Measures of Text Difficulty:

Testing their Predictive Value for Grade Levels and Student Performance

Jessica Nelson, Charles Perfetti,
David Liben, and Meredith Liben

Report submitted to the Gates Foundation, February 1, 2012

Table of Contents

1 Executive Summary

The Common Core State Standards draw attention to the need for students to engage with texts of appropriate complexity throughout schooling. This goal requires valid and reliable measures of text complexity that can guide curriculum decisions, assist assessment development, and support the efforts of educational publishers to meet complexity guidelines. This report addresses the extent to which current measures of text complexity meet this requirement.

The study assessed the capabilities of six text difficulty metrics to predict reference measures of text difficulty. These six metrics were as follows: Lexile (MetaMetrics), ATOS (Renaissance Learning), Degrees of Reading Power: DRP Analyzer (Questar Assessment, Inc.), REAP (Carnegie Mellon University), SourceRater (Educational Testing Service), and the Pearson Reading Maturity Metric (Pearson Knowledge Technologies). Additionally, the study included a seventh metric (Coh-Metrix, University of Memphis) that provides multiple indices of text difficulty. All these metrics use measures of word difficulty (frequency, length) and sentence length. Some metrics add other features of words, sentence syntax, and text cohesion, creating a broader range of text and linguistic measures. To assess the value of these metrics in ordering texts according to difficulty, we acquired five sets of texts as reference measures. These included 1) the set of exemplar texts that were placed into grade levels by education experts and published as Appendix B of the Common Core Standards, 2) a set of standardized state test passages, 3) passages from the Stanford Achievement Test (SAT-9), 4) comprehension passages from the Gates-MacGinitie Reading Test, and 5) passages from the MetaMetrics Oasis platform used for student practice. In addition, Rasch scores, which reflect student performance, were obtained for both the SAT-9 and Gates-MacGinitie passages. Thus, reference measures included both measures of grade level and measures of student performance against which to test the text difficulty metrics.

The general form of these tests was the rank order correlation (Spearman's *Rho*) between the text difficulty measures provided by the metrics and those provided by the reference measures. These correlations constitute the main results of the study.

The results established the value of the text difficulty metrics in predicting student performance (Rasch Scores) on the Stanford and Gates-MacGinitie passages. These correlations were impressively high

for both tests, between .74 and .81 for five of the six metrics for the Gates-MacGinitie. (The exception was the REAP metric, which tended to produce low correlations across most reference measures.) More variability was observed for grade level measures, especially for the Common Core exemplar texts and the standardized state tests. For example, correlations for the latter ranged across the metrics from .59 to .79. Generally, for these grade level measures, the metrics that included the broader range of linguistic and text measures produced higher correlations than the measures that used word difficulty and sentence length measures.

Two other sources of variability were observed. The metrics produced higher correlations for informational texts than narrative texts across the two reference measures that made this distinction. However, on one of these two comparisons, the Reading Maturity Metric did well on both text types. The second source of variability was the discrimination among grade levels over the entire range of grades. The metrics tended to discriminate better among the lower grades than among the higher grades.

The results have implications for education. One is the viability of text difficulty metrics as guides to curriculum and assessment standards. The metrics studied can support the goal of the Common Core Standards to increase student achievement by reducing the large gap that currently exists between typical high school level and college texts (ACT, Inc., 2006; ACT, Inc., 2009). In addition to the practical value of the metrics that provide a single quantitative index of text difficulty, the finer grain analysis of texts, which could be of value for curriculum decisions and for research on text complexity, is demonstrated by measures (e.g. Coh-Metrix) that provide multi-dimensional descriptors of text complexity.

2 Background

This study was undertaken in support of the Common Core State Standards' emphasis on students reading texts of appropriate complexity. This emphasis and the research base for it are described in detail in Appendix A of the Common Core Standards for English Language Arts (CCSSO, 2010).

In order for stakeholders to identify and select texts of appropriate complexity for each grade and band level and to better understand the nature of complex text, measures of text complexity that are validated by research are needed. Furthermore, there is a critical need for these tools to help stakeholders identify what makes texts complex, what makes reading difficult for students, and whether these two are the same.

At the time the Standards were released (June 2010), the need for further research into text complexity measurement was acknowledged by the Council of Chief State School Officers (CCSSO, 2010), one of the initiators of the Common Core Standards. Seven groups who had developed text analysis tools were identified and all agreed to participate in this study, undertaken between September 2010 and August 2011.

