reading can be evaluated by examining how well Lexile measures relate to other measures of reading comprehension and text difficulty.

Lexile Framework Linked to other Measures of Reading Comprehension. The Lexile Framework for Reading has been linked to several standardized tests of reading

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

Table 2 presents the results from linking studies conducted with The Lexile Framework for Reading. For each of the tests listed, student reading comprehension scores can also be reported as Lexile measures. This dual reporting provides a rich, criterion-related frame of reference for interpreting the standardized test scores. When a student takes one of the standardized tests, in addition to receiving his norm-referenced test results, he can receive a reading list that is targeted to his specific reading level.

Table 2. 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)	1–8	16,363	0.72 to 0.88*
Proficiency Assessment for Wyoming Students (PAWS)	3, 5, 8, and 11	3,871	0.91

Notes: Results are based on final samples used with each linking study.
*TAAS, TAKS and CRCT were not vertically equated; separate linking equations were derived for each grade.
**CST was linked using a set of Lexile calibrated items embedded in the CST research blocks. CST items were calibrated to the Lexile scale.

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 is. 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 3*).

Table 3. 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 3*, 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 publisher 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 (Poznansky, 1990; Stenner, 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 equation did not fit poetry items or non-continuous 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 non-continuous prose items were removed and correlations were recalculated. *Table 4* contains the results of this analysis.

Table 4. 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 4* 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.

Forecasting Comprehension with the Lexile Framework

An important feature of the Lexile Framework is that it also provides criterion-referenced interpretations of every measure. A criterion-referenced interpretation of a test score compares the specific knowledge and skills measured by the test to the student's proficiency with the same knowledge and skills. Criterion-referenced scores have meaning in terms of what the student knows or can do, rather than in relation to the scores produced by some external reference (or norm) group.

When a reader's measure is equal to the task's calibration, then the Lexile scale forecasts that the individual has a 75% comprehension rate on that task. When 20 such tasks are given to this reader, one expects three-fourths of the responses to be correct. If the task is more difficult than the reader is able, then the probability is less than 75% that the response of the person to the task will be correct. Similarly, when the task is easier compared to a reader's measure, then the probability is greater than 75% that the response will be correct.

There is empirical evidence supporting the choice of a 75% target comprehension rate, as opposed to, say, a 50% or a 90% rate. Squires, Huitt, and Segars (1983) observed that reading achievement for second-graders peaked when the success rate reached 75%. A 75% success rate also is supported by the findings of Crawford, King, Brophy, and Evertson (1975), Rim (1980), and Huynh (1998). It may be, however, that there is no one optimal rate of reading comprehension. It may be that there is a range in which individuals can operate to optimally improve their reading ability.

Since the Lexile Theory provides complementary procedures for measuring people and text, the scale can be used to match a person's level of comprehension with books that the person is forecast to read with a high comprehension rate. Trying to identify possible supplemental reading materials for students has, for the most part, relied on a teacher's familiarity with the titles. For example, an eighth-grade girl who is interested in sports but is not reading at grade level may be interested in reading a biography about Chris Evert. The teacher may not know, however, whether a specific biography is too difficult or too easy for the student. The Lexile Framework provides a reader measure and a text measure on the same scale. Armed with this information, a teacher, librarian, media specialist, student, or parent can plan for success.

Readers develop reading comprehension skills by reading. Skill development is enhanced when their reading is accompanied by frequent response requirements. Response requirements may be structured in a variety of ways. An instructor may ask oral questions as the reader progresses through the prose or written questions may be embedded in the text, much as is done with *Scholastic Reading Inventory* items. Response requirements are important; unless there is some evaluation and self-assessment, there can be no assurance that the reader is properly targeted and comprehending the material. Students need to be given a text on which they can practice being a competent reader (Smith, 1973). The above approach does not complete a fully articulated instructional theory, but its prescription is straightforward. Students need to read more and teachers need to monitor this reading with some efficient response requirement. One implication of these notions is that some of the time spent on skill sheets might be better spent reading targeted prose with concomitant response requirements (Anderson, Hiebert, Scott, and Wilkinson, 1985). This approach has been supported by the research of Five (1986) and Hiebert (1998).

