



1400L I Lexile: Matching readers to

**Passport Reading Journeys™**  
**Benchmark Assessments**  
Edition 4

Development and Technical Guide

**MetaMetrics, Inc.**

1000 Park Forty Plaza Drive, Suite 120  
Durham, NC 27713  
[www.Lexile.com](http://www.Lexile.com)

Date: October 2009

## Introduction

Passport Reading Journeys™ is a targeted intervention program designed for students in middle and high school (Voyager Expanded Learning, 2005). In fact, nearly eight million students in fourth through twelfth grade struggle to read at grade level. This program is designed to help adolescent students who struggle with reading and need help building upon essential skills.

The Passport Reading Journeys program builds on the Voyager Passport™ reading intervention program for students in kindergarten through grade 6. Highlights of the Passport Reading Journeys program include:

- Teacher-directed whole-group instruction, flexible small-group instruction, and independent practice activities.
- Built-in Reading Connected Text assessments and benchmark assessments powered by The Lexile Framework® for Reading.
- Research-based online software proven to help striving readers master vocabulary, fluency, and comprehension skills.
- Action-packed video segments that get students excited about lesson topics.
- Writing in response to reading instruction.
- A rich library of leveled magazines, books, audio books, and online materials for self-selected reading activities.
- Easy-to-use lesson plans with detailed language and instructional strategies.
- Ongoing professional development and year-round implementation support.

(Voyager Expanded Learning, 2005).

During the summer and fall of 2004, Voyager Expanded Learning met with MetaMetrics, Inc. to discuss ways that an assessment could be developed for use with middle school students reading well below grade level to assess initial reading level and to monitor reading ability development. The result was the development of the Voyager benchmark assessments for Grades 7 and 8 by MetaMetrics. Soon after the implementation of the Grades 7 and 8 assessments, development of a Grade 6 series of assessments began. The design of the test and the development of passages and items was the same for all three grades. During the spring of 2006, MetaMetrics, Inc. developed the Grade 6 set of assessments. The *Voyager Reading Assessment Technical Guide* (2006) describes the design and development of this set of assessments.

When data on the implementation of the full set of Passport Reading Journeys Benchmark Assessments became available in the fall of 2007, it became clear that a new series of assessments was needed to measure the growth of students who were closer to on-grade level in terms of their reading ability. The Passport Reading Journeys program was being used with students whose reading ability was above the 50<sup>th</sup> percentile for their grade level based on their results of the benchmark assessments. To better meet the needs of these students and to monitor growth in their reading ability, the series of assessments for Passport Reading Journeys was redesigned and expanded. This development occurred during the summer and fall of 2007. The new set of assessments includes a screener test and a high and low set of benchmark tests to better target assessment to each student's individual reading level.

In the spring of 2008, the Passport Reading Journeys Benchmark Assessments were expanded to include a set of assessments designed for students in Grade 9. The design of the Grade 9 assessments is the same as for the Grade 6 through 8 assessments, with a screener test and high and low benchmark tests for three administration points. The development of the Grade 9 assessments occurred during the summer and fall of 2008.

All of the assessments measure reading to gain information. The passages convey factual information or are of general human interest. They are all written to represent content-rich areas such as social studies, science, history, technology, and general interest.

Results for the Passport Reading Journeys Benchmark Assessments are calculated using the Lexile<sup>®</sup> scale and the Lexile Framework for Reading, a scientifically based scale of reading ability. The Lexile scale provides accurate feedback, helping measure progress and forecast student development.

This technical guide for the Passport Reading Journeys Benchmark Assessments will provide users with a broad research foundation. Such a base is essential when deciding if and how the benchmark assessments should be used and what kinds of inferences about readers are appropriate.

Ongoing analysis of item performance has increased the amount of information provided in the technical guide and has resulted in some changes to the test items from the first introduction of the Passport Reading Journeys Benchmark Assessments. In addition, ongoing review of reporting guidelines has resulted in some adjustments to the initial testing and reporting protocol. These changes have been noted in each updated edition of the technical guide.

## **Features of Passport Reading Journeys Benchmark Assessments**

Passport Reading Journeys Benchmark Assessments are research-based, scientifically valid, and reliable. Several specific features of the benchmark assessments are noteworthy.

- ◆ Passages are written to be appropriate for students in middle and high school.
- ◆ Two levels for each benchmark assessment are available. A short screener test is administered prior to each benchmark assessment to enable more focused targeting of the benchmark assessment to the reader's ability.
- ◆ The native-Lexile item format and the passage-native Lexile format used on the benchmark assessments 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).
- ◆ The benchmark assessments 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 the benchmark assessment passages and test items is the same equation that is used to measure books/texts. Thus, readers and texts are placed on the same metric.

- ◆ 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 benchmark assessment items. A multi-stage review process was used to ensure conformance with the item-writing specifications.
- ◆ The benchmark assessments are appropriate for individual, small group, and large group administration settings.
- ◆ The test format supports quick administration in an un-timed, low-pressure format.
- ◆ No extensive or specialized preparation is needed to administer the benchmark assessments although proper interpretation and use of the results requires an understanding of the Lexile Framework for Reading.
- ◆ The Voyager benchmark assessments support rapid objective scoring by computer.

## **Purposes and Uses of Passport Reading Journeys Benchmark Assessments**

Voyager benchmark assessments are designed to measure a reader's ability to comprehend expository texts of increasing difficulty. The results of the benchmark assessments can be used to measure students' reading ability and where they are in terms of a developmental progression.

One outcome of the benchmark assessments is the location of the reader on the Lexile scale. After a reader has been measured it is possible to forecast how well the reader will comprehend thousands of books and articles that have also been measured in the Lexile metric and placed on the Lexile scale. 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%. Examples of text at various points on the Lexile scale are shown on the Lexile Map (Appendix A).

## **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.*

## 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, however, only deals with two: reader ability and text difficulty.

Within the Lexile Framework, text difficulty is determined by examining such characteristics as word frequency and sentence length. Text measures typically range from 0L to 1800L. Within any one classroom there will be a range of reading materials.

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

### The Semantic Component

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

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

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

The word frequency measure used was the raw count of how often a given word appeared in a corpus of 5,088,721 words sampled from a broad range of school materials (Carroll, Davies, and Richman, 1971). A “word family” included: (1) the stimulus word; (2) all plurals (adding “-s” or changing “-y” to “-ies”); (3) adverbial forms; (4) comparatives and superlatives; (5) verb forms (“-s,” “-d,” “-ed,” and “-ing”); (6) past participles; and (7) adjective forms. Correlations were computed between algebraic transformations of these means and the rank order of the test items. Since the items were ordered according to increasing difficulty, the rank order was used as the observed item difficulty. The mean log word frequency provided the highest correlation with item rank order ( $r = -0.779$ ) for the items on the combined form.