As a condition of participating, each group committed to offering transparency in revealing both the text features it analyzed and the general means of analysis. Each group also agreed to make available a version of its analysis tool that could be adapted for public access at the individual user level and be relatively user-friendly in that role. Appendix D lists each tool and how to access the public version of the analyzer. Furthermore, it was required that the analysis tool be valid, reliable, and able to calibrate text difficulty by grade or band level to match the Common Core Standards' demand for appropriate text complexity by grade (band) levels.

What follows is the report on the research and results of the study of quantitative measures of text difficulty.

3 Research Study Questions and Aims

The goal of this research was to evaluate text analysis tools that can measure text complexity quantitatively with reliability and validity.

Besides the central question of which tools function best for this purpose, other questions have surfaced. One is whether additional features of text, such as vocabulary and cohesion features, can be measured to yield practical and predictive information about text beyond sentence length and word difficulty. Another is the question of how well objective features that make text complex are the same features that make text difficult for readers. Does this predictability change at different grade levels? Last, narrative literature offers particular challenges to quantitative assessment (CCSSO, 2010, p. Appendix A), so it was of particular interest to examine the predictive abilities of the analyzer tools with both informational and narrative text.

4 Study Methods

4.1 Performance Evaluation Methods

We assessed the measures provided by the text analysis tools (henceforth referred to as "metrics") by computing the correlations between each metric and an independent second estimate of passage difficulty, which we refer to as a "reference measure". Reference measures included grade levels and scores based on student comprehension of the passages acquired for five sets of text passages described below.

For these correlations, we report the non-parametric Spearman's rank correlation coefficient (*rho*) rather than the Pearson's product moment correlation (*r*). (For reference, Pearson's correlations are provided in Appendix B). The rank order correlation accomodates a wide range of possible underlying data distributions. Thus, *rho* is less sensitive to outliers, indifferent to non-normality in the data, and makes no assumption that the reference measures comprise an equal interval scale. It assumes only that the relation between the two measures can be described by a monotonic function. Thus, *rho* describes the extent to which each metric ranks text difficulty in the same order as the reference measure.

We used a Fisher r-to-z transformation to compute 95% confidence intervals for the correlation coefficients (Caruso & Cliff, 1997). The confidence intervals are interpreted as the range of the "true" correlations to be expected in the populations of reference measures being sampled. The confidence intervals are entirely dependent on the sample size (e.g. number of texts) and the observed value of *rho*. Datasets with more texts, as well as higher values of *rho*, will have shorter confidence intervals, and shorter confidence intervals are less likely to show overlap. Generally, in the data we report below, there is substantial overlap in the confidence interval for one metric and the confidence interval for any of the others metrics.

In addition to these correlations, we describe the degree of automaticity of each tool. Although all tools will compute a measure for any given text, some degree of text "cleaning" prior to applying the tool can provide more meaningful results. For example, images, headings, misspellings, lists, footnotes, and non-ASCII characters may need to first be removed or corrected. This may be done either by hand or

automatically. If significant manual effort is required for the tool to work, the tool will not be as scalable for broader use as a more automatic tool will be.

Below, we first provide a description of each of the metrics that were evaluated in the study. We then provide a description of each of the sets of texts used as reference measures and the correlations of these measures with each metric. Finally, we summarize the study results.

4.2 The Metrics

Seven research groups provided metrics for analysis. All of these measures are intended to index text complexity or text difficulty (both terms are used by the metrics) using word level factors (frequency or word length) and sentence level or syntactic difficulty (estimated using sentence length), which are variations on traditional readability formulae. The metrics vary in the extent to which they use additional linguistic and text features as predictors.