As the reader improves, new titles with higher text measures can be chosen to match the growing reader ability. This results in a constantly growing person-measure, thus keeping the comprehension rate at the most productive level. We need to locate a reader's "edge" and then

expose the reader to text that plays on that edge. When this approach is followed in any domain of human development, the edge moves and the capacities of the individual are enhanced.

What happens when the "edge" is over-estimated and repeatedly exceeded? In physical exertion, if you push beyond the edge you feel pain; if you demand even more from the muscle, you will experience severe muscle strain or ligament damage. In reading, playing on the edge is a satisfying and confidence-building activity, but exceeding that edge by over-challenging readers with out-of-reach materials reduces self-confidence, stunts growth, and results in the individual "tuning out." The tremendous emphasis on reading in daily activities makes every encounter with written text a reconfirmation of a poor reader's inadequacy.

For individuals to become competent readers, they need to be exposed to text that results in a comprehension rate of 75% or better. If an 850L reader is faced with an 1100L text (resulting in a 50% comprehension rate), there will be too much unfamiliar vocabulary and too much of a load placed on the reader's tolerance for syntactical complexity for that reader to attend to meaning. The rhythm and flow of familiar sentence structures will be interrupted by frequent unfamiliar vocabulary, resulting in inefficient chunking and short-term memory overload. When readers are correctly targeted, they read fluidly with comprehension; when incorrectly targeted, they struggle both with the material and with maintaining their self-esteem. *Within the Lexile Framework, there are no poor readers—only mistargeted readers who are being over challenged.*

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 to be 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.

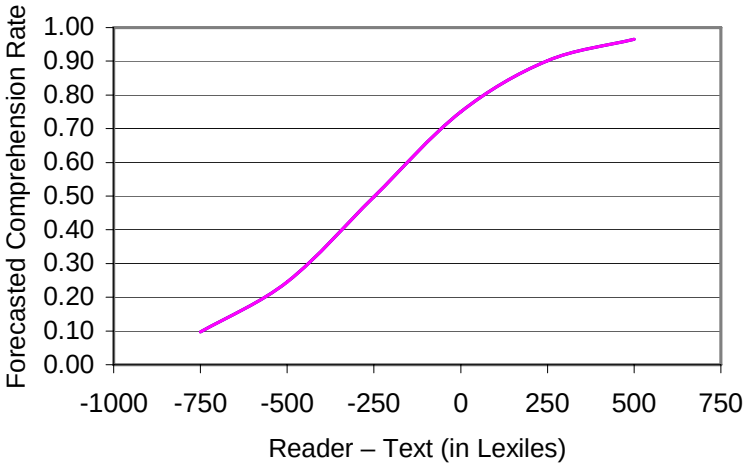
Suppose instead that the text and reader measures are not the same. It is the difference in Lexiles between reader and text that governs comprehension. If the text measure is less than the reader measure, the comprehension rate will exceed 75 percent. If not, it will be less. The question is "By how 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 (Equation 1 on page 7). 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.

Figure 2 shows the general relationship between reader-text discrepancy and forecasted comprehension rate. When the reader measure and the text measure are the same (difference of 0L on the *x*-axis) then the forecasted comprehension rate is 75%. In the example in the preceding paragraph, the difference between the reader measure of 600L and the text measure

of 350L is 250L. Referring to *Figure 2* and using +250L (reader minus text), the forecasted comprehension rate for this reader-text combination would be 90%.

Figure 2. Relationship between reader-text discrepancy and forecasted reading comprehension rate.



Tables 4 and *5* show comprehension rates calculated for various combinations of reader measures and text measures.

Table 4. Comprehension rates for the same individual with materials of varying comprehension difficulty.

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

Table 5. Comprehension rates of different ability persons with the same material.

Person Measure	Calibration for *Sports Illustrated*	Forecast Comprehension Rate
500L	1000L	25%
750L	1000L	50%
1000L	1000L	75%
1250L	1000L	90%
1500L	1000L	96%

The subjective experience of 50%, 75%, and 90% comprehension as reported by readers varies greatly. A 1000L reader reading 1000L text (75% 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 sound like they comprehend what they are reading. A 1000L reader reading 1250L text (50% 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 difficulty. Finally, a 1000L reader reading 750L text (90% 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.