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

## The Syntactic Component

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

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

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

## Calibration of Text Difficulty

The research study on semantic units (Stenner, Smith, and Burdick, 1983) was extended to examine the relationship of word frequency and sentence length to reading comprehension. In 1987(a), Stenner, Smith, 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 across the 9 tests 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, and 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 should hold when the same two items are administered to a second person. If two different items are administered to the second person, there is no way to know which set of items is harder and which set is easier. The problem is that the location of the scale is not known. General objectivity requires that scores obtained from different test administrations be tied to a common zero—absolute location must be sample independent (Stenner, 1990). To achieve general objectivity, the theoretical logit difficulties must be transformed to a scale where the ambiguity regarding the location of zero is resolved.

The first step in developing a scale with a fixed zero was to identify two anchor points for the scale. The following criteria were used to select the two anchor points: they should be intuitive, easily reproduced, and widely recognized. For example, with most thermometers the anchor points are the freezing and boiling points of water. For the Lexile scale, the anchor points are text from seven basal primers for the low end and text from *The Electronic Encyclopedia* (Grolier, Inc., 1986) for the high end. These points correspond to the middle of first grade text and the midpoint of workplace text.

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

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

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 Evidence for the Lexile Framework for Reading

The 1999 *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education) state that “validity refers to the degree to which evidence and theory support the interpretations of test scores entailed in the uses of tests” (p. 9). In applying this definition to the Lexile Framework for Reading, the question that should be asked is: What evidence supports the use of the Lexile Framework to describe text difficulty and reader ability? Because the Lexile Framework addresses reading comprehension, an important aspect of validity evidence that should be brought to bear is evidence showing that the construct being addressed is indeed reading comprehension. This type of validity evidence has traditionally been called construct validity. One source of construct validity evidence for The Lexile Framework for Reading can be evaluated by examining how well Lexile measures relate to other measures of reading comprehension.

*Lexile Framework Linked to other Measures of Reading Comprehension.* The Lexile Framework for Reading has been linked to several standardized tests of reading comprehension. When assessment scales are linked, a common frame of reference can be used to interpret the test results. This frame of reference can be “used to convey additional normative information, test-content information, and information that is jointly normative and content-based. For many test uses, ... [this frame of reference] conveys information that is more crucial than the information conveyed by the primary score scale” (Petersen, Kolen, and Hoover, 1989, p. 222). Linking the Lexile Framework to other measures of reading comprehension produces a common frame of reference: Lexile measure.

*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 (TAKS)	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 Reading/Literature Knowledge and Skills Test	3, 5, 8, and 10	3,180	0.89
Mississippi Curriculum Test (MCT)	2, 4, 6, and 8	7,045	0.90
Georgia Criterion Referenced Competency Test (CRCT and GHSCT)	1 – 8, and 11	16,363	0.72 to 0.88*
Wyoming Performance Assessment for Wyoming Students (PAWS)	3, 5, 7, and 11	3,871	0.91
Arizona Instrument to Measure Progress (AIMS)	3, 5, 7, and 10	7,735	0.89
South Carolina Palmetto Achievement Challenge Tests (PACT)	3 – 8	15,559	0.87 to 0.88*

Notes: Results are based on final samples used with each linking study.

\*TAAS and TAKS are not vertically equated; separate linking equations were derived for each grade.

*Lexile Framework and the Difficulty of Basal Readers.* In a study conducted by Stenner, Smith, Horabin, and Smith (1987b) Lexile calibrations were obtained for units in 11 basal series. It was presumed that each basal series was sequenced by difficulty. So, for example, the latter portion

of a third-grade reader is presumably more difficult than the first portion of the same book. Likewise, a fourth-grade reader is presumed to be more difficult than a third-grade reader 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 (Poznansky, 1990; Stenner, Wright, and Linacre, 1994). Two of the tests provided no item parameters, but in each case items were ordered on the test in terms of difficulty (e.g., PIAT). For these two tests, the empirical difficulties were approximated by the difficulty rank order of the items. In those cases where multiple questions were asked about a

single passage, empirical item difficulties were averaged to yield a single observed difficulty for the passage.

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

The last three columns in *Table 3* show the raw correlation between observed (O) item difficulties and theoretical (T) item calibrations, with the correlations corrected for restriction in range and measurement error. The Fisher Z mean of the raw correlations ( $r_{OT}$ ) is 0.84. When corrections are made for range restriction and measurement error, the Fisher Z 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?

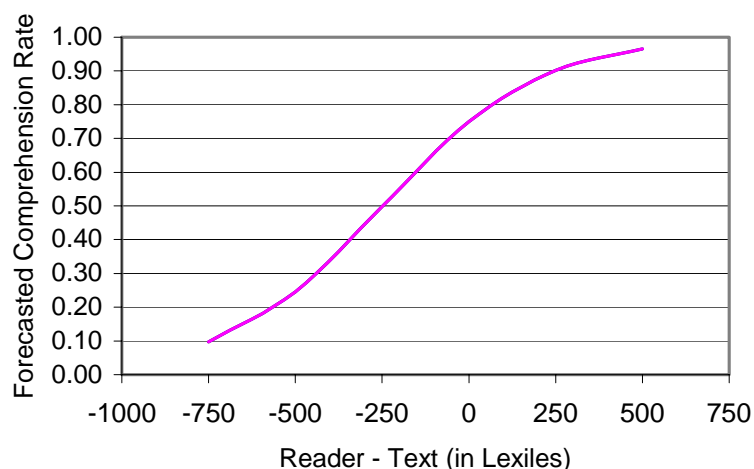
The 75-percent comprehension rate for a reader-text pairing can be given an operational meaning by imagining the text to be carved into item-sized slices of approximately 125-140 words with a question embedded in each slice. A reader who answers three-fourths of the questions correctly has a 75-percent 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 (Equation 1 on page 7). This equation describes the relationship between the measure of a student's level of reading comprehension and the calibration of the items. Unfortunately, comprehension rates calculated by this procedure would be biased because the calibrations of the slices in ordinary prose are not all the same. The average difficulty level of the slices and their variability both affect the comprehension rate.

*Figure 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) 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	Forecasted Comprehension
1000L	500L	<i>Tornado</i> (Byars)	96%
1000L	750L	<i>The Martian Chronicles</i> (Bradbury)	90%
1000L	1000L	<i>Reader's Digest</i>	75%
1000L	1250L	<i>The Call of the Wild</i> (London)	50%
1000L	1500L	<i>On the Equality Among Mankind</i> (Rousseau)	25%

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

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

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

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

## 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 helped to ensure that the items are valid measures of reading comprehension. Item writers are provided with training materials concerning the embedded completion item format and guidelines for selecting passages, developing statements, and selecting options. The item writing materials also contain incorrect items that illustrate the criteria used to evaluate items and corrections based on those criteria. The final phase of item writer training is a short practice session with three items.

