1340L I Lexile: Matching readers to text

Total Reader Technical Manual

MetaMetrics, Inc.

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

December 2006

Table of Contents

Page

List of Tables

List of Figures

Page

Introduction

Total Reader is an online benchmarking system for grades 3 through 12 utilizing The Lexile Framework® for Reading to target student reading levels. Total Reader matches student reading ability with text readability using a common objective scale and provides long-term progress tracking for students, schools, and districts.

Total Reader consists of three main sections to help students and teachers monitor reading progress: the Reading Zone, the Suggested Reading pages, and the Reporting section. The Reading Zone encompasses the main student testing experience. After taking an initial diagnostic test, students complete on-going, progress-monitoring testlets. Each testlet consists of a reading passage with associated test items. Students select passages based on their individual interests and reading level. The computer-adaptive system allows students access to only those reading selections that are within their individual targeted reading range (the suggested range of Lexiles at which the reader should be reading; from 50L above her or his Lexile measure to 100L below). While reading the selection, students are presented with cloze items, or embedded completion statements, which they complete by clicking on the best response from four choices. Once the passage is completed, students answer a final monitoring question that triggers the scoring process, and their updated Lexile measure is computed. Using this structure, Total Reader continually generates updated Lexile measures, and students are always aware of their reading level.

The Suggested Reading pages allow students to access appropriate resources that are within their Lexile range. All students can access the Lexile Titles Database to search for titles. At schools with specified licenses, students can also access Lexile-calibrated database resources.

The Reporting section allows students, teachers, and administrators to monitor longitudinal reading progress. The student Progress Report consists of basic profile information (name, grade, school, district, class association), graphic representations of both Lexile progress and program usage, Lexile history, and passage history (with data on genre selection, passage scores, date taken, etc.). In addition to the student Progress Report, the administrative reports provide more detailed information on class, school, and district-level performance. Data on student usage, improvement ranges, rankings, as well as low, high, and average Lexile measure per class/grade are provided. Additional information on recommended reading ranges per student Lexile measure is also provided. Access to the varying levels of these reports is dependent upon the administrative level of the user. There are currently seven levels of access: student, class, school, district, region, state, and system.

The foundation upon which Total Reader rests is The Lexile Framework for Reading, a scientifically based scale of reading ability. All measures within Total Reader are calculated using the Lexile Analyzer, the Lexile scale, and the Lexile Reader/Writer engine developed by MetaMetrics, Inc. A Bayesian scoring algorithm is used to provide measures that monitor progress in reading development. With these tools, Total Reader provides accurate information to help students and teachers measure progress and forecast student development.

This technical guide for the Total Reader system should provide users with a broad research foundation. Such a base is essential when deciding if and how the Total Reader assessment results should be used and what kinds of inferences about readers are permissible.

Background

In January 2002, President George W. Bush signed the No Child Left Behind Act of 2001. The act rewards schools, districts, and states that improve student achievement, and it sanctions those entities that fail. It mandates that each state develop a system to measure the progress of all students and subgroups of students in meeting state-determined grade-level standards. It suggests that the assessments states use will provide an important diagnostic tool that can aid in achieving improvement. Many states have responded to the Act's provisions by adapting their existing assessment programs, while others have developed new assessments. These assessments are typically administered only once a year. Many schools and districts, however, find that they would benefit from more frequent monitoring of students' achievement and progress, particularly in reading. Such results would help them more accurately assess student progress in an ongoing manner and also prepare them for the results of their high-stakes assessments.

In 2003, Dr. Annette Bohling, Wyoming's Deputy Superintendent of Educational Quality and Accountability, approached EdGate, about Wyoming's need for an online reading assessment program. Bohling was interested in having EdGate develop a system that would provide ongoing monitoring and evaluation of students' achievement in reading. Wyoming had recently developed an initiative to use the Lexile Framework for Reading as a means to help teachers identify student reading levels and then provide instructional support to improve that level of proficiency. What was missing was an efficient, reliable method for producing a student's Lexile measure and giving the student access to materials that would help him or her increase his or her reading level. The system would need to be portable (Web-based), easy to use, and provide extensive reporting capabilities that would manage data about student reading achievement at the class, school, district, and state levels.

Dr. Jimmy Kim, a researcher at the University of California-Irvine, demonstrated in a randomized field study that low-income students are not destined to summer loss. Dr. Kim showed that low-income students’ skills could, in fact, grow over the summer if they were able to select books at their interest level and reading level. Their gains in reading were comparable to gains one would expect in summer school (Kim, 2005, 2006). Since motivation is key to voluntary reading, two critical features of book selection are interest and reading level, and both were addressed in Kim’s study. Dr. Kim used a tool that many states use to make sure that students are appropriately challenged.

To develop an assessment program to meet Wyoming's needs, EdGate modeled a program similar to the work of Kim and worked with MetaMetrics, Inc. (developers of The Lexile Framework for Reading). MetaMetrics licensed to EdGate the Lexile Analyzer, a computer program that analyzes the difficulty of text, and the Inline Reader/Writer engine to develop and score testlets (reading assessments). EdGate developed the Total Reader user interface and the reading passages that comprise the bulk of the Total Reader system.

EdGate launched a pilot version of Total Reader in Wyoming summer schools in June 2004. A revised version of Total Reader was released to the state of Wyoming in September 2004. Total Reader was launched nationally in November 2004.

Features of Total Reader

Several specific features of the Total Reader system are noteworthy.

- Passages are authentic: they are sampled from materials students would read in and out of the classroom.

- The native-Lexile item format used on the diagnostic reading assessments and the passage native-Lexile item format used in the Total Reader testlets are extensions of the “embedded completion” item format that has been shown to measure the same core reading competency that is measured by norm-referenced, criterion-referenced, and individually administered reading tests (Stenner, Smith, Horiban, and Smith, 1987a).

- More than a decade of research went into defining the rules for sampling text and writing embedded completion items. These rules were precisely followed in developing the Total Reader diagnostic items and testlet items.

- Total Reader testlets are linked to the Lexile scale and, as such, the item calibrations used to convert a raw score (number correct) into the Lexile metric are provided by the Lexile Theory. The calibration equation used to calibrate Total Reader testlet passages and test items is the same equation that is used to measure books/texts. Thus, readers and texts are placed on the same scale.

- The Total Reader test format supports quick administration in an un-timed, low-pressure format.

- Total Reader testlets are administered individually online, scored immediately, and results are returned to the student at the end of the session.

- Total Reader employs a computer-adaptive algorithm to administer the test by adapting the level of each testlet to the specific level of the reader. This methodology continuously targets the reading level of the student and thus produces more precise measurements when compared with “fixed-form” assessments.

- Total Reader supports rapid objective scoring by computer.

- Total Reader uses a Bayesian scoring algorithm such that past performance is used to predict future performance. Bayesian methodology provides a paradigm for combining prior information with current data, both subject to uncertainty, to come up with an estimate of current status, which is again subject to uncertainty. This methodology connects the administration of each testlet to the administration of every other testlet and thus produces more precise measurements when compared with independent assessments.

- No extensive or specialized preparation is needed to administer Total Reader testlets although proper interpretation and use of the results requires an understanding of the Lexile Framework for Reading.

Using the Lexile Framework for Reading

For Teachers, parents, and students can use the tools provided by the Lexile Framework to plan instruction. When teachers provide parents and students with lists of titles that match the students' Lexile measures, they can then work together to choose appropriate titles that also match the students' interest and background knowledge. The Lexile Framework does not prescribe a reading program, but is a tool that gives educators more control over the variables involved when they design reading instruction. The Lexile Framework yields multiple opportunities for use in a variety of instructional activities. After becoming familiar with the Lexile Framework, teachers are likely to think of a variety of additional creative ways to use this tool that match students to books that students find challenging but not frustrating.

The Lexile Framework is a system that helps match readers with literature appropriate for their reading skills. When reading a book within his or her Lexile range (50L above his or her Lexile measure to 100L below), the reader should comprehend enough of the text to make sense of it, while still being challenged enough to maintain interest and learning.

Remember, there are many factors that affect the relationship between a reader and a book. These factors include content, age of the reader, interest, suitability of the text, and text difficulty. The Lexile measure of a text, a measure of text difficulty, is a good starting point in the selection process with other factors then being considered. The Lexile measure should never be the only factor considered when selecting a text.

Purposes and Uses of Total Reader

Total Reader is designed to measure a reader's ability to comprehend expository texts of increasing difficulty. The results of Total Reader can be used to measure where students stand in the development of their reading ability.

One outcome of Total Reader is the location of the reader on the Lexile Map (Appendix A). Once a reader is measured, it is possible to forecast how well the reader will comprehend thousands of books that have been measured in the Lexile metric. Readers and texts are similarly measured in the same Lexile metric making it possible to directly compare a reader and text. When reader and text measures match, the Lexile Framework forecasts 75% comprehension. The operational definition of 75% comprehension is that given 100 items from a text, the reader will be able to correctly answer 75. When the text has a Lexile measure 250L higher than the reader measure, the Framework forecasts 50% comprehension. When the reader measure exceeds the text measure by 250L, the forecasted comprehension is 90%.

The data provided by the Total Reader reports can help educators make more guided decisions about materials selection, particularly in cases where differentiated instruction is the goal. It is an ideal tool for students who require extra attention in reading, such as IEP or ESL students. Administrators can also use the class and school data to evaluate curriculum and document accountability. It is yet one more tool available to document adequate yearly progress.

Limitations

Instructional decisions are best made when using multiple sources of evidence about a student. Other sources include standardized test data, instructional group placement, lists of books read, and, most importantly, teacher judgment. *One measure of student performance, taken on one day, is never sufficient to make high-stakes student-level decisions such as summer school placement or retention.*

Theoretical Framework of Reading Ability and 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.

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:

$$\text{RGL} = 0.39(\text{average number of words per sentence}) + 11.8(\text{average number of syllables per word}) - 15.59$$

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.

Like other readability formulas, the Lexile Framework examines two features of text to determine it's readability—semantic difficulty and syntactic complexity. Within the Lexile Framework, text difficulty is determined by examining the characteristics of word frequency and sentence length. Text measures typically range from 200L to 1700L, 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.

The Semantic Component. It is clear that most operationalizations of semantic difficulty 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. 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) 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.

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. 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 1/100th of the difference between freezing (0 degrees) and boiling (100 degrees) water. For the Lexile scale, the unit size was defined as 1/1000th 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 1/1000th 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 numerours 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 1 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 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 Tests (Reading Comprehension)	3, 4, 5, 8	956	0.90
TerraNova (CTBS/5)	2, 4, 6, 8	2,713	0.92
Texas Assessment of Academic Skills (TAAS)	3–8	3,623	0.73 to 0.78*
Metropolitan Achievement Test (Eighth Edition)	2, 4, 6, 8, and 10	2,382	0.93
Gates-MacGinitie Reading Test (Version 4)	2, 4, 6, 8, and 10	4,644	0.92
Utah Core Assessments	3–6	1,551	0.73
Texas Assessment of Knowledge and Skills	3, 5, and 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 Knowledge and Skills	3, 5, 8, and 10	3,180	0.89
California Standards Test (CST)	2 though 12	55,564	NA**

Notes: Results are based on final samples used with each linking study.
*TAAS and TAKS are not vertically scaled; 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 hypothesized 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 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 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 (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 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 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 correlations 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

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?