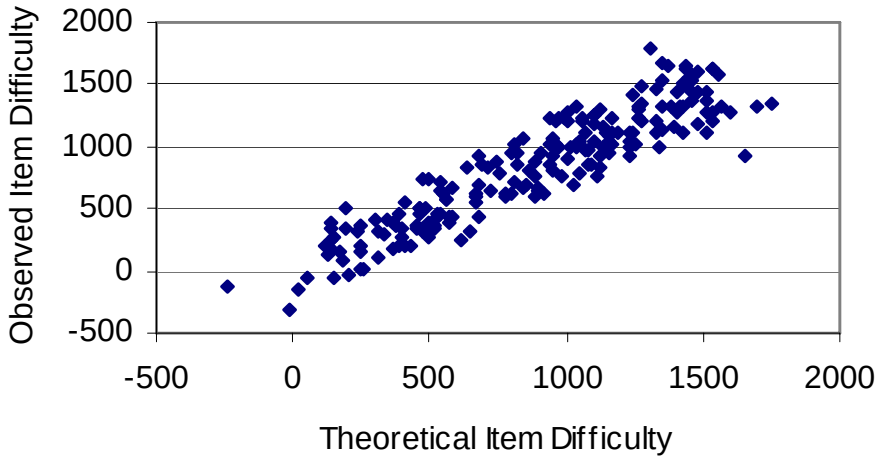
Text Measure Error Associated with the Lexile Framework

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. The result would be an identical text measure to the first because there is no sampling error when a complete text is calibrated.

Study 1. 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 scatterplot of the data is presented in *Figure 3*.

Figure 3. 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 Lexile Titles Database (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.

Study 2. A second study was conducted during 2002 to examine ensemble differences across items. An ensemble consists of the all of the items that could be developed a selected piece of text. The Lexile measure of a piece of text is the mean difficulty

Participants. Participants in this study were students from four school districts in a large southwestern state. These students were participating in a larger study that was designed assess reading comprehension with the Lexile scale. The total sample included 1,186 grade 3 students, 893 grade 5 students, and 1,531 grade 8 students. The mean tested abilities of the three samples were similar to the mean tested abilities of all students in each grade on the state reading assessment. Though 3,610 students participated in the linking study, the data records for only 2,867 of these students were used for determining the ensemble item difficulties presented in this paper. The students were administered one of four forms at each grade level. The reduction in sample size is because one of the four forms was created using the same ensemble items as another form. For consistency of sample size across forms, the data records from this fourth form were not included in the ensemble study.

Instrument. Thirty text passages were response-illustrated by three different item writing teams resulting in three items nested within each of 30 passages for a total of 90 items. All three teams employed a similar item-writing protocol. The ensemble items were spiraled into test forms at the grade level (3, 5, or 8) that most closely corresponded with the item's theoretical calibration.

Winsteps (Wright & Linacre, 2003) was used to estimate item difficulties for the 90 ensemble study items. Of primary interest in this study was the correspondence between theoretical text calibrations and the 30 ensemble means and the consequences that theory misspecification holds for text measure standard errors.

Results. Table 32 presents the ensemble study data in which three independent teams wrote one item for each of thirty passages for ninety items. Observed ensemble means taken over the three ensemble item difficulties for each passage are given along with an estimate of the within ensemble standard deviation for each passage.

Table 32. Analysis of 30 item ensembles providing an estimate of the theory misspecification error.

Item Number	Theory (T)	Team A	Team B	Team C	Mean[a] (O)	SD[b]	Within Ensemble Variance	T-O
1	400L	456	553	303	437	126	15,909	-37
2	430L	269	632	704	535	234	54,523	-105
3	460L	306	407	483	399	88	7,832	61
4	490L	553	508	670	577	84	6,993	-87
11	510L	267	602	468	446	169	28,413	64
5	540L	747	825	654	742	86	7,332	-202
6	569L	909	657	582	716	172	29,424	-147
7	580L	594	683	807	695	107	11,386	-115
8	620L	897	805	497	733	209	43,808	-113
9	720L	584	850	731	722	133	17,811	-2
12	720L	953	587	774	771	183	33,386	-51
13	745L	791	972	490	751	244	59,354	-6
14	770L	855	1017	958	944	82	6,717	-174
16	770L	1077	1095	893	1022	112	12,446	-252
15	790L	866	557	553	659	180	32,327	131
21	812L	902	1133	715	917	209	43,753	-105
10	820L	967	740	675	794	153	23,445	26
17	850L	747	864	674	762	96	9,257	88
22	866L	819	809	780	803	20	419	63
18	870L	974	1197	870	1014	167	28,007	-144
19	880L	1093	733	692	839	221	48,739	41
23	940L	945	1057	965	989	60	3,546	-49
24	960L	1124	1205	1170	1166	41	1,653	-206
25	1010L	926	1172	899	999	151	22,733	11
20	1020L	888	1372	863	1041	287	82,429	-21
26	1020L	1260	987	881	1043	196	38,397	-23
27	1040L	1503	1361	1239	1368	132	17,536	-328
28	1060L	1109	1091	981	1061	69	4,785	-1
29	1150L	1014	1104	1055	1058	45	2,029	92
30	1210L	1275	1291	1014	1193	156	24,204	17