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

Item writers are also given extensive training related to "sensitivity" issues. Part of the item writing materials address these issues and identify areas to avoid when selecting passages and developing items. The following areas are covered: violence and crime, depressing situations/death, offensive language, drugs/alcohol/tobacco, sex/attraction, race/ethnicity, class, gender, religion, supernatural/magic, parent/family, politics, animals/environment, and brand names/junk food. These materials were developed based on material published 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 were edited and maintained in the item bank.

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

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

## Development of the Passport Reading Journeys Benchmark Assessments

The Passport Reading Journeys Benchmark assessments were designed to measure reading ability. Voyager Expanded Learning identified criteria for the development of the assessment:

- Simplified test administration that could be accomplished through a paper-based environment while utilizing item formats to closely match state assessments.
- Minimum number of items per test form and minimum administration time while still ensuring minimal measurement error when determining each student's reading ability.
- Development of multiple test forms for monitoring student growth in reading.

Design and development of the first edition Grade 7 and Grade 8 benchmark assessments took place during 2004 and 2005, with Grade 6 following in 2006. *The Voyager Reading Assessment Technical Guide* (2006) describes the complete design and development process of the first edition forms. The design and development of the second edition assessments took place during 2007 and are described in *The Voyager Reading Assessment Technical Guide — Edition 2* (2008). This edition, *The Voyager Reading Assessment Technical Guide—Edition 3* (2008), adds a description of the Grade 9 assessments to the previous technical guide editions. The following sections of this technical guide describe each stage of the development process.

### Assessment Specifications

Usage information and an analysis of student test data from the first edition Passport Reading Journeys Benchmark Assessments identified a need for an additional set of tests written for students at a higher ability level than those for whom the first edition tests were designed. The first edition assessments were designed to measure reading ability for students who were reading in the 5<sup>th</sup> to 50<sup>th</sup> percentile of grade-level students. Some students who were enrolled in the Passport Reading Journeys Program were reading at a level above the 50<sup>th</sup> percentile and were not well targeted by the first edition assessments. Voyager Expanded Learning, in conjunction with MetaMetrics, Inc., developed a specification for a series of tests that includes a screener and two levels of benchmark tests to better target higher and lower ability students.

At each grade level, students are assessed upon entry to the program to determine initial reading level, midway through the program to monitor progress, and again at the end of the instructional program to determine overall growth. At each assessment point, the student receives a Lexile measure to help guide program reading assignments.

The specifications for the second edition set of assessments and the Grade 9 assessments called for a 10-item screener test and two 30-item benchmark tests for use at each assessment period. Student results on the initial screener test determine which benchmark test the student will be administered. The combined results from the screener and benchmark test determine the student's Lexile measure. *Figure 2* shows the relationship between the screener test, and Benchmark 1 test, and the total score for the student.

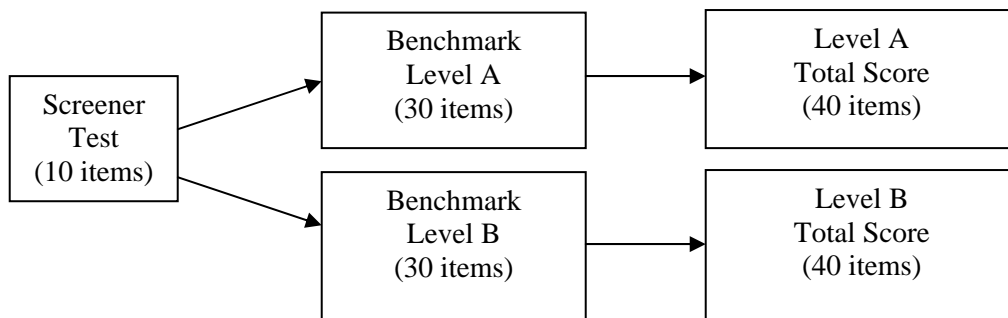


Figure 2. Screener and benchmark test progression for Benchmark 1.

At each grade level, the screener test was designed to measure grade-level reading ability from approximately the 5<sup>th</sup> to the 85<sup>th</sup> percentile of grade-level readers. The Lexile measure for the percentiles is based on previous work conducted by MetaMetrics, Inc. with the Lexile Framework for Reading. Based on the number correct for the screener items administered before Benchmark 1, the student is assigned to either Benchmark 1 Form A or Benchmark 1 Form B for the remaining portion of the assessment. The Benchmark Form A tests were targeted at approximately the 40<sup>th</sup> to 85<sup>th</sup> percentile for each grade; each test in the grade-level progression was designed to be of graduated difficulty based upon the administration cycle. The Benchmark Form B tests were targeted at approximately the 5<sup>th</sup> to 50<sup>th</sup> percentiles, and, like the Benchmark A forms, each test was designed to be of graduated difficulty. The Benchmark Form B tests measure a similar range of ability as the first edition Passport Reading Journeys Benchmark Assessments. The reading levels for the screener tests were established by Lexile level according to the Lexile measures presented in *Table 6*. The reading level targets for the benchmark tests are shown in *Table 7*.

Based on performance on Benchmark 1, students are assigned to either Benchmark 2 Form A or Form B. *Figure 3* shows the relationship between the Benchmark 1 scores, and the assignment to Benchmark 2 forms.