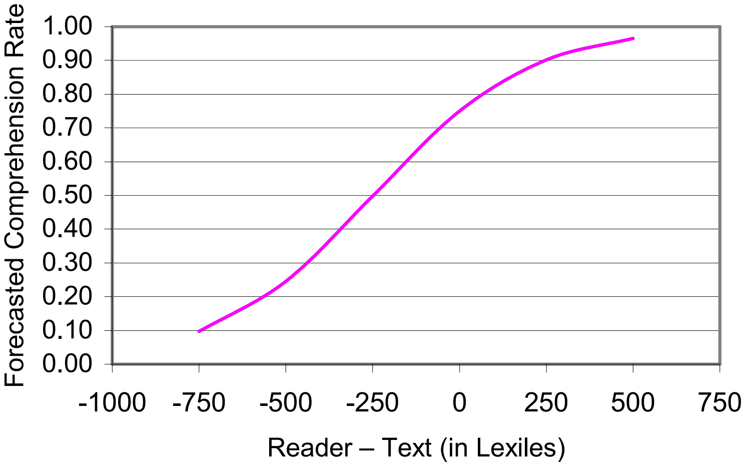
The 75-percent comprehension rate for a reader-text pairing can be given an operational meaning by imagining the text to be carved into item-sized slices of approximately 125-140 words with a question embedded in each slice. A reader who answers three-fourths of the questions correctly has a 75-percent 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. 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 1 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 1* and using +250L (reader minus text), the forecasted comprehension rate for this reader-text combination would be 90%.

Figure 1. 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 Grade 10 Textbook	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 media specialist, 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.

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.

Inline Reader/Writer Engine

Educators are profoundly aware of the need to assess a student's current reading ability and to monitor growth in reading over time. To be most effective, educators need a classroom-based assessment system that meets the following criteria:

- uses a single metric to estimate students' current status and monitors growth within and across years,
- aggregates the necessary number of performances to predict outcomes on future high-stakes testing,
- integrates assessment and instruction so that testing experiences use text based on what is being taught in the classroom to estimate readers' ability,
- reports in a metric common to those used on standardized norm- and criterion-referenced tests of reading ability, and
- informs day-to-day and year-to-year instructional decision-making, both in reading and in content area instruction.

Currently, the limited number of classroom-based progress monitoring tests falls short of achieving all of these goals. Clearly, educators' effectiveness in improving students' reading and content knowledge is enhanced by using an assessment system that attains each of the goals.

The Inline Reader/Writer (IRW) engine was developed by MetaMetrics, Inc. to meet the criteria for an effective assessment system identified by teachers. The genesis for IRW was comments from educators who desired an automated way to: (1) audit students' completion of well-targeted reading assignments, (2) assess their level of performance, and (3) encourage and facilitate writing by students. The first version of IRW was designed to audit students' completion of reading assignments by recording the total amount of time spent on reading assignments. Performance was monitored by requiring students to complete auto-generated embedded cloze tasks and writing a summary of the assigned reading. MetaMetrics conducted two pilot tests of IRW (version 1) to determine (a) the feasibility of its use in the classroom and (b) the satisfaction with using the Web-based application by educators and students. The beta test of IRW occurred in two phases. EBSCO provided digitized articles for use in the first version of IRW.

Description of the Inline Reader/Writer Engine and Application

The Inline Reader/Writer engine is a Web-based application developed to be used by educators and students. Teachers can assign, or students can select, text to be read based on what is being taught in the classroom and the match between student readability and text readability. Students read the assigned or selected text and respond to embedded cloze items. The text is "clozed" by taking out certain words and replacing them with a blank line. For research conducted by MetaMetrics, students complete the cloze by selecting the best response from four choices. Words selected to complete the cloze are selected at the same difficulty level as that of the text. For example, in the article *Listening to the Wildlife in the Everglades* (1230L), words with Lexile measures within 150L of the text measure are identified (e.g., 1080L to 1380L). Students are presented with the word from the text and three additional words that are syntactically correct, and are asked to identify the best word to complete the sentence given the context of the passage.

The cloze-item format has been shown in multiple studies to measure the same reading construct as norm- and criterion referenced tests (Stenner, Horabin, Smith, Smith, 1988). In addition, the cloze procedure has been shown to require more re-reading of the passage and an increase in the use of context clues, both characteristics of better readers. A clock runs in the background that is used to estimate total time spent reading (i.e., silent fluency) and time spent answering the embedded sentence cloze items (i.e., response latency).

Finally, students are also asked to write a summary of the text they just read. The summary is conceptualized as an additional facet of reading. Students may submit their summary for analysis and suggestions from spelling and grammar checkers. Specifically, students receive feedback in relation to the following areas: (1) *grammar*, including subject-verb agreement, ill-formed verbs and missing possessive errors; (2) *usage*, including confused words and nonstandard verbs or words; (3) *mechanics*, including missing capitalization, end punctuation and internal punctuation, and (4) *style*, including sentences beginning with coordinating conjunctions and sentences with passive voice.

The following specific features of the IRW application are fundamental to conducting research related to psychometric properties:

1. **Student management suite** – Educators enter (or upload from the school's or district's student information management system) key information about each participant into the student information management tool at the initiation of the study, including a unique identification number, gender, birth date, grade, and ethnicity. In addition, key outcomes from each student's IRW experience are maintained here.

2. **Content-relevant assessment experiences** – Educators assign or students select articles from the article database using topical keywords based on the content currently being taught, the student's estimated Lexile measure of ability, and the Lexile measure of text. This match between student and text Lexile measures insures well-targeted reading and differentiated instruction.

3. **Assessment environment** – Educators may allow students to engage in their assessment experience at a time and location that best meets the needs of the teacher and student. Because the assigned reading task is stored electronically on a Web site that teachers and students can access from any Web-enabled laptop or personal computer, assessment can be conducted at any location.

4. **Real time data collection** – IRW collects data during each assessment experience for the following elements hypothesized to contribute to a valid estimate of reader ability:

   - *Silent fluency* – number of words read per minute (total number of words read between clozes divided by the time spent reading those words),
   - *Response latency* – time spent responding to each type of embedded cloze item, four choice response or prompted production items,
   - *Count correct* – total number of cloze items answered correctly (four-choice response), and
   - *Written summary* – a rating of students' production of what was comprehended in writing.

The Lexile Framework for Reading: A Conceptual Framework that Helps IRW Bridge Assessment and Instruction

IRW was developed based on the five fundamental concepts contained in The Lexile Framework. First, a reader's ability is posited to be a *general trait* estimated from performance on a test or instructional experience in which count correct data are collected. Second, text readability is a trait. Third, a reader's comprehension is a *state* that results from his or her construction of meaning from local text (i.e., materials specific to a reading assignment such as a chapter in a textbook or an article from a periodical). Fourth, a reader's expected comprehension rate is modeled as the difference between a reader's ability and a text's readability which are both estimated in Lexiles. Fifth, well-targeted reading experiences occur when a readers' ability and the readability of text are matched (i.e., text is within +50L to –100L of student's reading ability). The match assures that readers achieve a sufficient rate of comprehension given a good match between readers and the semantic and syntactic features of text (65% to 80% comprehension) (Moats, 2000).

The Lexile Framework purports to measure in a common unit (Lexiles) the semantic and syntactic components of the written English symbol system. The Lexile scale allows for the accurate measurement of reader ability, text readability, task demand, and rater/observer severity (in its simplest expression, a single multiple-choice task type is used, thus eliminating the need for a rater/observer parameter). With task type restricted to the native-Lexile format, the Lexile Framework reduces the semantic and syntactic features of written language to a text facet which estimates text readability and a reader facet which estimatess reader ability to understand written language. Therefore, text can be ranked in terms of its increasing difficulty (i.e., increasing semantic difficulty and syntactic complexity) and students can be placed on a developmental continuum based on their ability to comprehend increasingly difficult text. One important benefit of the Lexile Framework for educators is the precise matching of student with text that will occur because text readability and student reading ability are placed on the same developmental scale. Placing students and text on the same scale also benefits educators because they can model reading comprehension with a high degree of precision. Thereby, allowing classroom teachers, parents or caregivers, and students to select text with a high degree of confidence that the text will likely be understood by the reader. This process of matching students with text allows educators to manage reading comprehension.

Expected comprehension rate is modeled as the difference between reader ability and text readability. Text readability and reader ability are defined in terms of one another. A text's readability (e.g., *Harry Potter and the Goblet of Fire*, 880L) is defined as the amount of reader ability required to answer 75% of the reading items correctly *if* the book were turned into a long reading test. Thus, we imagine that each of the, say, 1,000 paragraphs in the Harry Potter novel were turned into a native-Lexile reading item. The result would be a 1,000-item reading test with each item possessing a calibration given by the Lexile Analyzer.

A Rasch equation is used to answer the question, "What Lexile reading ability is needed to answer correctly 750 of the 1,000 native-Lexile reading items comprising the Harry Potter novel?" The answer is that a reader ability of 880L is needed for 75% comprehension (750/1000) of the Harry Potter novel. Similarly, reader ability is defined in terms of the text that a reader can read with 75% comprehension. A 1200L reader can read text like *USA Today* (1200L) with 75% comprehension and would have 92% comprehension of a Harry Potter novel, but would have only 60% comprehension of the typical College Board SAT text (1330L). A useful habit is when given a text measure (880L), imagine what reader characteristics match this text readability (e.g., a sixth grader on level for grade). Similarly, when given a reader measure (1200L), imagine the texts that this reader can read with 75% comprehension (*Cold*

Mountain, 1210L; *The Trumpeter of Krakow*, 1200L). The complementary relationship of reader and text pivoting on the choice of 75% comprehension makes the Lexile Framework compelling. The realization that differences in text readability can be traded off for differences in reader ability while holding comprehension constant is a protean concept with important implications for reading theory and practice.

The Many-Facet Rasch Model: A Measurement Perspective Applied to the Assessment of Reading

The objective in reading is to define simultaneously the meaning of five concepts: (1) amount of reader ability, (2) degree of text readability, (3) reader comprehension rate, (4) amount of task demand, and (5) amount of rater severity (Stenner & Wright, 2004). Much contemporary reading theory and research focuses only on reader ability (see *Handbooks of Reading Research*, 1984, 1996, and 2000). The model proposed for research related to IRW posits a many-facets approach to reading which generalizes the two-facet model (Linacre, 1987; Rasch, 1980). This approach enables the measurement of readers, texts, tasks, and raters on a common scale. If data fit this many-facet Rasch model, then (a) each facet is estimated independently of the other facets, (b) additive conjoint measurement of each facet is obtained, and (c) differences in measurements on one facet can be traded off for equal differences on other facet(s) to hold the probability of success constant.