Table 1: Overview of Metrics

Research Group	Metric(s)
MetaMetrics	Lexile
Renaissance Learning	Advantage / TASA Open Standard (ATOS)
Questar Assessment, Inc.	Degrees of Reading Power: DRP Analyzer
The REAP Project: Carnegie Mellon	REAP (REAder-specific Practice) Readability
Educational Testing Service (ETS)	SourceRater
Pearson Knowledge Technologies (PKT)	Pearson Reading Maturity Metric
Coh-Metrix: University of Memphis	Narrativity, Referential Cohesion, Syntactic Simplicity, Word Concreteness, Deep Cohesion

Two of the text tools (Pearson's Reading Maturity Metric and SourceRater's grade level estimate) describe text properties along several dimensions in addition to providing the single text difficulty score that allowed correlations to be computed in this study. Coh-Metrix computes only a multi-dimensional analysis of texts (each dimension with an associated normalized score) and, thus, did not meet our study's requirement of a single metric that could be correlated with a reference measure. Accordingly, we consider the Coh-Metrix data in a separate section in which we describe how each of the five dimensions varies with grade level and student comprehension performance.

Each metric is described in more detail below.

4.2.1 Lexile®

4.2.1.1 Self Description

"The Lexile® Framework for Reading is a scientific approach to measuring reading ability and the text demand of reading materials. The Lexile Framework includes a Lexile measure and the Lexile scale. A Lexile measure represents both the complexity of a text, such as a book or article, and an individual's reading ability. Lexile measures are expressed as numeric measures followed by an "L" (for example, 850L) and are placed on the Lexile scale. The Lexile scale is a developmental scale for measuring reader ability and text complexity, ranging from below 200L for beginning readers and beginning-reader materials to above 1700L for advanced readers and materials. Knowing the Lexile measures of a reader and a text helps to predict how the text matches the reader's ability—whether it may be too easy, too difficult, or just right. All Lexile products and services rely on the Lexile measure and Lexile scale to match reader with text.

"The Lexile® Framework for Reading (Lexile.com) evaluates reading ability and text complexity on the same developmental scale. Unlike other measurement systems, the Lexile Framework determines reading ability based on actual assessments, rather than generalized age or grade levels. Recognized as the standard for matching readers with texts, tens of millions of students worldwide receive a Lexile measure that helps them find targeted readings from the more than 400 million articles, books, and websites that have been measured. Lexile measures connect learners of all ages with resources at the right level of challenge and monitor their progress toward state and national proficiency standards."

4.2.1.2 Variables Used

- → Word Frequency
- → Sentence Length

4.2.1.3 Text Cleaning / Automaticity

Figures, tables, equations, titles, headings, footnotes/endnotes, numbered lists, non-standard characters, and pronunciation guides must be removed or altered manually prior to analyzing the texts. Misspellings can be optionally detected automatically and corrected by hand to improve accuracy. Non-standard prose such as plays, interviews, poetry, recipes, or lists, which all have non-standard punctuation, cannot be accurately processed. Other texts of any length, starting with a single sentence, can be processed.

4.2.2 ATOS

4.2.2.1 Self Description

"Released in 2000, ATOS is the product of an intensive research and development process that sought to develop a more accurate and user-friendly quantitative readability system. ATOS includes two formulas: ATOS for Text (which can be applied to virtually any text sample, including speeches, plays, and articles) and ATOS for Books. Both formulas take into account three variables that research determined to be the most important predictors of text difficulty: words per sentence, average grade level of words, and characters per word. (Grade level of words is established via the Graded Vocabulary List, which is believed to be the most extensive tool of its kind, developed and modified using existing word lists, word frequency studies, vocabulary test results, and expert judgment.) ATOS for Books also includes adjustments for book length and variations in internal structure, two factors shown to significantly impact the understandability of books. ATOS is provided by Renaissance Learning as a free and open system. ATOS research and formulas are published in a technical report, and users may submit text for analysis free-of-charge at Renaissance's web site. Because ATOS is the default readability system incorporated in the Accelerated Reader (AR) program used in approximately 50,000 schools, it is arguably the most widely-used system for matching students with books in the US. ATOS can be reported in many different scales. The default is grade equivalent, which means both student achievement and books can share the same scale, one that is easy for educators, parents, and students to interpret."

4.2.2.2 Variables Used

- Word length
- Word grade level
- Sentence length (with adjustments for extreme sentence length in the ATOS for books formula)
- Book length (in ATOS for books formula)

4.2.2.3 Text Cleaning / Automaticity

No text cleaning is required to automatically compute the ATOS metric, nor are corrections or changes to the text made by the analyzer. Cleaning the texts can be done manually to improve the accuracy of the ATOS output (for example, correcting misspellings), but this was not done for texts analyzed for this study. Only texts without recognizable sentences cannot be analyzed. There is no minimum or maximum text length that can be processed — files with as many as 3,000,000 words have been processed successfully.