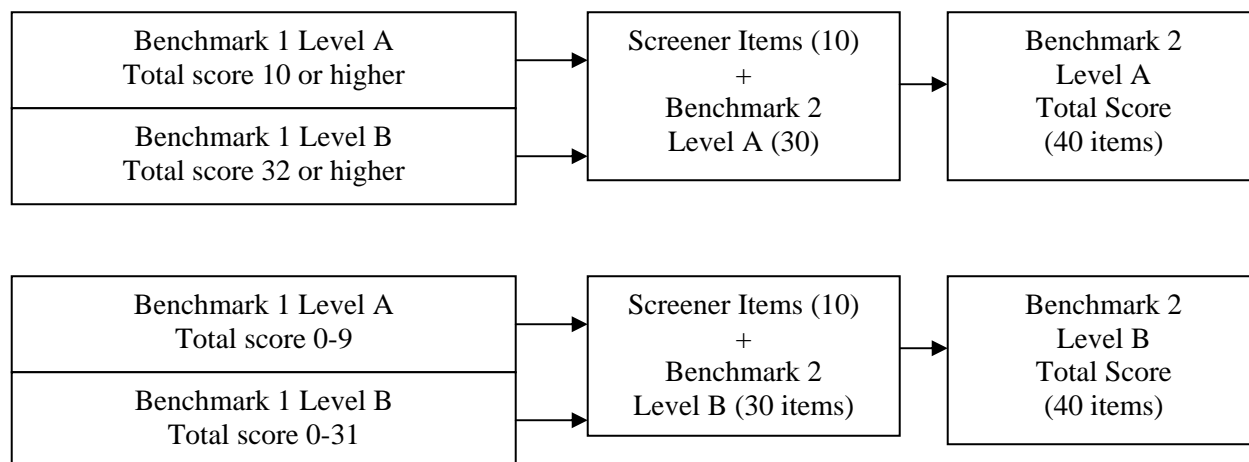
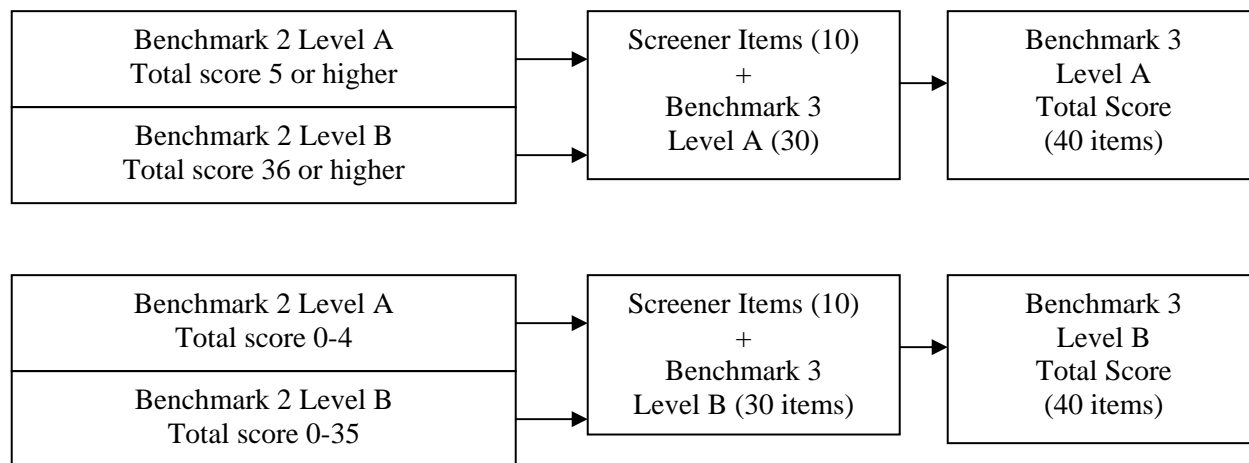


Figure 3. Progression from Benchmark 1 to Benchmark 2.

Similarly, students are assigned to a Benchmark 3 form based on their performance on Benchmark 2. *Figure 4* shows the relationship between the Benchmark 2 scores and the assignment to Benchmark 3 forms.



*Figure 4.* Progression from Benchmark 2 to Benchmark 3.

For the Benchmark 2 and 3 assessments, the screener items function to provide a set of broad range reading items to supplement the larger set of targeted items at each form level.

*Table 6.* Lexile targets by reading level for screener tests.

Grade/Level	Reading Range (5 <sup>th</sup> to 85 <sup>th</sup> Percentile)	Test 1	Test 2	Test 3
6/Beginnings	400L to 1110L	400L to 1040L	450L to 1070L	520L to 1110L
7/Journeys I	500L to 1170L	500L to 1090L	550L to 1140L	620L to 1170L
8/Journeys II	550L to 1220L	550L to 1150L	600L to 1180L	650L to 1220L
9/Journeys III	650L to 1260L	650L to 1200L	700L to 1220L	750L to 1260L

**Table 7.** Lexile targets by reading level for benchmark tests.

Grade /Level	Form Level	Test 1	Test 2	Test 3
6 / Beginnings	A	800L to 1040L	880L to 1070L	900L to 1110L
	B	400L to 800L	450L to 850L	520L to 900L
7 / Journeys I	A	900L to 1095L	950L to 1135L	970L to 1170L
	B	500L to 890L	550L to 930L	550L to 960L
8 / Journeys II	A	950L to 1145L	1000L to 1180L	1020L to 1220L
	B	550L to 950L	600L to 1000L	650L to 1050L
9 / Journeys III	A	1020L to 1200L	1045L to 1220L	1075L to 1260L
	B	650L to 950L	700L to 1000L	750L to 1100L

## Passage Development

The reading passages on all of the Voyager Expanded Learning Benchmark Assessments were specified to be informational in nature. Informative passages were described as text that conveyed factual information or described how to perform a task, were nonfiction, and were selected from content-rich text (e.g., social studies, science, and history). Five authors were commissioned to develop the passages.

The following criteria were established for the development of passages for the Voyager Benchmark Assessments:

- Passages should be approximately 300 words, and be self-contained and complete (have a clear beginning, middle, and end).
- Content should be informative and instructional, and the focus of the topic chosen should relate to science and social studies. At the same time, the content of the passages should be relevant and interesting to a typical middle school student (interesting, exciting, adventurous, surprising aspects of the topic).
- Passages should be written to reflect somewhat sophisticated topics and subject matter, while staying at the targeted Lexile level. Examinees who take the Level B forms of the Benchmark Tests typically represent the 5<sup>th</sup> to 50<sup>th</sup> percentile of middle school readers. Examinees who take the Level A forms are stronger readers – typically the 40<sup>th</sup> to 85<sup>th</sup> percentile of middle school readers.

- **Standard English.** Reading passages should use conventions appropriate for students at the targeted grade level.
- **Bias Free.** All passages and items should be free from bias based on race, gender, age, ethnicity, religion, disability, sexual orientation, or socioeconomic status. No group should have an advantage over another because of values, vocabulary, phrasing, or assumptions in a passage. Stereotypes of ethnic or gender groups in passages and items should be avoided.
- To the degree possible, prior knowledge should not be required for the examinee to understand or appreciate the passage. References to events, people, and places should be explained within the passage unless considered common knowledge. Figurative language should be explained within the passage or be defined through context.
- All passages should avoid topics that may be offensive to, or induce an emotional reaction from, an examinee, parent, or citizen group (e.g., violence, abuse, terminal illness, and poverty).
- All passages and items should be free of registered trademarks and brand names. Common business names should also be avoided.
- **Use of Real-World Contact Information.** Generally, contact information should not be given in a passage. However, when necessary to include fictional contact information (e.g., a customer service phone number in an appliance manual), it should be modeled after real-world contact information. To ensure that the contact information is not real, these guidelines were followed:

#### **Phone numbers**

Local phone numbers should appear as 555 + (0100–0199), (e.g., 555–0185).

Long distance numbers should appear as (live area code [except 555] + 555 + 0100–0199), (e.g., 319–555–0177).

Toll-free numbers should appear as 1 + 888 + (100–199) + four digits (e.g., 1–888–190–4455).

"Vanity" numbers may be used following the above guidelines (e.g., 1–888–100–1–FUN).

#### **Addresses**

Complete addresses should use a mismatched city and zip code.

Verify to the extent possible that the named business or individual cannot be reached in the named city.

#### **URLs and E-Mail Addresses**

URLs should end in *voyager.com* or other address that is acceptable to Voyager and not proprietary (e.g., *www.cityzoo.voyager.com*).