A many-facet Rasch model is specified below assuming dichotomous observations. The assumption of dichotomous data (1, 0) serves only to simplify the presentation. As noted elsewhere, a full range of data types including partial credit, rating scale, binomial counts, and counts per unit of time, among others, can be accommodated in this model:

$$\log \frac{p_{ntar}}{1 - p_{ntar}} = B_N - T_i - A_j - R_k \qquad \text{(Equation 2)}$$

where

p_{ntar} is the probability of reader *n* answering a question correctly on text *i* under response requirement (e.g., task) *j* as rated by rater *k*.
B_N is the ability of reader *N*.
T_i is the readability of text *i*.
A_j is the demand of task type *j*.
R_k is the severity of rater *k*.

The lack of interaction terms in this formulation is deliberate and in keeping with the requirements of conjoint additive measurement. The model requires that the data must be demonstrably additive in structure (i.e., quantitative) if the data are to be useful in making reader measures. Additive conjoint measurement (Luce & Tukey, 1964) is a formal statement of the measurement model behind all derived measurement in physics (e.g., temperature, density, viscosity). The many-facet Rasch model is an example of conjoint measurement applied in a probabilistic (i.e., stochastic) framework.

The many-facet Rasch model provides a direct test of the axioms underlying conjoint additive measurement (Perline, Wright, & Wainer, 1979). When the axioms hold, each facet (reader, text, task, and rater) can be measured on a common interval scale. Most importantly, differences between facet pairs can be computed. For example, “comprehension rate” can be modeled as a function of the difference between reader ability and text readability. There are numerous philosophical, mathematical, and practical implications of the way that substantive reading theory and the many-facet Rasch model are combined in the Lexile Framework for Reading.

Inline Reader/Writer Research

IRW has been developed through a series of research studies ranging from small-scale feasibility studies to school-wide implementations in middle and high schools. The following sections describe the specific research studies conducted during the development of version 1 of IRW.

Study 1: Feasibility Study. The first pilot test was designed to determine the feasibility of IRW in a small number of classrooms. The goal of the study was to determine if a finite number of articles (30) from EBSCO could be delivered to readers (adults and students) via a Web-based version of IRW. Articles utilized in IRW ranged from 520L to 1490L. The primary interest of this study was to ascertain the impact of technical barriers that might be faced due to the wide array of computers and operating systems used in today's classrooms. A secondary interest of this study was the extent of educator's and student's satisfaction with their IRW experience.

The beta test was conducted from mid-October through mid-December 2002. Approximately, 80 educators were contacted to participate in the initial phase of the beta test. Six classroom teachers and eight students took part (see *Tables 6* and *7*) in the initial study.

In general, few technical obstacles were identified that could not be solved by e-mail or telephone support. The most common technical problem cited by educators was difficulty using IRW within Netscape. No educators or students who used Internet Explorer reported difficulty using IRW. Users of Netscape faced no difficulty after technical assistance was provided. In general, educators and students strongly agreed or agreed with the ease of use and relevance of IRW to classroom assignments.

Table 6. Teacher Inline Reader/Writer evaluation results (*N* = 6).

	Strongly Disagree	Disagree	Agree	Strongly Agree
1. My students were more on task while they used Inline Reader/Writer compared to when they read textbooks or magazines in class.			100%	
2. The students said they enjoyed using Inline Reader/Writer.		50%	50%	
3. Pictures need to be included in the articles.		17%	50%	33%
4. The directions for the reading section were easy to understand.			67%	33%
5. The report will be an effective tool for communicating with parents about the progress of their child.			67%	33%
6. Students were able to write summaries with little difficulty.		33%	67%	
7. The directions for the summary writing section were easy to understand.		17%	83%	
8. I understood the report.		17%	17%	50%
9. I would use Inline Reader/Writer 2-3 times per week with my students.		50%	17%	33%
10. The fill-in-the-blanks helped me to concentrate on what I was reading.		17%	50%	33%
11. I would like to be able to write in more feedback about the summaries.	17%	17%	17%	50%
12. My students used Inline Reader/Writer with little assistance from me after their first use.			67%	33%
13. Inline Reader/Writer will help me to differentiate instruction in my class (after more content articles are in the database).			67%	
14. I would like the report to include instructional recommendations based on the results.			33%	67%
15. Inline Reader/Writer will help with reading and writing across the curriculum.			33%	67%

Table 7. Student Inline Reader/Writer evaluation results (*N* = 8).

	Strongly Disagree	Disagree	Agree	Strongly Agree
1. Inline Reader/Writer was easy to use.		13%	63%	25%
2. The directions were clear.		13%	38%	50%
3. Pictures need to be included in the articles.	13%	38%	25%	25%
4. I was motivated to use Inline Reader/Writer.		13%	75%	13%
5. I understood what I was reading.		25%	13%	50%
6. Writing the summary was not too difficult.		13%	88%	
7. I could have written a good summary in less than 100 words.	25%	25%	25%	25%
8. I understood the report.	13%	25%	38%	13%
9. I would like to use Inline Reader/Writer 2-3 times per week.	13%		38%	50%
10. The fill-in-the-blanks helped me to concentrate on what I was reading.	13%	13%	50%	25%
11. The feedback on my written summary was useful.		13%	38%	38%
12. Inline Reader/Writer would help improve my reading if I used it 2 – 3 times per week.			50%	50%
13. Writing the summary was easier to do because of the fill-in-the-blank tasks.		38%	38%	13%
14. I am a good writer.		25%	38%	38%
15. I am a good reader.			25%	75%

Study 2: Large-Scale Feasibility Study. In 2003, a large-scale beta test of IRW was conducted from late March to mid May in Pinellas County (Florida) School District. Ten teachers and 153 students from grades three through eight participated in the study. Students completed 353 IRW experiences (read article, respond to cloze statements, and write a summary). EBSCO provided approximately 2,000 articles for use during this test of IRW. These articles were a subset of the total number of articles the Pinellas County School District licensed from EBSCO. The goals of this study were to determine the feasibility of IRW on a relatively larger scale (i.e., more students with multiple uses over a longer period of time) and to gauge educators and students satisfaction with IRW. Students spent an average of eight minutes reading each article and responding to the cloze statements (range from 4 minutes to 19 minutes) and spent an average of 23 minutes writing a summary of the completed reading assignment.

Teachers in the study took part in a focus group upon completion of the beta test. Classroom teachers reported the following strengths for IRW:

- Simple task to select and assign reading.
- Improved ability to differentiate instruction.
- Effective linkage of assessment results and instruction.
- Ease of tools to audit completion and comprehension of reading assignments.
- Enhanced student motivation for reading and completing reading assignments.
- Increased time on task.

- Decreased behavior problems in the classroom.

Two middle school teachers ceased using their reading curriculum in favor of having students use IRW to read topical articles targeted to each student's reading ability. Teachers explained that IRW provided students with articles targeted to their reading ability that were more content-focused and relevant to the instruction being delivered in the classroom. All of the teachers interviewed (elementary and middle school) commented that students were spending more time on task, especially writing. Teachers noted that students who typically wrote one- or two-sentence summaries were now writing summaries of a length recommended in IRW. Spending more time engaged in a task usually avoided was viewed by the educators as a very favorable outcome

Students also participated in focus groups. Students were excited to talk about their experiences with IRW. The following comments describe their experience:

- Filling-in-the-blank helped me pay attention.
- Easy to use.
- I liked it more than reading from a book.
- I could do it from home.

Summary of Feasibility Studies. The results of these studies suggested that IRW provides educators with the means to link assessment and instruction. More importantly, the application enables educators to differentiate instruction. Teacher observations of students provided evidence that students were more motivated to read and more dedicated to the tasks of reading and writing when engaged in well-targeted reading assignments. This engagement results in a success spiral: Students engaged in content relevant reading assignments targeted to their reading ability enjoy greater success; and, this success results in more time spent reading and writing which in turn leads to substantive growth in reading ability and content knowledge.

Linking Inline Reader/Writer with the Lexile Framework for Reading

Reading Riches, a reading and writing motivation program, was implemented in two large school districts during the 2004-2005 school year. Reading and writing were assessed with Inline Reader/Writer technology. The goal of the study was to link the IRW item type (i.e., auto-generated embedded sentence cloze task) to the Lexile scale. The following section describes the version of IRW used by participants and the results of the linking study.

In the Reading Riches program, students and educators accessed key information about three of the pillars of reading ability: (1) fluency, (2) vocabulary, and (3) comprehension. Educators facilitated one-on-one, small-group discussions, conducted whole-group instruction about what students learned from the article or series of articles, discussed vocabulary words that were missed, and examined the relationship between time spent reading and choosing answers to complete the clozes and performance using the results collected from reading content relevant articles. The results of a single Inline Reader/Writer session or multiple sessions served as the basis for educators and students to discuss their performance in the context of reading content relevant text.

Inline Reader/Writer Linking Study. In the spring of 2005, MetaMetrics conducted a linking study to determine if IRW could produce Lexile measures and if so, check for what adjustments to measures the new item type required.

A sample of 1,498 students in grades 5 through 12 were administered both a Lexile linking test comprised of native Lexile items and an online administration of IRW passages targeted at 50%, 75%, and 90% comprehension for the typical student at each grade. There were two forms of IRW passages, one with conditioned items (Form B) and one with unconditioned items (Form A). Conditioning involved human interaction with the foils to edit out any perceived foibles caused by the computer algorithm. The unconditioned items were completely computer generated without any post editing. The specifications for each form are presented in *Table 8* below.

Table 8. IRW linking study test form specifications.

Grade	Form	Article 1		Article 2		Article 3	
		Lexile	# of items	Lexile	# of items	Lexile	# of items
5	A	500L	5	810L	15	900L	12
	B	520L	8	800L	16	890L	8
6	A	650L	12	900L	7	1100L	8
	B	680L	6	900L	10	1100L	8
7*	A	780L	8	1000L	8	1200	11
	B	770L	9	1000L	11	1200L	10
8	A	800L	10	1100L	13	1300L	8
	B	869L	14	1100L	9	1300L	10
9/10	A	950L	9	1200L	13	1400L	8
	B	950L	11	1200L	9	1300L	10
11/12	A	1000L	8	1250L	10	1450L	9
	B	1000L	12	1250L	8	1440L	9

A computer error was discovered in the implementation of the Grade 7 unconditioned form, therefore, Grade 7 data was removed from further analyses. The first analyses examined the point measure correlations to see if the IRW items were performing as expected. The correlations presented in *Table 9* were lower than have been observed in the past for other reading item types, even when controlled for potential artifacts like range restriction (MetaMetrics, 1999a, 1999b, 2000, 2006b).

Table 9. Mean point-biserial and point measure correlations from various reading research studies.

	Date	Number of Items	Point Measure Correlation	Point-Biserial Correlation
PASeries Reading	2004	342	0.42	
Native-Lexile items (Duval, FL)	1999	427		0.42
Native-Lexile items (Miami, FL)	1999	300		0.42
IRW	2005	280	0.33	
Lingos Vocabulary Assessment	2000	65		0.39

The lower correlations for IRW are to be expected from a computer-generated item type. The correlations for the conditioned and unconditioned IRW items were similar with the unconditioned items being slightly higher at .33 (.32 for the conditioned form). Given the lower point biserials, however, the person data from the linking study was examined to determine if the lower item performance actually affected person measures. Most of the 1,498 students in the study took two native-Lexile tests, one in the winter, one in the spring, as well as the IRW field study form. This design enabled a within-grade "roundabout" to be employed to determine if the correlations between tests comprised of IRW items produced measures as highly correlated with native-Lexile item tests as two native-Lexile item tests are correlated with each other.