Total MSE = Average of $(T-O)^2$ = 12022; Pooled within variance for ensembles = 7984; Remaining between ensemble variance = 4038; Theory misspecification error = 64L
Barlett's test for homogeneity of variance produced an approximate chi-square statistic of 24.6 on 29 degrees of freedom and sustained the null hypothesis that the variances are equal across ensembles.

Note. All data is reported in Lexiles.
a. Mean (O) is the observed ensemble mean.
b. SD is the standard deviation within ensemble.

The difference between passage text calibration and observed ensemble mean is provided in the last column. The RMSE from regressing observed ensemble means on text calibrations is 110L. *Figures 12a* and *12b* shows a plot of observed ensemble means against theoretical text calibrations.

Figure 13a. Plot of observed ensemble means and theoretical calibrations (RMSE = 111L).

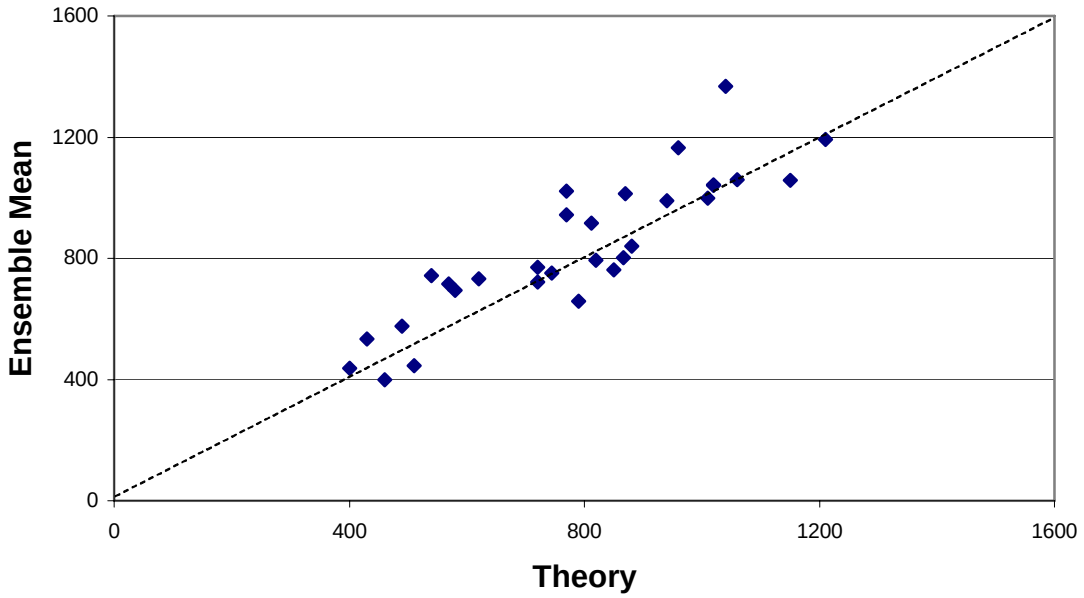
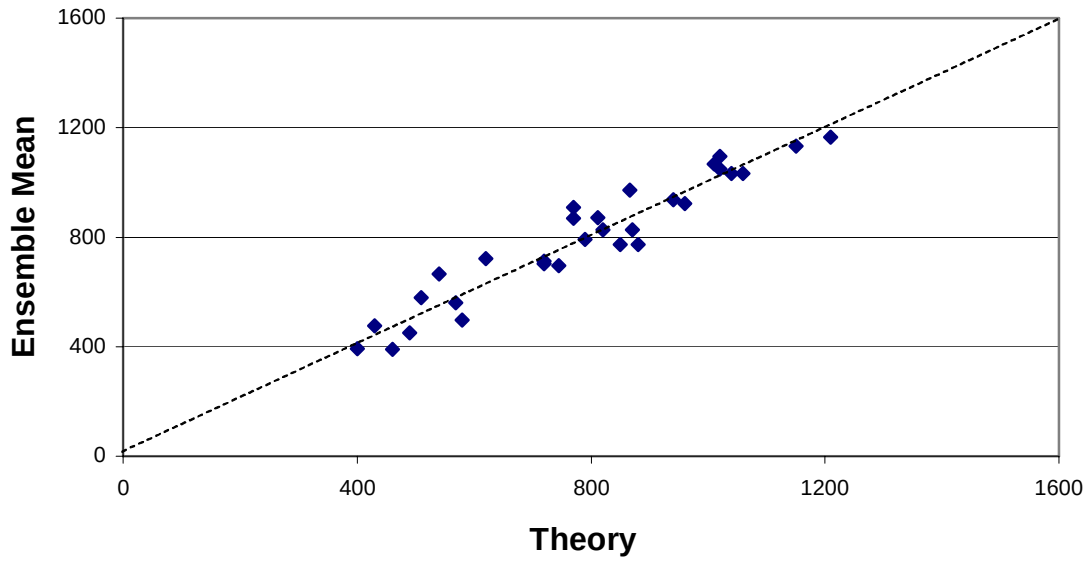