E-mail addresses should end in *@voyager.com* or other address that is acceptable to Voyager and not proprietary (e.g., *rtb@voyager.com*).

- All source material used for the development of passages must be documented, and copies of book, magazine, or web pages, and any other source material should be sent

back to MetaMetrics with the passage. Some helpful websites (for content and style) include:

<http://news.nationalgeographic.com/kids/>

<http://www.timeforkids.com/TFK/news>

<http://www.factmonster.com/>

Although the content of the texts used in the passages could be altered if necessary, it was important to select text for items that was void of sensitive issues. The following guidelines were 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.

1. Violence/crime: Avoid weapons, fights, arrests, illegal activities, abuse, and murders.
2. Depressing situations or death: Avoid sickness, death, and other negative situations.
3. Offensive language: Avoid use of curse words or words used to cover up a harsher curse; avoid oaths such as "Oh God!", words that belittle others, or other insulting words such as "backwards" or "ugly."
4. Drugs/alcohol/tobacco: Avoid any mention of drugs, alcohol, tobacco, and anything associated with these topics such as rehab, bars, etc.
5. Sex/attraction: Avoid issues that call for a discussion of sex, sexual orientation, or relationships of either a romantic or sexual nature.
6. Race: Avoid racial slurs, belittling words, stereotypes (e.g., referring to Native Americans as Indians), and unbalanced representations of a race (e.g., mentioning African Americans only in the context of slavery).
7. Class: Avoid mentioning economic and social differences and avoid stereotypes.
8. Gender: Use gender free language (e.g., firefighter instead of fireman); avoid using male pronouns to refer to both sexes; show both genders in a variety of roles; avoid stereotypical portrayals of men or women.
9. Religion: Avoid selections that promote or demean a religious belief; avoid the assumption that people share a common belief; avoid mention of a reference to any holidays of a religious nature (e.g., Christmas, Halloween).
10. Supernatural/magic: Avoid mention of witches, goblins, wizards, and other supernatural beings; avoid magic in general.
11. Parents/family: Avoid selections that question parents, authority, or judgment; avoid negative relationships within the family; avoid raising the issue of alternative families.
12. Politics: Avoid controversial issues (e.g., unions, strikes) and selections that portray political bias.
13. Animals/environment: Avoid topics about hunting and/or cruelty to animals (e.g., fur coats, trapping animals) and be sensitive to environmental issues and animal rights.
14. Brand names/junk food: Avoid mentioning either.

All of the passages and items from the first edition of the Passport Reading Journeys Benchmark Assessments were used in the second edition assessments. Student data from test administrations of the first edition forms were analyzed using RUMM2020, Version 4.1 (Andrich, Sheridan, & Luo, 1997) to determine whether the items were performing as expected. Based on these analyses, approximately 19 percent of the first edition items were revised slightly for use in the second edition forms.

To complete development of the second edition assessments, a total of 47 extended passages were developed for use with the passage-native Lexile items. An additional 69 passages were

developed for use with the native-Lexile items in the screener tests and Test 1 of the benchmark series for each grade. To complete development of the Grade 9 assessments, an additional 32 extended passages and 50 native-Lexile passages were developed. All passages were reviewed for alignment with the specifications and for potential developmental inappropriateness. The passages were also reviewed for reading demand (Lexile measure) and revised where needed to reflect the specifications needed for test development.

Passages used for the passage-native Lexile items (five items per passage) were approximately 300 words in length. Passages used with the native-Lexile items (one item per passage) were an average of 100 words in length. The guidelines for passage length were established to help ensure that the overall length of each test was uniform, and that the reading demand of each test form allowed administration within a single class period. A frame was written for each extended passage to introduce the passage and to direct the reader in the assessment task. Each frame was 1 to 2 sentences in length.

## Item Development

The Passport Reading Journeys assessments measure reading comprehension by focusing on skills readers use when studying written materials sampled from various content areas. These skills include understanding the importance of details in the passage, drawing conclusions, and making comparisons and generalizations. These reading assessments do not require prior knowledge of ideas outside of the passage, vocabulary taken out of context, or formal logic.

There is evidence to support the conclusion 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 re-reading of the passage and an increase in the use of context clues. The traditional cloze procedure is based on the deletion of every 5<sup>th</sup> to 7<sup>th</sup> word (or some variation) regardless of part of speech. It can also consist of selectively deleting certain categories of words (Bormuth, 1967, 1968, 1970). Selective deletions have shown greater instructional effects than random deletions.

The item formats used with the Passport Reading Journeys Benchmark Assessments can be described as variants of the selective deletion cloze format—the native-Lexile item format and the passage-native Lexile item format. These item formats are similar to the fill-in-the-blank format. When properly written, these formats directly assess 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 these formats, 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. These formats are “well-suited for testing a student’s ability to evaluate” (Haladyna, 1994, p. 62). In addition, these formats are also useful as an instructional tool.

There are two main advantages to using these item formats. The first is that the level of reading of the statement and the four options is controlled to ensure that their difficulty level is easier than the most difficult word in the passage. The second advantage of these formats is that while typical passages are used, the statement is as short as or shorter than the briefest sentence in the passage. These two advantages help ensure that the statement is easier than the accompanying passage.

The statement portion of the 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, provide 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 statement is read by itself, any of the four options could be plausible.

The following criteria were used to develop native-Lexile items and passage-native Lexile items. The statement should:

1. Require the student to draw an unambiguous conclusion or inference from the passage.
2. Be clear as to what or whom the statement question is about.
3. Not use the exact or nearly the same wording as what appears anywhere in the passage.
4. Attempt to avoid the use of negatives.

The answer choices should:

1. Be reasonably grade level/Lexile targeted (300L below to 100L above as a general guideline).
2. Logically complete the statement to force passage dependence for answering correctly. (All foils should make sense in context of the statement, but only the correct choice should make sense in context of the paragraph.)
3. Be one word or a short phrase.
4. Not be homonyms, as this may merely confuse the reader. Avoid using antonyms; if two choices are opposite there is a high probability that one is correct.
5. Contain words from the passage only if all of the answer choices do as well.
6. Be balanced; if correct choice is a word or phrase containing a positive connotation, at least one other choice should be positive so the correct choice does not stand out. Although, with higher-level texts it is best to try and make all of the words positive or negative.
7. Not use negative sentence structure.
8. Vary in form as the Lexile level of the item increases, for example, the answers should not all be written including the same phrasing.
9. Be selected in accordance to sensitivity restrictions.

*Item Writer Training.* Item writers were experienced item-development specialists who had experience with the everyday reading ability of students at various levels. The use of individuals with these types of experiences helped to ensure that the items are valid measures of reading ability. Item writers were provided with training materials concerning the native-Lexile item format and the passage-native Lexile item format and guidelines for selecting passages, developing statements, and selecting options. The item writing materials also contained incorrect items that illustrate the criteria used to evaluate items and corrections based on those criteria. The final phase of item writer training was a short practice session with three items.