Table 10. Descriptive statistics for test forms in linking study (IRW forms standardized to 45 items on a form).

	Native to Native		Native to IRW (Unconditioned Form)		Native to IRW (Conditioned Form)	
Grade	Correlation	*N*	Correlation	*N*	Correlation	*N*
5			.80	46	.76	38
6	.59	197	.59	91	.35	106
8	.73	169	.51	86	.57	83
9	.81	331	.82	295	.77	290
10	.76	254				
11	.66	140	.66	130	.63	140
12	.73	130				
Mean	**.72**		**.68**		**.62**	

Only in Grade 8 does the IRW unconditional item correlation not compare to the native item correlation. In Grade 5, only one native form was administered and therefore there is no native to native correlation with which to compare; but, a correlation of .80 for the native to IRW unconditional is high for a within-grade raw score correlation. These results support a premise that IRW items and native Lexile items measure the same construct, reading comprehension. The data suggests that measures produced by IRW items will link suitably to the Lexile scale despite the lower point-biserial correlations. With the conditioned items performing lower compared to the unconditioned items and with the fact that they require human intervention, conditioning IRW items was not considered necessary in further research.

Checking for theory fit of the items with the Lexile Theory was the last stage of the IRW item-validation process. The goal for the IRW engine is to produce items that can produce Lexile measures based on theory alone. Given the multiple-items-per-passage nature of IRW items, *PASeries* Reading's passage-native Lexile items provide the best interpretative framework for accessing how the theory is doing on IRW items. The results are presented in *Table 11* below.

Table 11. Comparison of RMSEs on passage means for various item formats.

Item Type	Number of Passages	RMSE	Within passage SD
Passage Native from *PASeries* Reading	42	151L	159L
IRW	15	152L	207L

The theoretical prediction for IRW items (RMSE 152L) is nearly identical to that of the passage natives (152L). The within-passage variability on the IRW items is much higher, but no adjustment to the Bayesian scoring algorithm when computing measures within IRW is made at this time. *Table 12* compares reader Lexile measures produced using theoretical Lexile measures with observed Lexile measures. The observed measures were anchored on the native Lexile items' theoretical measures.

Table 12. Difference in mean reading ability when computed by theory and when observed (weighted).

	Number	Mean	SD
Theory	15	1023L	268L
Observed	15	1041L	161L

Description of the Total Reader Assessment System

AS described in the introduction to this paper, Total Reader includes an area for students and an area for teachers and administrators. Students access Total Reader by navigating to the Total Reader Web site and logging in. Once students log in, they have the option, from the student home page, to navigate to one of the following three areas:

Reading Zone: This is the area where students take the diagnostic test and complete testlets to assess and update their Lexile measure.

Progress: Here students can see reports documenting their current Lexile measure, their performance on the most recent testlet completed, their Lexile history, and their passage history.

Suggested Reading: Once students have their Lexile measure, they can go to their Suggested Reading page to access information about books and other resources that match their current reading ability. The student's Suggested Reading page also includes a link to the Lexile Titles Database. Students can search this free database for texts by keyword, title, author, Lexile range, and other criteria. Students can identify books in their Lexile range that match their interests and then locate those books in a library or bookstore.

Test Administration and Scoring

Diagnostic Test. The first time students log in to Total Reader and enter the Reading Zone, they take a diagnostic test to determine their initial Lexile measure. Students must complete the diagnostic test within one uninterrupted session or they will have to take the diagnostic test again.

The diagnostic test consists of five age-appropriate native-Lexile items from the Lexile Item Bank (described on page 17). The native item format consists of a passage of text followed by an item written by the item author. The item consists of an embedded completion statement (stem) and four foils. For each item, the student is presented four possible substitutions at the bottom of the screen; by clicking a word, the student fills in the text with the best substitution and continues reading. The embedded completion statement is similar to the fill-in-the-blank format and should assess the student's ability to form a generalization based on the passage or draw an inference from the passage. The difficulty of the items for each grade level is centered at the 55th percentile (based on prior research by MetaMetrics, Inc.) and has a range of 400L.

This group of native-Lexile items provides a valid baseline Lexile measure. When the student completes the diagnostic, the student receives an initial Lexile measure. Students receive an initial Lexile measure based on the number of correct responses on the diagnostic test. Individual scores are calculated by first summing the number of correct responses (omitted items and multiple responses are counted as incorrect). The number correct is then converted to a Lexile measure using correspondence tables developed specifically for each level of the diagnostic test.

Testlets. Testlets are age-appropriate Lexile-leveled reading passages and associated embedded completion items. The embedded item format is a variant of the native item format and consists of a passage with multiple items within the passage. Items are embedded within the passage by computer (auto-generated). Each item consists of an embedded completion statement (stem) and four foils. The embedded completion statement is similar to the fill-in-the-blank format and should assess the student's ability to form a generalization based on the passage, draw an inference from the passage, or establish logical connections between the ideas in the passage. A unique characteristic of this item format is its emphasis on the *intersentential* relationship that exists between the item and the passage.

After completing the initial diagnostic, students access testlets from within the Reading Zone. There are two interfaces within Total Reader, one designed for younger students (grades 3-5) and one designed for older students (grades 6-12). Students can choose between fiction and nonfiction passages and then access a number of broad subject areas, such as Animals, Work and Careers, Family, and Environment. Each subject area contains age-appropriate titles. For example, the upper-level interface presents passages within a subject area called "Society and Social Issues," whereas the younger-level interface is called "Our Society." Students select a subject area and see a list of all the testlet titles within their Lexile range (+50L/-100L) within that particular subject area. Selecting a title allows students to read a brief, inviting introduction to the testlet. Students can then choose to take the testlet or return to the Subjects page to make another selection. Depending upon word count, testlets include between three to twelve cloze items.

Testlet passages have been carefully analyzed and organized into both Lexile and age-appropriate categories. Because of the multiple levels of categorization of passages, a 650L third grader looking in the "Cars and Motors" category will have different titles to select from than a 650L twelfth grader looking in the same category.

A student's preliminary Lexile measure is updated only after a student has completed a minimum of forty items and a minimum of three testlets using a Bayesian scoring algorithm (for more information, see the discussion beginning on page 43). This initial set of testlets must be completed within a four-week span. Thereafter, each time students complete new testlets, their Lexile measures are re-calculated, though the measures may not change with every testlet. As their Lexile measures increase, students access progressively more challenging passages. To prevent the chance of ever re-reading a passage, students may complete a testlet only once; after they've accessed the passage (whether or not they have completed all the items associated with it), the selection disappears from their list of choices.

Administration Time. Total Reader can be administered to individual students or groups of students (including entire classes). Each student using Total Reader must have access to an Internet-connected computer capable of displaying Macromedia Flash files.

A student's first Total Reader session requires that he or she complete the diagnostic test. It will typically take young students (grades 3 through 5) about 15 minutes to complete the diagnostic test. Older students (grades 6 through 12) should expect to spend about 10 minutes completing the same task. On subsequent sessions, a student will typically complete two testlets. Depending on the reading level of the student, such a session will typically last between 10 and 25 minutes. EdGate recommends that students complete no more than three testlets per week.

Total Reader Reports

User rights determine access to the different levels of reports. There are currently five different reports: student progress, class, targeted reading, school, and district. Each level of report is accessible to that level administrator or above (i.e., school reports are accessible by school, district, and state-level administrators only). Individual students may view their own reports. From all other reports, users can drill down to the individual student progress report.

The student *Progress Report* consists of basic profile information (name, grade, school, district, and class association), graphic representations of both Lexile progress and program usage, Lexile history, and passage history (with data on genre selection, passage scores, date taken, etc.).

Class Reports include detailed information on each teacher's class, including name, ID, grade, reading group (Advanced, Proficient, Needs Intervention), total passages completed, fiction/nonfiction ratios, benchmark (starting) Lexile measure, current Lexile measure, Lexile measure change, and whether the student is considered active in the system or not. The reports can be sorted by any of these categories and can be grouped by class. Reading groups are determined by the Lexile ranges of students, by grade, falling into the 50-75th percentile ranges for students on a national level (as determined by a MetaMetrics study). Advanced readers fall above this range, Proficient fall within, and those Needing Intervention fall below this range. The chart and Total Reader's groupings are meant to be a general guideline only; they do not correspond specifically with state assessments or NAEP cut-off scores. The purpose of grouping students into these three general categories is to allow teachers to see at a glance how their students compare with students, by grade, on a national level.

Targeted Reading Reports present suggested reading ranges based on each student's current Lexile measure (CL). Recommended readings are made for three difficulty ranges; Easy (CL from -200L to -100L), Targeted (CL from -100L to +50L), and Instructional (CL from +50L to +150L). These recommendations are consistent with The Lexile Framework for Reading and are intended to help teachers guide students to appropriate reading materials depending upon the specific reading task. These reports can be sorted by any of the heading categories and can be grouped by class.

School Reports are filtered by teacher and can be sorted by the following categories: grade, class name, total student count, percent using system, average/high/low Lexile measure, and percent showing improvement, decline, no change, and class ranking. These reports can also be grouped by teacher. School administrators can drill down to any class report; they can also access the Targeted Reading Reports.

District Reports include similar information to the School Reports, but they can be filtered by school and grouped by school or grade. They can also be sorted by each report category.

Interpreting Total Reader Scores

The Lexile Framework for Reading provides teachers and educators with tools to help them link the results of assessment with subsequent instruction. Tests such Total Reader that are linked to the Lexile scale provide tools for monitoring the progress of students at any time during the school year.

When a reader takes a Total Reader testlet, his or her results are reported as a Lexile measure. This means, for example, that a student whose reading skills have been measured at 500L is expected to read with 75-percent comprehension a book that is also measured at 500L. When the reader and text are matched (same Lexile measures), the reader is "targeted." A targeted reader reports confidence, competence, and control over the text. When a text measure is 250L above the reader's measure, comprehension is predicted to drop to 50 percent and the reader experiences frustration and inadequacy. Conversely, when a text measure is 250L below the reader's measure, comprehension is predicted to go up to 90% and the reader experiences total control and automaticity.

Total Reader scores can be used for a variety of purposes. First, teachers can examine scores from students in their class to help identify those at the very high and very low ends of the range. These students are most at risk for either not being challenged by the reading material in their classroom or not understanding the majority of the reading material. These students merit further specific evaluation. Second, teachers can examine student progress over time. Students' Lexile measures should, in general, increase over time. However, the rate of change may be gradual. For example, the average rate of progress for a middle school student is about 10 Lexiles per month. Therefore, it would not be unusual for a student's measure to remain static over several weeks.

Help Students Set Appropriate Learning Goals. Students' Lexile measures can be used to identify reading materials that they are likely to comprehend with 75% accuracy. Students can set goals of improving their reading comprehension, and plan clear strategies for reaching those goals, using literature from the appropriate Lexile ranges. Students can be re-tested during the school year to monitor their progress toward their goals.