4.2.3 Degrees of Reading Power®: DRP® Analyzer

4.2.3.1 Self Description

"The DRP Analyzer employs a derivation of a Bormuth mean cloze readability formula based on three measureable features of text: word length, sentence length, and word familiarity. DRP text difficulty is expressed in DRP units on a continuous scale with a theoretical range from 0 to 100. In practice, commonly encountered English text ranges from about 25 to 85 DRP units, with higher values representing more difficult text; DRP units = (1 – Bormuth value) x 100. The Bormuth formula was chosen for several reasons, including its low standard error of measurement and published validation and cross-validation data.

"The standardized procedures by which the DRP values are calculated are as important as the initial selection of the Bormuth formula, to be certain that all variables are counted consistently in every sample of text analyzed. Text cleaning and other rules determine, for example, what are considered common words, whether hyphenated words are counted as one word or two, and how initials, acronyms, and abbreviations, etc. are treated, ensuring that the DRP Analyzer provides consistent, reliable, and valid results.

"Standardized text sampling rules are also applied. If a book has between 150 and 1000 words of continuous text, the entire book is analyzed. For longer books, the overall readability is obtained by analyzing approximately 300-word samples of text taken from different parts of the book according to a sampling plan based on book length. The sample results are averaged to calculate the mean difficulty of book sections and the entire book.

"DRP reading comprehension tests combine the reading selection and the assessment in a single structure, and the result is an estimate of functional reading ability that has been empirically demonstrated across four decades to be highly valid and reliable. The measurement of student reading ability and the readability of instructional materials are reported on the same DRP scale, providing educators with instructionally relevant information about the most difficult texts the student can read proficiently at various comprehension levels (e.g., independent and instructional)."

4.2.3.2 Variables Used

- → Word length
- → Word difficulty
- → Sentence Length
- → Within-sentence punctuation

4.2.3.3 Text Cleaning / Automaticity

The DRP Analyzer requires the text to be free of non-standard characters, diagrams, headings, formulas and equations, numbered lists, etc. This pre-processing is done manually according to a consistently applied set of rules. Typographical errors that are actually present in the source text are left as-is. The DRP Analyzer can analyze texts ranging from 150 to 1000 words. The DRP values may not be as reliable for texts with fewer than 150 words. Texts with more than 1000 words are manually broken into shorter texts, and the whole text is analyzed in segments.

4.2.4 REAP (REAder-specific Practice) Readability Tool

4.2.4.1 Self Description

"The REAP Readability Tool was created for use in the REAP vocabulary learning system (Heilman et al., 2006). This system is designed to teach students vocabulary through exposure to new words in context, within documents that the student reads. The goal of the tool is to define the level of each document from the levels of the individual words it contains. The tool uses support vector machine regression to achieve a prediction and a simple bag-of-words approach (words are stemmed and function and short words are removed) (Collins-Thomson & Callan 2004) to determine level. It does not take into account higher-level attributes such as cohesiveness.

"As such, the tool provides a basic vocabulary difficulty estimate and can serve as a baseline to compare other, more sophisticated measures, determining the level of contribution of knowledge beyond the word level."

4.2.4.2 Variables Used

- → Word frequency
- → Word length
- → Sentence length
- → Sentence count
- → Parse tree of sentences and paragraphs
- → Frequency of node elements

4.2.4.3 Text Cleaning / Automaticity

REAP automatically removes function words and any words with fewer than 3 characters. No other text cleaning is required for the tool to run, and manual corrections of the text are not made. Texts of any length, starting with a single word, can be analyzed.

4.2.5 SourceRater

4.2.5.1 Self Description

"SourceRater is a comprehensive text analysis system designed to help teachers and test developers evaluate the complexity characteristics of stimulus materials selected for use in instruction and assessment. SourceRater includes two main modules: an Analysis Module and a Feedback Module.

"SourceRater's Analysis Module employs a variety of natural language processing techniques to extract evidence of text standing relative to eight construct-relevant dimensions of text variation, including: syntactic complexity, vocabulary difficulty, level of abstractness, referential cohesion, connective cohesion, degree of academic orientation, degree of narrative orientation, and paragraph structure. Resulting evidence about text complexity is accumulated via three separate regression models: one optimized for application to informational texts, one optimized for application to literary texts, and one optimized for application to mixed texts. The specific regression model to be employed in each new analysis can either be specified by the user, or determined via SourceRater's automated genre classifier.