Item writers were provided vocabulary lists to use during statement and option development. The vocabulary lists were compiled from word lists compiled by MetaMetrics based on vocabulary research related to determining the Lexile measures (difficulty) of words. The Lexile Vocabulary Analyzer (LVA) determines the Lexile measure of a word using a set of features related to the source text and the word's prevalence in the MetaMetrics corpus (MetaMetrics, 2006). The rationale used to compile the vocabulary lists was that the words should be part of a

reader’s “working” vocabulary if they had likely been encountered in easier text (those with lower Lexile measures).

Item writers were provided additional 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. These materials were developed based on material published by CTB/McGraw-Hill (*Guidelines for Bias-Free Publishing*) concerning universal design and fair-access—equal treatment of the sexes, fair representation of minority groups, and the fair representation of disabled individuals.

Item writers were first asked to independently develop items for two passages. The items were then reviewed by the development group for item format, grammar, and sensitivity. Based on this review, item writers received feedback and more training if necessary.

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

## Test Development

*Test Design Specifications.* Using the assessment design specifications presented in *Tables 6 and 7*, specific test specifications were developed for each test level.

The screener tests contain only native-Lexile items. Benchmark 1 tests contain a combination of native-Lexile items and passage-native Lexile items while Benchmarks 2 and 3 tests contain only passage-native Lexile items. The number and type of items on each test are presented in *Table 8*.

*Table 8.* Number of items by item type for each test.

Item Type	Screener Test	Benchmark Test 1	Benchmark Test 2	Benchmark Test 3
Native-Lexile Item	10	10	0	0
Passage-Native Items (5 items per passage)	0	20	30	30
Total Number of Items	10	30	30	30

The screener tests are used to determine the most appropriate Benchmark 1 test assessment for each student, so the range of difficulty of the screener test is large. For Benchmarks 2 and 3, the screener items serve to provide a small set of broad range items in conjunction with the larger set of targeted Benchmark Form items. In general, the screener test item difficulty range matches the combined range of the associated Benchmark Forms A and B. Based on

performance on the first wide-range screener test, the student is given the 30-item Benchmark 1 Form test that best matches his or her initial performance. The Benchmark Form A tests were targeted at approximately the 40<sup>th</sup> to 85<sup>th</sup> percentile for each grade and the Benchmark Form B tests were targeted at approximately the 5<sup>th</sup> to 50<sup>th</sup> percentiles. Each test form level (A or B) in the grade-level progression was designed to be of graduated difficulty based upon the administration cycle. Reading levels for the screener tests and are presented in *Tables 9 and 10*.

*Table 9. Specification of native-Lexile items for screener tests—Grades 6 and 7.*

Lexile Range	Grade 6 Screeners			Grade 7 Screeners		
	1	2	3	1	2	3
400L to 490L	2					
500L to 590L		2	1	1	1	
600L to 690L	2	1	1	1	1	1
700L to 790L	2	2	1	1	1	2
800L to 890L	2	2	3	2		
900L to 990L	1	2	2	3	5	3
1000L to 1090L	1	1	2	2	2	2
1100L to 1190L					1	2
1200L to 1290L						
Total Items	10	10	10	10	10	10
Mean	745L	800L	855L	853L	891L	920L

*Table 10. Specification of native-Lexile items for screener tests—Grades 8 and 9.*

Lexile Range	Grade 8 Screeners			Grade 9 Screeners		
	1	2	3	1	2	3
400L to 490L						
500L to 590L	1					
600L to 690L		1	1	1		
700L to 790L	2	1	1	2	2	1
800L to 890L	1	1		1	2	2
900L to 990L	3	2	2	2	1	1
1000L to 1090L	2	4	3	2	2	2
1100L to 1190L	1	1	2	2	2	3
1200L to 1290L			1		1	1
Total Items	10	10	10	10	10	10
Mean	900L	954L	986L	925L	977L	1018L

Each Benchmark 1 Test contains 10 native-Lexile items and 20 passage-native Lexile items. *Tables 11 and 12* show the Lexile zones of the native-Lexile items for Benchmark 1. *Tables 13 to 16* show the Lexile zones for the passage-native Lexile items for all benchmark tests.

*Table 11.* Specification of native-Lexile items for Benchmark Test 1—Grades 6 and 7.

Lexile Range	Grade 6		Grade 7	
	Form A	Form B	Form A	Form B
400L to 490L		3		
500L to 590L		3		4
600L to 690L		3		2
700L to 790L		1		2
800L to 890L	2			2
900L to 990L	6		5	
1000L to 1090L	2		5	
1100L to 1190L				
1200L to 1290L				
Total Native-Lexile Items	10	10	10	10

*Table 12.* Specification of native-Lexile items for Benchmark Test 1—Grades 8 and 9.

Lexile Range	Grade 8		Grade 9	
	Form A	Form B	Form A	Form B
400L to 490L				
500L to 590L		2		
600L to 690L		2		2
700L to 790L		2		2
800L to 890L		3		4
900L to 990L	2	1		2
1000L to 1090L	6		5	
1100L to 1190L	2		4	
1200L to 1290L			1	
Total Native-Lexile Items	10	10	10	10

At all grade levels, Benchmark Tests 2 and 3 contain passage-native Lexile items only, so the total number of passages for Benchmarks 2 and 3 is greater than the total number of passages for the Benchmark 1 Tests.

**Table 13.** Specification of passages for passage-native Lexile Items, Grade 6.

Lexile Range	Benchmark Test 1		Benchmark Test 2		Benchmark Test 3	
	Form A	Form B	Form A	Form B	Form A	Form B
400L to 490L		1				
500L to 590L		1		2		1
600L to 690L		1		2		2
700L to 790L		1		1		2
800L to 890L	2			1		1
900L to 990L	1		4		2	
1000L to 1090L	1		2		4	
<b>Total Passages</b>	<b>4</b>	<b>4</b>	<b>6</b>	<b>6</b>	<b>6</b>	<b>6</b>

**Table 14.** Specification of passages for passage-native Lexile Items, Grade 7.

Lexile Range	Benchmark Test 1		Benchmark Test 2		Benchmark Test 3	
	Form A	Form B	Form A	Form B	Form A	Form B
500L to 590L		1		1		
600L to 690L		1		1		2
700L to 790L		1		3		2
800L to 890L		1				1
900L to 990L	2		1	1	1	1
1000L to 1090L	2		4		2	
1100L to 1190L			1		3	
<b>Total Passages</b>	<b>4</b>	<b>4</b>	<b>6</b>	<b>6</b>	<b>6</b>	<b>6</b>