Monitor Reading Program Goals. As a student's Lexile measure increases, his reading comprehension ability increases for given literature, and the set of reading materials he can comprehend at 75% accuracy changes. Many school districts are required to write school improvement plans in terms of measurable goals. Schools also write grant applications in which they are required to state how they will monitor progress of the intervention funded by the grant. For example, schools that are recipients of Reading Excellence Act funds can use the Lexile Framework for evaluation purposes. Schools can use student-level and district-level Lexile information to monitor and evaluate interventions designed to improve reading skills.

Measurable goals can be clearly stated in terms of Lexile measures. Examples of measurable goals and clearly related strategies for reading intervention programs might include:

> *Goal:* At least half of the students will improve reading comprehension abilities by 100L after one year of use of an intervention.
>
> *Goal:* Students' attitudes about reading will improve after reading 10 books at their 75% comprehension level.

These examples of goals emphasize the fact that the Lexile Framework is not an intervention, but a tool to help educators plan instruction and measure the success of the reading program.

Communicate With Parents Meaningfully to Include Them in the Educational Process. Teachers can make statements to parents such as, "Your child will be able to read with at least 75% comprehension these kinds of materials which are at the next grade level" or "Your child will need to be able to move 400-500 Lexiles to prepare for college in the next few years. Here is a list of appropriate titles your child can choose from for reading this summer."

Challenge the Best Readers. A variety of instructional programs are available for the poorest readers, but few resources are available to help teachers challenge their best readers. The Lexile Framework links reading comprehension levels to reading material for the entire range of reading abilities, and can help teachers identify age-appropriate reading material to challenge the best readers.

Studies have shown that students who succeed in school without being challenged, often develop poor work habits and unrealistic expectations of effortless success as adults. Therefore, even though these problems are not likely to be evidenced until the reader is beyond school age, providing appropriate-level curriculum to the best students may be as important as it is to the poorest reading students.

Improve Students' Reading Fluency. Educational researchers have found that students who spend a minimum of three hours a week reading at their own level for their own purposes develop reading fluency that leads to improved mastery. Not surprisingly, researchers have also found that students who read age-appropriate materials with a high level of comprehension also learn to enjoy reading.

Teach Learning Strategies by Controlling Comprehension Match. The Lexile Framework permits the teacher to target readers with challenging text and to systematically off-target students when the teacher wants fluency and automaticity (i.e., reader measure is well above text measure) or wants to teach strategies for attacking "hard" text (i.e., reader measure is well below text measure). For example, metacognitive ability has been well documented to play an important role in reading comprehension performance. Once teachers know the kinds of texts that would be challenging for a group of readers, they can systematically target instruction that will allow students to encounter difficult text in a controlled fashion. The teacher can model appropriate learning strategies for students, such as rereading or rephrasing text in one's own words, so that students can then learn what to do when comprehension breaks down. Then students can practice these metacognitive strategies on selected text while the teacher monitors their progress.

Teachers can use Lexiles to guide a struggling student toward texts at the lower end of the student’s Lexile range (100L below to 50L above his or her Lexile measure). Similarly, advanced students can be adequately challenged by reading texts at the midpoint of their Lexile range, or slightly above. Challenging new topics or genres may be approached in the same way.

Reader-focused adjustment of the experience also relates to the student’s motivation and purpose. If a student is highly motivated for a particular reading task (e.g., self-selected free reading), the teacher may suggest books higher in the student’s Lexile range. If the student is less motivated or intimidated by a reading task, material at the lower end of his or her Lexile range can provide the basic comprehension support to keep the student from feeling overwhelmed.

Targeting Instruction to Students' Abilities. To encourage optimal progress with the use of any reading materials, teachers need to be aware of the difficulty level of the text relative to a child's reading level. A text that is too difficult, then, not only serves to undermine a child's confidence but also diminishes learning itself. A text that is too easy fosters bad work habits and unrealistic expectations that will undermine the later success of the best students.

When students confront new kinds of texts, the introduction can be softened and made less intimidating by guiding the student to easier reading. On the other hand, students who are comfortable with a particular genre or format can be challenged with more difficult readability levels, which will prevent boredom and promote the greatest rate of vocabulary and comprehension skills.

To become better readers, students need to be continually challenged—they need to be exposed to less frequent and more difficult vocabulary in meaningful contexts. A 75% comprehension level provides an appropriate level of challenge, but not too challenging. If text is too difficult for a reader, the result is frustration and a growing dislike for reading. If text is too easy, the result is often boredom. Reading levels promote growth and literacy by providing the optimal balance. Reading just 20 minutes a day can be so important.

Apply Lexiles Across the Curriculum. Over 450 publishers Lexile their titles, enabling educators to link all the different components of the curriculum to more effectively target instruction. With a student's Lexile measure, teachers can connect him or her to tens of thousands of books (www.Lexile.com), and tens of thousands of newspaper and magazine articles (through periodical databases) that also have Lexile measures.

Using Lexiles in the Classroom

- Develop individualized reading lists that are tailored to provide appropriately challenging reading.
- Enhance thematic teaching by building a bank of titles at varying levels that not only support the theme, but also provide a way for all students to successfully participate in the theme.
- Use as an additional organizing tool when sequencing materials. For example, choosing one book a month for use as a read-aloud throughout the school year. In addition to considering the topic, increasing the difficulty of the books throughout the year is effective. This approach is also useful when utilizing a core program or textbook that is set up in anthology format. (Many educators find they need to rearrange the order of the anthologies to best meet their students' needs.)
- Develop a reading folder that goes home with students and comes back for weekly review. The folder can contain a reading list of books within the student's Lexile range, reports of recent assessments, and a parent form to record reading that occurs at home.
- Choose texts lower in the student's Lexile range when factors make the reading situation more challenging, threatening or unfamiliar. Select texts at or above the student's range to stimulate growth when a topic is of extreme interest to a student, or when additional support such as background teaching or discussion is provided.
- Use the free Lexile Book Database (at www.Lexile.com) to support book selection and create booklists within a student's Lexile range to help the student make informed choices when selecting texts.
- Use the free Lexile Calculator (at www.Lexile.com) to gauge expected reading comprehension at different Lexile measures for readers and texts.

Using Lexiles in the Library

- Labeling books with Lexile measures helps students find books of interest at their appropriate reading level.
- Comparing student Lexile levels with the Lexile levels of the books and periodicals in the library or media center helps educators to analyze and develop the collection to more fully meet the needs of all students.
- Using the free Lexile Book Database (at www.Lexile.com) to support book selection and create booklists within a student's Lexile range helps guide student reading selections.

Using Lexiles at Home

- Ensure that each child gets plenty of reading practice, concentrating on material within his or her Lexile range. Ask the child's teacher or school librarian to print a list of books in the child's range, or search the Lexile Book Database.
- Communicate with the child's teacher and school librarian about his or her reading needs and accomplishments. They can use the Lexile scale to explain their assessment of the child's reading ability.
- When a reading assignment proves too challenging for the child, use activities to help. For example, review the words and definitions from the glossary, and the review questions at the end of a chapter before the child reads the text. Afterwards, be sure to return to the glossary and review questions to make certain the child understood the material.
- Celebrate the child's reading accomplishments. One of the great things about the Lexile Framework is that it provides an easy way for readers to keep track of their own growth and progress. Children and adults can set goals for reading – sticking to a reading schedule, reading a book at a higher Lexile measure, trying new kinds of books and articles, or reading a certain number of pages per week. When the child hits the goal, make an occasion out of it!

Limitations of the Lexile Framework. Just as variables other than temperature affect comfort, variables other than semantic and syntactic complexity affect reading comprehension ability. A student's personal interests and background knowledge are known to affect comprehension. We do not dismiss the importance of the information communicated by temperature simply because temperature alone does not communicate comfort level of an environment. Similarly, the information communicated by the Lexile Framework is valuable, even though other information also enhances instructional decisions. In fact, the meaningful communication that is possible when test results are linked to instruction provides the opportunity for parents and students to give input regarding interests and background knowledge.

Relationship between Total Reader results and grade levels. Lexile measures do not translate specifically to grade levels. Within any grade there will be a range of readers and a range of materials to be read. In a fifth-grade classroom there will be some readers that are far ahead of the rest (about 250L above the typical reader) and there will be some readers that are far below the rest (about 250L below the typical reader). To say that some books are "just right" for fifth graders assumes that all fifth graders are reading at the same level. The Lexile Framework can be used to match readers with texts at whatever level the reader is reading. Just because a student is an excellent reader does not mean that he or she would comprehend a text typically found at a higher grade level. Without the background knowledge the words would not have much meaning. A high Lexile measure for a grade indicates that the student can read grade-level appropriate materials at a higher comprehension level (say 90%).

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% 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.

Total Reader Training

Training is integral to accurately and reliably administering Total Reader. On the Total Reader Web site, general information is provided for to help educators implement the program.

Flash Tutorial. A Flash-based tutorial, walking users (teachers and students) through different aspects of Total Reader, is available on the Total Reader site.

Documentation. A frequently asked questions document (FAQ) is available on the Total Reader Web site. Other training documents are available as well, including instructions on how to export student data into Total Reader, a Total Reader Reference Manual, and a list of tips for using Total Reader.

Total Reader offers a number of online and face-to-face training opportunities for teachers and administrators through the EdGate Training Center. These sessions include extensive hands-on experience with Total Reader, including taking testlets and using and analyzing student data from the system. Examples include: train-the trainer workshops, train-the-teacher workshops and online training. IN addition, online courses for credit have been developed. In these courses, teachers participate in discussions, complete assignments, and generate lessons for use in their classrooms over the course of a five-week session.

A key area in a successful product implementation is follow-up. Follow-up is a category of activities to extend, clarify, and enrich the learning experience. Follow-up activities may include:

- Train-the trainer and train-the teacher workshops
- Technology projects/lesson plans posted to EdGate resources
- EdGate Newsletters
- Online resources web site
- Online courses for credit

Development of Total Reader Testlets

The Total Reader testlets are designed to measure and monitor general reading ability. Testlet specification began during January and February 2004, with passage and item development following closely behind during spring and summer 2004. Finally, the computer interface for Total Reader and the operational materials were completed during summer 2004. Additional testlets have been developed and added to the system on a continual basis since summer 2004. Each of these stages will be described in detail in the followings parts of this section and in the sections to come.

Total Reader Reading Passage Development

Within Total Reader, testlets consist of age-appropriate reading passages at a specified Lexile level and associated cloze items. All passages in Total Reader testlets come from well-regarded children's magazines, popular novels, nonfiction books, and other reputable sources. The following sources were used to identify licensed and pubic domain passages:

- ProQuest Database: Magazines including *Boy's Life*, *Psychology Today*, *Triquarterly*, *Harper's Magazine*, *Scholastic News*, and more.
- Public Domain Text: Stories from children's magazines and excepts from well-known novels and short stories, such as *Dracula*, *The War of the Worlds*, "The Occurrence at Owl Creek Bridge," and more.
- Trade Book Publishers: Full-text of books from publishers such as Griffin Publishing and Abdo Publishing.
- Children's Authors: Commissioned and reprinted work from skilled children's book, magazine, and textbook writers.