"SourceRater also includes an innovative Feedback Module designed to help users understand and compare the individual complexity drivers detected in individual texts (see Sheehan et al., 2010). Feedback includes graphical displays designed to highlight similarities and differences between the candidate text and a corpus of texts with known grade level classifications. Individual displays can facilitate efforts to (i) determine the specific aspects of text variation that may account for unexpectedly low or high grade-level classifications; (ii) identify areas of the text likely to be more or less problematic for struggling readers; and (iii) document text characteristics for presentation to technical review committees."

4.2.5.2 Variables Used

- Word frequency
- Word length
- Word meaning features (concreteness, imagability, etc.)
- Word syntactic features (tense, part of speech, proper names, negations, nominalizations, etc.)
- Word types (academic verbs, academic downtoners, academic word list)
- Sentence length
- Paragraph length

- Within-sentence and between-sentence cohesion measures
- Number of clauses (including type and depth)
- Text genre: informational, literary, or mixed (computed automatically or manually overridden, if preferred)

4.2.5.3 Text Cleaning / Automaticity

The analyzer requires paragraph markings to be correct, which may require manual correction. Non-standard characters, certain punctuation, and erroneous end-of-sentence markers are detected automatically and must be corrected manually. SourceRater can analyze texts of any length, but accuracy rates for texts under 100 words or over 3000 words have not been determined.

As SourceRater was under development over the course of this study, some of the features that are now available (including the ability to analyze mixed-genre texts and the inclusion of "messy text" filters) had not been implemented for the analysis of certain text sets.

4.2.6 Pearson Reading Maturity Metric

4.2.6.1 Self Description

"The new Pearson Reading Maturity Metric marks a major advance in the measurement of reading difficulty and text complexity. The most important innovation, called Word Maturity, uses the computational language model, Latent Semantic Analysis (LSA) to accurately estimate how much language experience is required to achieve adult knowledge of the meaning of each word, sentence and paragraph in a text. Using measures based on the average maturity and highest maturity words, the metric accurately estimates overall difficulty and complexity of the language used in the text.

"An example of a useful application of the metric is highlighting of the estimated most difficult words in a given reading. It also supports a number of related analyses, such as showing the changes in multiple senses of words as they mature, which can have significant effects on complexity.

"While the Word Maturity measure accounts for a substantial portion of the total variation in and accuracy of our overall measure of text complexity, a selection of other computational linguistic variables is also included to increase the predictive power of the Reading Maturity Metric, such as perplexity, sentence length, and semantic coherence metrics. An important demonstration of the metric's overall validity is its high correlation with that of human test-takers on well-established vocabulary and reading tests. A demonstration of the accuracy of

the method's underlying technologies is its agreement on essay scores equal to that between two expert graders. Similarly, the value of the basic AI technologies behind the metric is attested by its use in Pearson's widely acclaimed automatic essay scoring and reading comprehension technologies."

4.2.6.2 Variables Used

- Pearson Word Maturity Metric
- Word length (e.g. syllables per word)
- Sentence length
- Within-sentence punctuation
- Within and between-sentence coherence metrics
- Sentence and paragraph complexity (e.g. perplexity)
- Order of information

4.2.6.3 Text Cleaning / Automaticity

The Pearson Word Maturity Metric requires a consistent character encoding scheme, such as UTF-8, and non-text elements, such as illustrations, need to be removed before analysis. However, manual cleaning is typically not needed. The measures have been designed to be robust and invariant under the normal variability seen in texts, such as the presence of headings and footnotes. For this study, no manual text cleaning was used.