**Table 15.** Specification of passages for passage-native Lexile Items, Grade 8.

Lexile Range	Benchmark Test 1		Benchmark Test 2		Benchmark Test 3	
	Form A	Form B	Form A	Form B	Form A	Form B
600L to 690L		1		1		1
700L to 790L		1		2		1
800L to 890L		1		1		1
900L to 990L	1	1		2		2
1000L to 1090L	2		3		2	1
1100L to 1190L	1		3		4	
<b>Total Passages</b>	<b>4</b>	<b>4</b>	<b>6</b>	<b>6</b>	<b>6</b>	<b>6</b>

**Table 16.** Specification of passages for passage-native Lexile Items, Grade 9.

Lexile Range	Benchmark Test 1		Benchmark Test 2		Benchmark Test 3	
	Form A	Form B	Form A	Form B	Form A	Form B
600L to 690L		1				
700L to 790L		2		2		1
800L to 890L		1		2		1
900L to 990L				1		3
1000L to 1090L	2		2	1	1	1
1100L to 1190L	2		3		3	
1200L to 1290L			1		2	
Total Passages	4	4	6	6	6	6

The test forms were constructed according to the specifications previously described and the information in *Tables 8 through 16*. *Table 17* shows the mean, standard deviation, and Lexile range of items for the nine final Screener Tests of the Passport Reading Journeys Benchmark Assessments.

**Table 17.** Descriptive statistics for final screener tests.

Grade	Screener 1		Screener 2		Screener 3	
	Lexile measure Mean (SD)	Range	Lexile measure Mean (SD)	Range	Lexile measure Mean (SD)	Range
6	745.00 (189.40)	420L to 1020L	800.00 (171.33)	540L to 1030L	855.00 (167.55)	550L to 1060L
7	853.00 (174.61)	530L to 1090L	891.00 (188.76)	540L to 1140L	920.00 (179.63)	600L to 1130L
8	900.00 (178.08)	550L to 1130L	954.00 (159.87)	650L to 1180L	986.00 (170.37)	660L to 1220L
9	925.00 (180.08)	660L to 1180L	977.00 (100.5)	710L to 1220L	1018.00 (170.02)	770L to 1250L

*Tables 18 through 21* show the mean, standard deviation, and Lexile range of items and passages for the final test forms of the Passport Reading Journeys Benchmark Assessments.

Table 18. Descriptive statistics for Grade 6/Beginnings -- final benchmark tests.

Form	Benchmark Form 1		Benchmark Form 2		Benchmark Form 3	
	Lexile measure Mean (SD)	Range	Lexile measure Mean (SD)	Range	Lexile measure Mean (SD)	Range
A	923.67 (68.05)	810L to 1030L	971.67 (52.99)	900L to 1060L	1001.67 (54.27)	910L to 1070L
B	568.67 (109.79)	420L to 740L	656.67 (91.29)	540L to 820L	708.33 (89.72)	570L to 850L

Table 19. Descriptive statistics for Grade 7/Journeys I -- final benchmark tests.

Form	Benchmark Form 1		Benchmark Form 2		Benchmark Form 3	
	Lexile measure Mean (SD)	Range	Lexile measure Mean (SD)	Range	Lexile measure Mean (SD)	Range
A	1000.00 (54.39)	920L to 1090L	1046.67 (62.88)	950L to 1140L	1085.00 (62.63)	970L to 1170L
B	680.00 (128.01)	510L to 890L	736.67 (121.44)	580L to 950L	773.33 (111.67)	630L to 950L

Table 20. Descriptive statistics for Grade 8/Journeys II -- final benchmark tests.

Form	Benchmark Form 1		Benchmark Form 2		Benchmark Form 3	
	Lexile measure Mean (SD)	Range	Lexile measure Mean (SD)	Range	Lexile measure Mean (SD)	Range
A	1049.67 (56.90)	980L to 1140L	1085.00 (46.52)	1020L to 1150L	1110.00 (51.19)	1030L to 1190L
B	763.00 (114.96)	550L to 920L	811.67 (102.62)	660L to 950L	855.00 (131.93)	660L to 1060L

Table 21. Descriptive statistics for Grade 9/Journeys III -- final benchmark tests.

Form	Benchmark Form 1		Benchmark Form 2		Benchmark Form 3	
	Lexile measure Mean (SD)	Range	Lexile measure Mean (SD)	Range	Lexile measure Mean (SD)	Range
A	1097.67 (60.84)	1000L to 1200L	1128.33 (49.28)	1060L to 1210L	1163.33 (59.79)	1080L to 1250L
B	784.00 (86.01)	650L to 940L	861.67 (90.10)	740L to 1000L	900.00 (91.91)	760L to 1040L

The final review process for the forms was conducted in a three-stage process. First, the test and passage specifications were reviewed: Lexile measures of items and means and standard deviations of test forms, word counts across the forms, and distributions of correct responses. Next, the tests were taken to verify the answer keys and review the foils in relation to the passages and items. Finally, the overall tests were reviewed for flow and consistency. The following criteria were used to evaluate each set of tests for a grade level/span:

#### *Curricular Perspective*

1. Length
  - a. Is the location of longer and shorter passages similar across the forms?
  - b. Do the forms have about the same length and same number of words?
  - c. Is the number of words similar across the comparable passages across the three forms?

#### *Psychometric Perspective*

1. Is the range of Lexile measures consistent across the equivalent forms? (I.e., does one form have a passage from the top of one zone and the bottom of the next compared to another form that has the Lexile measures evenly spaced?)
2. Is the distribution of the placement of correct answers approximately equal (25% for each response position)?
3. Are there runs of the same correct response position? (I.e., more than 3 of any response position in a row.)
4. Is the same word used as the correct response for more than one item on a form?

Where necessary, passages and or items were revised and test forms were reconstituted to more closely reflect the test specifications. Since the Lexile Theory was used to estimate the difficulties of the passages and items, it was necessary for the test specifications to be adhered to as closely as possible.

#### **Form and Item Analysis**

## Scoring and Reporting

The Passport Reading Journeys Benchmark Assessment scores are reported on the Lexile scale. 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 scaled Lexile measure. The typical range of the Lexile scale is from 200 to 1700 Lexiles, although actual Lexile measures can range from below zero to above 2000 Lexiles. Reader Lexile measures are reported in 5-unit intervals. For reporting purposes, the lowest Lexile measure reported with Passport Reading Journeys assessments is “BR” for “Beginning Reader” (scores at or below 0L).

There are many reasons to use scale scores rather than raw scores to report test results. Scale scores overcome the disadvantage of many other types of scores (e.g., percentiles and raw scores), in that equal differences between scale score points represent equal differences in ability. Each question on a test has a unique level of difficulty; therefore, answering 23 questions correctly on one form of a test requires a slightly different level of ability from answering 23 items correctly on another form of the test. But, receiving a scale score (Lexile measure) of 675L on one form of a test represents a similar level of reading ability as receiving a scale score (Lexile measure) of 675L on another form of the test.

Correspondence tables were provided for each test form based upon the difficulties of the items on the form. The Lexile uncertainties (standard errors) for each score point were also provided.