Once content is identified, it is categorized according to the following criteria:

- Age-appropriate grade range (3 to 5, 6 to 8, 9 to 12, or across ranges).
- Lexile zone corresponding to the Lexile measure of the passage.
- Fiction or nonfiction genre.
- Self-contained, i.e., the reader should not need specific prior knowledge to understand the passage.
- Length: 200 to 1,200 words. For passages over 1,200 words, every effort is made to split the material into shorter pieces, with continuity retained through explanation in the frames. Pieces over 1,200 words are justified only if material is compelling enough, or if story flow would be unduly interrupted by being cut. The maximum word length is 1,500 words.

Table 13: Total Reader passage specifications.

Grade Range	Average Length (words)	Minimum Number of words	Maximum Number of words	Number of Items
3 to 4	300	200	500	2 – 6
5 to 7	490	250	730	3 – 9
8 to 10	600	350	850	4 – 10
11 to 12	725	450	1,000	5 – 12

All passages are copyedited to ensure that they adhere to standard written English conventions and style, are bias-free, and are self-contained. A frame (two to four sentences) is written for each passage. The frame is a short introduction to the story and is written to entice the reader to read the story. Reading passages (and their frames) use conventions appropriate for students at the targeted grade levels. Some fictional passages include non-biased colloquial expressions that are appropriate for the targeted grade.

The following guidelines were developed for determining the age-appropriate grade range for a passage.

- As a general rule, stories written about specific age groups should be targeted to students that age and younger. Exceptions may apply to passages containing content deemed suitable only for students older than those mentioned in the story (e.g., younger students like reading about older kids, but not vice-versa).
- High interest to students within particular grade spans.
- When appropriate, grades are assigned beyond these spans (e.g., 4-7).

When in doubt about the appropriateness of a given passage, editors are advised to ask another editor who has more teaching experience at that grade span.

The following guidelines are used to help ensure the creation of non-offensive and bias-free assessments. These guidelines were assembled from the results of MetaMetrics' collaboration with various partners in textbook and test publishing. The following content should avoided:

- A passage, or language in a passage, that is offensive to particular groups, minorities, etc.
- Passages condoning stereotypes.
- Sexually explicit or crude subject matter or content that promotes drug use.
- Violent or upsetting subject matter. Examples include:
    - graphic descriptions of malicious abuse or family abusive situations,
    - rape or explicit sex scenes,
    - graphic/detailed murder scenes, and
    - graphic/detailed drug use scenes.

- Content that might offend a particular religious or sectarian group.
- Passages containing a specific advertising pitch (excessively commercial content or promotion of a product or service). Note: brand names may be mentioned but should not be the focus of the passage.
- How-to or process-oriented passages (with extensive listing).
- Passages that refer to multiple Web sites or other addresses.
- Widely anthologized writing.
- Passages with substantial amounts of poetry or dialect.

Because material gleaned from the public domain can be either quite dated (pre-1923) or poorly edited, extra care is taken in the preparation of this material. Public domain material can be adapted. Adaptation consists of:

- Substantially revising the text to simplify the language.
- Updating obscure or archaic language and grammar.
- Shortening a passage by removing sections from the text body (in rare cases).

Any passages that are divided into parts (due to length) or excerpted from longer pieces adhere to the fundamental concept of test design called Source Targeting. The Lexile measure for each passage falls within 100L of the Lexile measure of the source text. This process uses information from the entire source of a reading passage to ensure that the estimated syntactic complexity and semantic demand of the passage are consistent with the "true" reading demand of the passage (longer text produces smaller uncertainty measure because the analysis is based on a broader sample of text).

The standard editorial process for passages consists of an initial screening for Lexile and age appropriateness, an online edit, a hard copy peer edit, and a testlet-creation/review phase. A final system review takes place once the passage has been loaded into the system. This final review insures that all items are functioning and all text is displaying properly within the Total Reader system.

Total Reader Item Development

The traditional cloze procedure for measuring reading comprehension is based on the deletion of every 5th to 7th word (or some variation) regardless of part of speech (Bormuth, 1967, 1968, 1970). It can also consist of selectively deleting certain categories of words. Selective deletions have shown greater instructional effects than random deletions (Greene, 2001). There is evidence to support that the cloze procedure reveals both text comprehension and language mastery levels. Some of the research done with metacognition shows that better readers use more strategies (and the appropriate strategy) when they read. The cloze procedure has been shown to require more rereading of the passage and an increase in the use of context clues.

The item format used with Total Reader is similar to the selective deletion format. This item 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. 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. This format is "well-suited for testing a student's ability to evaluate" (Haladyna, 1994, p. 62). In addition, this format is also useful as an instructional tool. In addition, this format is also useful as an instructional tool. All passages in Total Reader testlets come from well-regarded children's magazines, popular novels, nonfiction books, and other reputable sources.

The embedded item format is a variant of the native item format and consists of a passage with multiple items within the passage. Each item consists of an embedded completion statement (stem) and four foils. From the four foils, the reader is asked to select the "best" foil that completes the statement. Items should be written so that the correct response is not stated directly in the passage, but requires comprehension. The correct answer will not be suggested by the item itself. Rather, the examinee must determine the correct answer through comprehension of the passage.

Logic written into the Total Reader engine uses a real-time iterative process to automatically select words to cloze. The Total Reader engine begins by parsing, segmenting, and identifying words contained in the passage. This process involves breaking the text into word types that are recognized by a part-of-speech "tagger." Each word and its surrounding words are analyzed and their part-of-speech identified. Words are then tagged with their associated Lexile measure using a lookup dictionary. The Lexile measure of the clozed word from the passage comes from the LVA word list compiled by MetaMetrics based on vocabulary research related to determining the difficulty of words. The Lexile Vocabulary Analyzer (LVA) determines the Lexile measure of a word based on text difficulty using a set of features related to a word and its presence in the MetaMetrics' corpus (MetaMetrics, 2006a). The rationale was that these words should be part of a reader's "working" vocabulary since they had likely been learned while reading easier (lower Lexile measure) text. The goal is to have cloze items targeted to the Lexile measure of the source text, no less than 10 words apart (i.e., minimum distance), and no more than 70 words apart (i.e., density).

The Total Reader engine was developed such that the item writer can specify the minimum distance between clozes. The goal is to have one item per 100 words of text. Because of the nature of public domain material, the text distance is set for every 60 or 80 words. The goal is to have 7 to 13 cloze items per 1,000 words [see *Table 13* for the specific number of items developed for various grade ranges]. No token is selected for cloze in the first 80 words of the passage. This allows students to orient themselves to the content, retrieve relevant background knowledge, and develop a flow to the reading task. Three semantically and syntactically appropriate foils at the same Lexile level (based on difficulty in the LVA word list) of the word that is clozed (i.e., the correct word to complete the sentence) are provided to students as they read the passage.

Once a passage is passed through the Total Reader engine and items are auto-generated, content experts review the testlets to verify the content marking hierarchy, the Lexile measure of the passage, the display of the title and copyright information, the appropriateness and display of the frame, and the functionality of each of the items generated within the Flash engine. When necessary, items are revised by hand. If the auto-generated items as a set do not function well, then the passage is passed through the Total Reader engine again and the items are reviewed.

Scoring and Reporting

The two main purposes of Total Reader are to measure student reading comprehension and to monitor growth in reading comprehension throughout the school year. In order to meet these goals, a developmental scale must be used to report the results. The Total Reader testlet assessments are scaled to the Lexile scale. The Lexile scale is described in this report in the section entitled The Lexile Framework for Reading. This section describes the procedures and the analyses used to score and report the Total Reader testlets.

Scoring Total Reader Testlets

The Bayesian Paradigm—Basic Principles. Bayesian methodology provides a paradigm for combining prior information with current data, both subject to uncertainty, to come up with an estimate of current status, which is again subject to uncertainty. Uncertainty is modeled mathematically using probability.

Within Total Reader, when a student is ready to encounter an assessment instrument, i.e. testlet, the prior information can come from the student's grade level, the initial diagnostic test, or a Lexile measure from a previous assessment. The current data in this context is the performance on the test, which can be summarized as the number of items answered correctly out of the total number of items attempted.

Both prior information and current data are represented via probability models reflecting uncertainty. The need for incorporating uncertainty when modeling prior information is intuitively clear. The need for incorporating uncertainty when modeling test performance is, perhaps, less intuitive. Once the test has been taken and scored, and assuming that no scoring errors were made, the performance, i.e. raw score, is known with certainty. Uncertainty arises because test performance is associated with but not determined by the ability of the student and it is that ability, rather than the test performance per se, that we are endeavoring to measure. Thus, although we are certain about the test performance once the results are in, we remain uncertain about the ability that produced the performance.

The uncertainty associated with prior knowledge is modeled by a probability distribution for the ability parameter. This distribution is called the prior distribution and it is usually represented by a probability density function (e.g. the normal bell-shaped curve). The uncertainty arising from current data is modeled by a probability function for the data when the ability parameter is held fixed. When roles are reversed so that the data are held fixed and the ability parameter is allowed to vary, this function is called the likelihood function. In the Bayesian paradigm, the posterior probability density for the ability parameter is proportional to the product of the prior density and the likelihood, and this posterior density is used to obtain the new ability estimate along with its uncertainty.

Modeling Growth and its impact on the prior. When no prior assessments are available for a student, the prior distribution can be inferred from the student's grade. Once a posterior has been obtained from current data, that posterior can serve as the prior for an immediate repeat assessment. If a substantial amount of time has passed since the last assessment, however, then allowance should be made for an uncertain amount of growth in reading ability since the last assessment. This allowance is accomplished by means of a growth model, which estimates as a function of elapsed time both the growth in reading and the augmentation in uncertainty.

Bayesian Scoring Process—Overview of Flow.

1. *Get old values*. Student comes into the Reading Zone with a Lexile measure from a previous testlet.

   If no previous Lexile measure is available, then use default values from the grade-level file. Prior ability estimates for the diagnostic tests are set at approximately the 50th percentile based on prior research with the Lexile Framework**.**

2. With either a prior test score from the diagnostic test or the grade-level file, assume the maximum prior uncertainty of 225L.

3. *Compute new values*. This step combines the prior information with the data from student performance on the prior test to produce a posterior density. This value is used to create the new Lexile measure for the student. The new Lexile measure for the student will be incorporated into the prior information for the scoring of subsequent assessments. Within Total Reader, the subsequent assessment after the diagnostic test administration is the first administration of a testlet. For each subsequent administration of a testlet, all of the information on the student's reading ability from the previous test administrations is incorporated into the student's prior.

4. A student's updated Lexile measure is reported only after he or she has completed a minimum of 40 items and a minimum of three testlets. The initial set of testlets must be completed within a four-week administration window. Thereafter, each time a student completes a testlet, his or her Lexile measure is reported. Total Reader automatically offers the student progressively more challenging testlets as his or her Lexile measure increases.

Conditions.