Figure 13b. Plot of simulated "true" ensemble means and theoretical calibrations (RMSE = 64L).



Note, that some of the deviations about the identity line are because ensemble means are poorly estimated given that each mean is based on only three items. The bottom panel in Figure 2 depicts simulated data when an error term [distributed ~$N(0, \sigma = 64L)$] is added to each

theoretical value. Contrasting the two plots in *Figures 10a* and *10b* provides a visual depiction of the difference between regressing observed ensemble means on theory and regressing "true" ensemble means on theory. An estimate of the RMSE when "true" ensemble means are regressed on the Lexile Theory is 64L ($\sqrt{110^2 - 89^2}$ = $\sqrt{4{,}038}$ = 63.54). This is the average error at the passage level when predicting "true" ensemble means from the Lexile Theory.

Since the RMSE equal to 64L applies to the expected error at the passage/slice level, a text made up of n_i slices would have an expected error of $64 \div \sqrt{n_i}$. Thus, a short periodical article of 500 words (n_i = 4) would have a SEM of 32L ($64 \div \sqrt{4}$), whereas a much longer text like the novel *Harry Potter: Chamber of Secrets* (880L, Rowling, 2001) would have a SEM of 2L ($64 \div \sqrt{900}$). *Table 15* contrasts the SEMs computed using the old method with SEMs computed using the Lexile Framework for several books across a broad range of Lexile measures.

Table 33: Old method text readabilities, resampled SEMs, and new SEMs for selected books.

Book	Number of Slices	Lexile Measure	Resampled Old SEM[a]	New SEM
The Boy Who Drank Too Much	257	447L	102	4
Leroy and the Old Man	309	647L	119	4
Angela and the Broken Heart	157	555L	118	5
The Horse of Her Dreams	277	768L	126	4
Little House by Boston Bay	235	852L	126	4
Marsh Cat	235	954L	125	4
The Riddle of the Rosetta Stone	49	1063L	70	9
John Tyler	223	1151L	89	4
A Clockwork Orange	419	1260L	268	3
Geometry and the Visual Arts	481	1369L	140	3
The Patriot Chiefs	790	1446L	139	2
Traitors	895	1533L	140	2

Three slices selected for each replicate: one slice from the first third of the book, one from the middle third and one from the last third. Resampled 1,000 times. SEM = SD of the resampled distribution.

Lexile Item Bank

The Lexile Item Bank contains over 10,000 items that have been developed between 1986 and 2003 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. 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 are compiled from spelling books one grade level below the level the item would typically be used with. The rationale is that these words should be part of a reader's "working" vocabulary since they should have 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 on universal design and fair-access—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 ongoing 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.