*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).

The measures that are reported for an individual student should reflect the purpose for which they will be used. If the purpose is accountability (at the student, school, or district level), then actual measures should be reported at all score points. If the purpose is instructional, then the scores should be capped at the upper bounds of measurement error (90<sup>th</sup> percentile point based on prior research by MetaMetrics with the Lexile Framework). In an instructional environment where the purpose of the Lexile measure is to appropriately match readers with books, all scores at or below 0L should be reported as “BR” (Beginning Reader); no student should receive a negative Lexile measure.

*Test Use Guidelines.* Students should not be administered a specific test form more than once within any one year. When a student takes the same assessment form a second time within the span of one year, we are unsure as to how to interpret change in Lexiles: (1) because the student’s reading ability has improved/grown, or (2) because the student remembers some of the items and has experience with the testing environment.

Assessment practices should be in accordance with the generally accepted ethical standards of the education profession. Accordingly, any practice that increases students’ scores should simultaneously represent an increase in students’ mastery (i.e., increasing students’ abilities to perform skills or demonstrate knowledge in real world situations) of the content domains tested. For more information, refer to *Standards for Educational and Psychological Testing* (AERA, APA, NCME, 1999).

## 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. Uncertainty has been examined with the Passport Reading Journeys Benchmark Assessments in terms of examinee error.

*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).

Whenever a model is used to explain the relationship between parameters, some of the differences between observed and theoretical measures cannot be explained. Voyager benchmark assessments were 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. *Tables 22 and 23* describe the uncertainties due to model misspecification for the Voyager benchmark assessments. Complete correspondence tables for all possible scores on each assessment have been provided to Voyager Expanded Learning.

Table 22. Uncertainties by Lexile range, grade, and test form.

Lexile Range	Grade 6		Grade 7		Grade 8		Grade 9	
	Level A	Level B	Level A	Level B	Level A	Level B	Level A	Level B
200L to 295L		70L		77L		85L		89L
300L to 395L	83L	65L	96L	69L	105L	74L		79L
400L to 495L	74L	64L	82L	66L	89L	68L	95L	70L
500L to 595L	67L	65L	72L	64L	76L	65L	81L	65L
600L to 695L	62L	68L	66L	66L	68L	64L	71L	63L
700L to 795L	63L	76L	63L	70L	64L	66L	65L	65L
800L to 895L	65L	88L	63L	77L	62L	71L	63L	69L
900L to 995L	71L	103L	66L	89L	64L	79L	63L	76L
1000L to 1095L	80L		72L	103L	70L	92L	67L	88L
1100L to 1195L			85L		79L		74L	
1200L to 1295L					92L		86L	

Table 23. Average uncertainty (25% - 75% correct), by grade and test form.

Test Form	Grade 6		Grade 7		Grade 8		Grade 9	
	Level A	Level B	Level A	Level B	Level A	Level B	Level A	Level B
1	66L	67L	65L	67L	65L	67L	66L	66L
2	65L	66L	65L	67L	65L	66L	65L	66L
3	65L	66L	65L	67L	65L	67L	65L	66L

## Validity

The 1999 *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education) state that “validity refers to the degree to which evidence and theory support the interpretations of test scores entailed in the uses of tests” (p. 9). Validity evidence provides information about how well a test will fulfill its intended 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). Because a test score from a Voyager assessment will be used as a measure of the reading ability of a student and will be used to target reading instruction, validity evidence should primarily focus on the degree to which the Voyager assessments measure reading comprehension of appropriate reading material. For convenience, the various sources of validity evidence—content, criterion-related, and construct validity evidence—will be described as if they are unique, independent components rather than interrelated parts. At this time the primary sources of validity evidence come from an examination of the content of the Voyager assessments and the degree to which the assessments can be said to measure reading comprehension (construct validity). As more data are collected and more studies are completed, additional validity evidence will be described.

### Content Validity Evidence

Validity evidence concerning the content of a test relates to the degree to which the test content is supportive of the intended interpretations of the test scores. The Voyager assessments have been designed to measure reading comprehension ability of nonfiction texts. To this end the tests were constructed with nonfiction texts. In addition, the text difficulty of the reading passages was analyzed using the Lexile Analyzer to ensure that the difficulty of the text was appropriate for the students taking the placement and progress monitoring tests. The difficulty of the item vocabulary was also matched to the difficulty of the passage. All passages were designed to reflect authentic material, and students are 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). The passages and items were thoroughly reviewed prior to placement on a test.

### Construct Validity Evidence

Evidence for construct validity of the Voyager assessments is provided by the extensive body of research supporting the Lexile Framework for Reading. The development of the Voyager assessments utilized tools for text measurement such as the Lexile Analyzer and procedures for item development that have been shown to result in effective measures of reading comprehension. All of the items on the Voyager assessments are native items in the family of items upon which the research on the Lexile Framework were based. The section in this technical report entitled *The Lexile Framework for Reading* provides a detailed description of the framework and evidence to support that tests based upon the framework measure reading comprehension.

## References

- American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.
- Andrich, D., Sheridan, B., & Luo, G. (1997). *Rasch Unidimensional Measurement Models (RUMM2020)*.
- 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.
- 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.
- 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*. Newbury Park, CA: Sage Publications, Inc.
- Klare, G.R. (1963). *The measurement of readability*. Ames, IA: Iowa State University Press.

- Lieberman, 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.
- MetaMetrics, Inc. (2006, August). *Lexile Vocabulary Analyzer: Technical report*. Durham, NC: Author.
- Miller, G.A. & Gildea, P.M. (1987). How children learn words. *Scientific American*, 257, 94-99.
- 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.
- 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., Smith, M., & Burdick, D.S. (1983). Toward a theory of construct definition. *Journal of Educational Measurement*, 20(4), 305-315.
- Stenner, A.J., Smith, D.R., Horabin, I., & Smith, M. (1987a). Fit of the Lexile Theory to item difficulties on fourteen standardized reading comprehension tests. Durham, NC: MetaMetrics, Inc.
- Stenner, A.J., Smith, D.R., Horabin, I., & Smith, M. (1987b). Fit of the Lexile Theory to sequenced units from eleven basal series. Durham, NC: MetaMetrics, Inc.
- Voyager Expanded Learning. (2005). Passport Reading Journeys™ Overview. Retrieved September 28, 2005, from <http://www.voyagerlearning.com/journeys/overview.jsp>.
- Wright, B.D. & Linacre, J.M. (1994, August). *The Rasch model as a foundation for the Lexile Framework*. Unpublished manuscript.
- Wright, B.D. & Stone, M.H. (1979). *Best Test Design*. Chicago: MESA Press

# Appendices

Appendix A: The Lexile Framework® for Reading Map..... A-1

## Appendix A

### **The Lexile Framework<sup>®</sup> for Reading Map**

[Note: The Map is a separate PDF and should be printed on 11 × 17 paper.]