1. Use the grade-level priors of the grade-level of the student.
2. Negative growth (negative differences in days since last test) is not permitted. If a student takes a test that is not scored and then takes another test either (1) the first test should not be scored or (2) the first is scored and the second test is re-scored. If the first test is scored, the information will need to be used as the priors for the second test when re-scoring. Zero time (i.e., tests taken on the same day) will follow the standard process. Zero time means that sigma old will be automatically used as sigma update.
3. Changes in answer key and item difficulty should result in a re-score of the test impacted. All tests taken after that test will need to be have the Bayesian Score and forecast measure recalculated.

Reporting Total Reader Testlets

Conventions for Reporting. 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 person measures should be rounded to the nearest 5L to avoid over interpretation of the measures. If the Lexile measure is xxx2.5 or higher or xxx7.5 or higher, round up to the next higher 5 Lexiles. For example, if the Lexile measure from the program is 572.51, it should be reported as 575; if the Lexile measure from the program is 577.42, it should be reported as 575. As with any test score, uncertainty in the form of measurement error is present.

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 lower and upper bounds of measurement error (e.g., at the 10th and 90th percentile points). In an instructional environment, all scores at or below 0 should be reported as "BR" (Beginning Reader); no student should receive a negative Lexile measure.

Reliability

If use is to be made of some piece of information, then the information should be reliable—stable, consistent, and dependable. In reality, all test scores have some error (or level of uncertainty). This uncertainty in the measurement process is related to three factors: (1) the statistical model that was used to compute the score, (2) the questions that were used to determine the score, and (3) the condition of the reader when the questions used to determine the score were collected. Once the level of uncertainty in a test score is known then it can be taken into account when using the test results.

Reliability, or the consistency of scores obtained from an assessment, is a major consideration in evaluating any assessment procedure. Two sources of uncertainty has been examined with Total Reader—reader error and text measure error.

Standard Error of Measurement

Uncertainty and Standard Error of Measurement. Because of the presence of measurement error associated with test unreliability, there is always some uncertainty about a student's true score. This uncertainty is known as the standard error of measurement (SEM). The magnitude of the SEM of an individual student's score depends on the following characteristics of the test:

- the number of test items—smaller standard errors are associated with longer tests,
- the quality of the test items—in general, smaller standard errors are associated with highly discriminating items for which correct answers cannot be obtained by guessing, and
- the match between item difficulty and student ability—smaller standard errors are associated with tests composed of items with difficulties approximately equal to the ability of the student (targeted tests)

(Hambleton, Swaminathan, and Rogers, 1991).

Total Reader was developed using the Rasch one-parameter item response theory model to relate a reader's ability and the difficulty of the items. There is a unique amount of measurement error due to model misspecification (violation of model assumptions) associated with each score on the assessment. The computer algorithm that controls the administration of the assessment uses a Bayesian procedure to estimate each student's reading comprehension ability. This procedure uses prior information about readers and students to control the selection of questions and the recalculation of each student's reading ability after responding to each question.

Calculating Initial Uncertainty for an Individual. Assume that maximum uncertainty is 1.1056 (within grade standard deviation in logits from previous research). The assumption is that after three years uncertainty about a student’s ability is again at the maximum level. Time is measured in days and σ_{old} is measured in logits. The formula for updating the uncertainty estimate for an individual is

$$\sigma_{update} = \frac{(1.1056 - \sigma_{old})(t_2 - t_1)}{1095.75} + \sigma_{old} \qquad \text{(Equation 3)}$$

Example: If the previous uncertainty (σ_{old}) = 0.2778, then the difference is 0.8278. This value is divided by 1095.75 days (number of days in three years) and the change in certainty is 0.000755 per elapsed day since last assessment.

Updating Uncertainty. Average SEMs are presented in *Table 14*. The values can be used as a general guideline when looking at the results of Total Reader. It bears repeating: *because each student's results are based on prior information, the error associated with any one score or student is also unique*.

Table 14. Ranges of uncertainty for various levels of prior uncertainty.

Prior Sigma (in Lexiles)	Updated Sigma (in Lexiles)
10	9-10
20	19-20
30	27-30
40	35-40
50	37-50
100	64-91
300	81-195

As can be seen from the information in *Table 17*, when the test is well targeted (prior sigma is small), the student can respond to fewer test questions and not increase the error associated with the measurement process. When only the grade level of the student is known, the more questions the student responds to, the less error in the score associated with the measurement process.

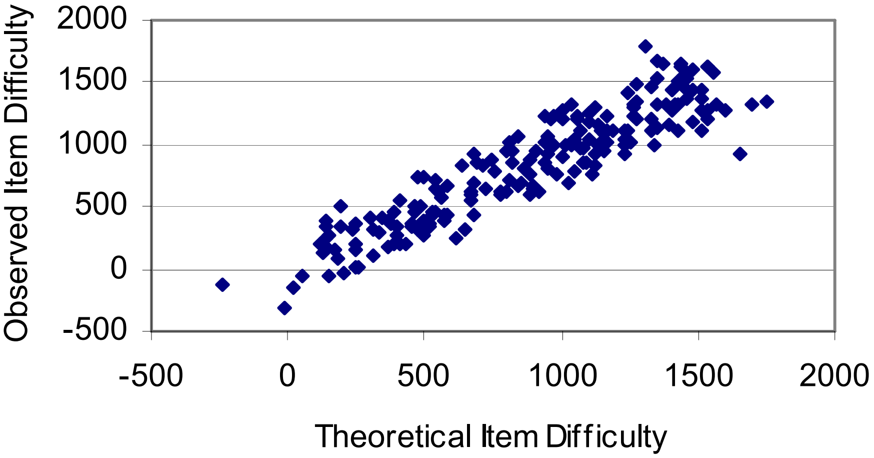
Text Measure Error Associated with the Lexile Framework

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 scatter plot of the data is presented in *Figure 2*.

Figure 2. 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 15*).

Table 15. 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 study was conducted during 2002 to examine ensemble differences across items (Stenner, Burdick, Sanford, and Burdick, 2006). 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 16 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 16. 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 $(T\text{-}O)^2$ = 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 3a* and *3b* shows a plot of observed ensemble means against theoretical text calibrations.

Figure 3a. Plot of observed ensemble means and theoretical calibrations (RMSE = 111L).

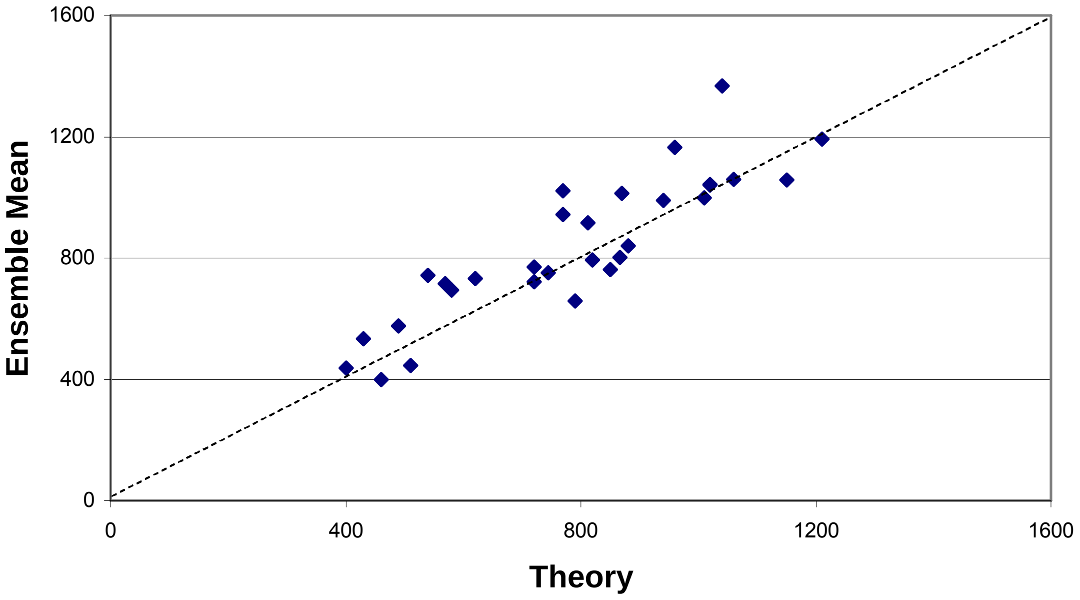
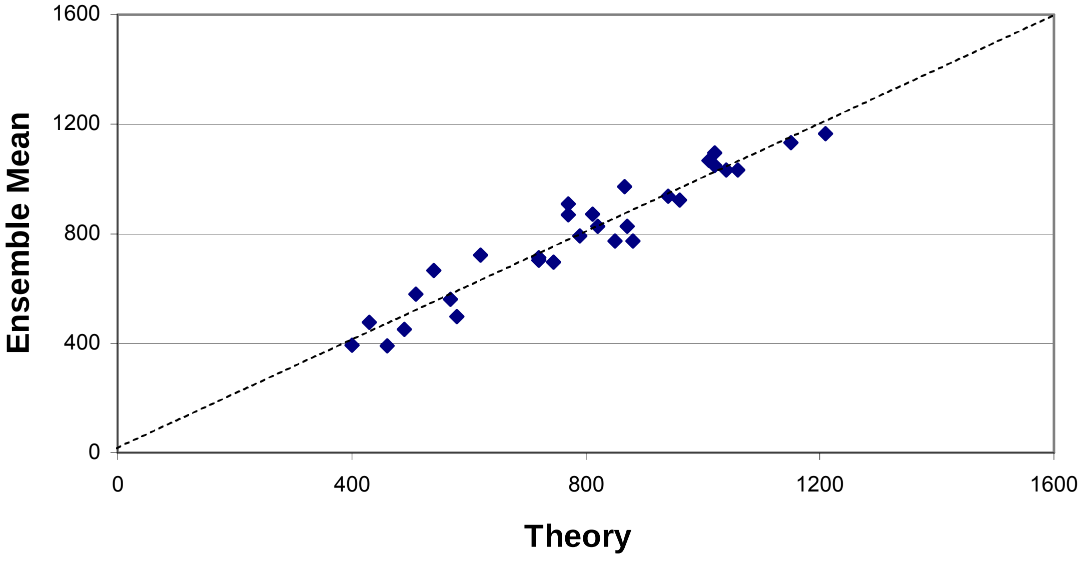


Figure 3b. Plot of simulated "true" ensemble means and theoretical calibrations (RMSE = 64L).



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 3* depicts simulated data when an error term [[distributed ~*N*(μ = 0L, σ = 64L)] is added to each theoretical value. Contrasting the two plots in *Figure 3* 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 ($\sqrt{MSE_{Total} - MSE_{Within}} = \sqrt{12,022 - 7,984}$ = $\sqrt{4,038}$ = 64). 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 = $64 \div \sqrt{4}$ = 32L, whereas a much longer text like the novel *Harry Potter: Chamber of Secrets* (880L, Rowling, 2001) would have a SEM = $64 \div \sqrt{900}$ ≈ 2L. *Table 17* 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 17: 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.

Validity

The validity of a test is the degree to which the test actually measures what it purports to measure. Validity provides a direct check on how well the test fulfills its function. "The process of ascribing meaning to scores produced by a measurement procedure is generally recognized as the most important task in developing an educational or psychological measure, be it an achievement test, interest inventory, or personality scale" (Stenner, Smith, and Burdick, 1983). For convenience, the various components of test validity—content, criterion-related validity, and construct—will be described as if they are unique, independent components rather than interrelated parts.

Content Validity

The content validity of a test relates to the adequacy with which important content has been sampled and the adequacy with which the content is evidenced in the test. Content validity was built into Total Reader during the development process. All texts are authentic and developmentally appropriate, and the student is asked to respond to the text in ways that are appropriate for the genre (for example, with nonfiction texts, the student is asked specific questions related to the content rather than asked to make inferences about what will happen in the text). Educators selected and reviewed the texts and reviewed the questions.

Construct Validity

The construct validity of a test is the extent to which the test may be said to measure a theoretical construct or trait, such as reading comprehension. Anastasi (1982) identified a number of ways that the construct validity of a test could be examined. One technique is to examine developmental changes in test scores for traits expected to increase with age. Another technique is to examine the "correlations between a new test and other similar tests . . . [the correlations are] evidence that the new test measures approximately the same general areas of behavior as other tests designated by the same name" (p. 145). The third technique is to examine the convergent and discriminate validity evidence (Campbell and Fiske, 1959). It is necessary to "show not only that a test correlates highly with other variables with which it should theoretically correlate, but also that it does not correlate significantly with variables from which it should differ" (Anastasi, 1982, p. 147). The following sections provide evidence of the construct validity of Total Reader.

Developmental Nature. Reading is a skill that is expected to develop with age—as students read more, their skills improve, and therefore they are able to read more complex material. Total Reader has been utilized in 12 states (Arkansas, Arizona, California, Florida, Mississippi, North Carolina, Ohio, Oregon, Vermont, Washington, West Virginia, and Wyoming) and 41 districts during the 2004-2005 and 2005-2006 school years. When the criterion of at least 100 days of elapsed time between initial experience and last experience was invoked, a total of 2,460 students have used the program. Of these students, 78.94% were in grades 3 through 12 in Wyoming. *Table 18* describes the distribution of Lexile measures by grade for Wyoming students. The results show a developmental increase in scores across the grade levels.

Table 18. Distribution of Lexile measures of Wyoming students from Total Reader, grade (*N* = 1,968).

Grade	*N*	Mean Lexile measure	Standard Deviation of Lexile measures
3	79	560L	177L
4	115	645L	151L
5	333	827L	177L
6	356	855L	189L
7	349	938L	176L
8	431	976L	173L
9	136	979L	193L
10	56	1076L	185L
11	88	1190L	137L
12	25	1216L	163L

The data was further examined to estimate growth in reading ability using linear and quadratic regression equations. Additional students were removed from analyses if they did not have at least 20 data points, and then a correlation of at least 0.50 and then 0.70 was required. A sample of 993 students met the inclusion criteria. The resulting linear regression slope was slightly more than 0.50L/day (about 100L of growth between fall and spring) which is consistent with prior research conducted by MetaMetrics, Inc. (see Williamson, Thompson, and Baker 2006). The mean R-squared coefficient was 0.7559 which indicates that correlation between reading ability and time is approximately 0.87.

Ecological Validity

One other aspect of validity—ecological validity—is the "degree to which the behaviors observed and recorded in a study reflect the behaviors that actually occur in natural settings" (Alleydog.com, 2006). For a study to process ecological validity, the methods and materials must be similar to "real-life" situations that are being examined (Brewer, 2000; Useability First; 2006; Wikipedia, 2006). With Total Reader, a student's reading experience is being monitored and users have provided anecdotal evidence as to the relationship of Total Reader to classroom reading instruction.

> Total Reader is excellent at promoting careful reading and using content and context to achieve meaning. It builds and strengthens vocabulary. The reports and graphs also provide a visual incentive for students to compete against themselves. In Total Reader, you have a wonderful product! (*Nancy Coleman, 6th Grade Reading Specialist, Hayes Middle School, Lansing, MI*)

> We rely on the Lexile Framework with a product called Total Reader to better match students with appropriate reading-level books. By using this tool to choose classroom texts and other reading materials, teachers are better equipped to differentiate instruction in the classroom by addressing the needs of all students. [It] is not an instructional program, but is useful in managing instructional interventions and for differentiating instruction. (*"Standards-Based, Technology-Rich Teaching in Wyoming", by Mark Hoffman, The NETC Circuit, NWREL's Northwest Educational Technology Consortium*)

I really like the program, it is easy to use for the children, and it requires very little time. The subject categories appeal to them. (*Odile Yeramian, homeschooling mother*)

My admin team was quite impressed with the data that we are getting from Total Reader and the accessibility of this data. The new features allow me to look at the data in many ways. I can see this becoming a valuable tool for my district. Kudos to you guys! (*Ellen Repstad, District Administrator, Mt. Abraham Union High School*)

I am very excited about the look and feel of the new assessment, Total Reader. The reports are really good. I am happy with the progress. (*Annette Bohling, J.D.*, *Deputy State Superintendent of Educational Quality and Accountability*)

What are the advantages of Total Reader? Many! There are over 3,000 reading [passages] to choose from on this site. The ability to set every student in the school against the Lexile scale in the same program is highly beneficial. We can see the reading progress of every student each year until they graduate. Total Reader is an over-all great tool for grades 3-12. This program offers reading consistency across the board. (*Molly Potas, DISTRICT HORN, Park County School District #16, Vol. 1, Issue 3, March 2006*)

Total Reader can be a highly useful tool in both informing instructional practices as well as tying students daily reading experiences to the year end high stakes assessment. (*Marie Sweeney, Reading Specialist, Portland, OR*)

Total Reader gives us the opportunity to develop both a comprehensive data collection system and related insight into the minds of our struggling readers. Furthermore, our students become extremely motivated and excited to read when they see their Lexile measures. Total Reader is a tool that is great to use for reluctant readers. Not only is the tool exceptional, but the customer service is impeccable. The program has been a win-win situation for both our students as well as our budget! (*Lauren Ellis, Estacada High School, OR*)

References

Alleydog.com. (2006). Ecological validity. Retrieved December 26, 2006, from http://www.alleydog.com/glossary.

Anastasi, A. (1982). *Psychological Testing* (Fifth Edition). New York: MacMillan Publishing Company, Inc.

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

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

Bormuth, J.R. (1967). Comparable cloze and multiple-choice comprehension test scores. *Journal of Reading*, February 1967, 292-299.

Bormuth, J.R. (1968). Cloze test readability: Criterion reference scores. *Journal of Educational Measurement*, *3*(3), 189-196.

Bormuth, J.R. (1970). *On the theory of achievement test items*. Chicago: The University of Chicago Press.

Brewer, M. (2000). “Research deign and issues of validity.” In H. Reis and C. Judd (Eds.), *Handbook of research methods in social and personality psychology*. Cambridge: Cambridge University Press.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, *56*, 81-105.

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). *Peabody Picture Vocabulary Test*-Revised, Forms L and M. Circle Pines, MN: American Guidance Service.

Fountas, I.C. & Pinnell, G.S. (1996). *Guided Reading: Good First Teaching for All Children*. Portsmouth, NH: Heinemann Press.

Greene, Jr., B.B. (2001). Testing reading comprehension of theoretical discourse with cloze. *Journal of Research in Reading*, *24*(1), 82-98.

Haladyna, T.M. (1994). *Developing and validating multiple-choice test items*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Hambleton, R.K., Swaminathan, H., & Rogers, H.J. (1991). *Fundamentals of item response theory* (*Measurement methods for the social sciences*, Volume 2). Newbury Park, CA: Sage Publications, Inc.

Kim, J.S. (2005). Project READS (Reading Enhances Achievement During Summer): Results from a randomized field trial of a voluntary summer reading intervention. Paper presented at Princeton University, Education Research Section, November 7, 2005.

Kim, J.S. (2006, Winter). Effects of a voluntary summer reading intervention on reading achievement: Results form a randomized field trial. *Educational Evaluation and Policy Analysis*, *28*(4), 335-355.

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

Klare, G.R. (1984). Readability. In P.D. Pearson (Ed.), *Handbook of reading research* (Volume 1, 681-744). Newark, DL: International Reading Association.

Kolen, M.J. & Brennan, R.L. (2004). *Test equating: Methods and practices* (second edition). New York: Springer-Verlag.

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. (1987). An extension of the Rasch model to multi-facet situations. Chicago: University of Chicago, Department of Education.

Luce, R.D., & Tukey, J.W. (1964). Simultaneous conjoint measurement: A new type of fundamental measurement. *Journal of Mathematical Psychology*, *1,*1-27.

MetaMetrics, Inc. (1999a). Duval County (FL) Schools: SRI data [data set].

MetaMetrics, Inc. (1999b). Miami-Dade County (FL) Schools: SRI data [data set].

MetaMetrics, Inc. (2000). Lingos feasibility study [data set].

MetaMetrics, Inc. (2006a). Lexile Vocabulary Analyzer technical report. Durham, NC: Author.

MetaMetrics, Inc. (2006b). *PASeries* Reading technical manual. Durham, NC: Author.

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

Moats, L.C. (2000). *Speech to print: Language essentials for teachers*. Baltimore: Paul H. Brooks Publishing Co.

Perline, R., Wright, B.D., & Wainer, H. (1979). The Rasch model as additive conjoint measurement. *Applied Psychological Measurement, 3*(2), 237-255.

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.

Rasch, G. (1980). *Probabilistic models for some intelligence and attachment tests*. Chicago: The University of Chicago Press. (First published in 1960).

Salvia, J. & Ysseldyke, J.E. (1998). *Assessment* (Seventh Edition). Boston: Houghton Mifflin Company.

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

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

Stenner, A.J., Burdick, H., Sanford, E.E., & Burdick, D.S. (2006). How accurate are Lexile text measures? *Journal of Applied Measurement*, *7*(3), 307-322.

Stenner, A. J., Horabin, I., Smith. D. R., & Smith, M. (1988*).* Most comprehension tests do measure reading comprehension: A response to McLean and Goldstein. *Phi Delta Kappan*, 765-769.

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., Horiban, 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., Horiban, I., & Smith, M. (1987b). Fit of the Lexile Theory to sequenced units from eleven basal series. Durham, NC: MetaMetrics, Inc.

Stenner, A.J. & Wright, B.D. (2002). Readability, reading ability, can reading comprehension. In B.D. Wright and M.H. Stone (Eds.), *Making measures* (2004). Chicago: Phaneron Press.

Usability First. (2006). Usability glossary: Ecological validity. Retrieved December 26, 2006, from http://usabilityfirst.com/glossary.

Wikipedia Foundation, Inc. (2006). Ecological validity. Retrieved December 26, 2006, from http://en.wikipedia.org/wiki.

Williamson, G.L., Thompson, C.L., & Baker, R.F. (2006, March). North Carolina's growth in reading and mathematics. Paper presented at the annual meeting of the North Carolina Association for Research in Education (NCARE), Hickory, NC.

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

Wright, B.D., & Linacre, J.M. (2003). *A user's guide to WINSTEPS Rasch-Model computer program (version 3.38).* Chicago: Winsteps.com.

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

Zakaluk, B.L. & Samuels, S.J. (1988). *Readability: Its past, present, and future*. Newark, DL: International Reading Association.

Appendices