4.2.7 Coh-Metrix Text Easability Assessor

4.2.7.1 Self Description

"The Coh-Metrix Text Easability Assessor analyzes the ease or difficulty of texts on five different dimensions: narrativity, syntactic simplicity, word concreteness, referential cohesion, and deep cohesion. For a given text, each of these dimensions is given an "ease score" compared to thousands of other texts. Narrativity measures whether the passage is story like and includes events and characters. Syntactic simplicity refers to the complexity or ease of the sentence syntax. Word concreteness measures whether the words in the passage are imageable versus abstract. Two important types of cohesion are measured by Coh-Metrix using a variety of indices. Referential cohesion is the overlap between sentences with respect to major words (nouns, verbs, adjectives) and explicit ideas. A text has higher referential cohesion when sentences have similar words and ideas. A cohesion gap occurs when a sentence has no words or ideas that connect to other sentences in the text. When text cohesion is higher, students more easily understand the text and are better able to comprehend the relationships between ideas or events in the text. Deep cohesion assesses meaning at deeper levels, such as

causal and temporal relations between events, actions, goals, and states. In order to understand these deeper meanings, it is often important for texts to have connective words (such as “because,” “therefore,” “however”) to help glue these ideas together. This is especially important when the purpose of a text is for instruction, for example a textbook or an article being used to introduce a topic to students.”

4.2.7.2 Variables Used

- Word frequency
- Word length
- Word meaning features (concreteness, imagability, number of senses, etc.)
- Word syntactic features (part of speech, negations, etc.)
- Sentence length
- Sentence complexity
- Paragraph length
- Within-sentence and between-sentence cohesion measures

4.2.7.3 Text Cleaning / Automaticity

Non-standard characters and certain types of punctuation are automatically detected and altered in pre-processing. Otherwise, no changes are made to the texts. The Coh-Metrix Text Easability Assessor can analyze texts ranging from 200 to 1000 words. The assessor output may not be as reliable for texts with fewer than 200 words. If a text has more than 1000 words, shorter segments of text are automatically sampled from the full text for analysis. The maximum text length can be increased, but the time it takes to process the text will also increase.

5 Results

5.1 Results for Five Text Sets

There is no clear "gold standard" measure of text difficulty against which to compare the various metrics. Instead, we compared each metric against various reference measures based on grade level and student comprehension data for five sets of passages gathered for the study. These are defined and discussed in the sections following. Although there are limitations in the validity of these indicators as measures of text difficulty, the variety in their construction allows us to observe the robustness of the metrics and consider how different reference measures might affect their performance. For example, grade level or band level as determined by expert educators reflects teachers, librarians, and curriculum developers' conception of passage difficulty, whereas mean Rasch scores (estimated item difficulty) are computed from parameters for comprehension test items and for student performance. Estimates of text difficulty that are consistently predictive of such widely varying constructs will be useful for teachers, publishers, and parents in determining whether a text is likely to be at the appropriate difficulty level for instruction in a certain grade band.

5.1.1 Common Core Exemplar Texts

5.1.1.1 Initial Selection

The text samples selected for inclusion in Appendix B of the Common Core Standards for ELA (CCSS) were intended to exemplify the level of complexity and quality that the Standards require for all students in a given grade band. They are presented by band levels that consist of the following: Grades 2–3, Grades 4–5, Grades 6–8, Grades 9-10, and Grade 11 to College and Career Readiness (CCR). This set of texts was also intended to suggest the breadth of text types required to fulfill the Common Core Standards. It is important to emphasize that these texts were intended to signal the demand for increased complexity that the Common Core Standards hold as a central tenet.

The process of selecting texts for inclusion was as follows: A working group was assembled from among the constituencies guiding the writing of the Common Core Standards. This working group solicited contributions from teachers, librarians, curriculum specialists, educational leaders, and reading researchers who had experience

working with students in the grades for which the texts were recommended. These contributors were asked to recommend texts that they or their colleagues had used successfully with students in a given grade band and to justify and describe that use.

Reviewing the recommendations and assembling the final collection was done using the following considerations:

Complexity: Following the recommendations set forth in Appendix A of the CCSS, a three-part model for measuring complexity was used. The three parts were qualitative indices of inherent text complexity judged by human raters, quantitative measures using Lexiles® and Coh-Metrix features of Easability, and professional (educator) judgment for matching texts to an appropriate band level. Final selection was made by the working group and vetted broadly during the Standards vetting process.

Quality: The working group recognized that it was possible to have high-complexity text of low inherent quality, so it solicited text recommendations of recognized value. From the pool of submissions offered by outside contributors to the process, the working group selected classic or historically significant texts as well as contemporary works of comparable literary merit, cultural significance, and rich content.

Range: After identifying texts of appropriate complexity and quality, the working group applied other criteria to ensure that the samples presented in each band represented as broad a range of sufficiently complex, high quality texts as possible. The proportions of texts that were classified by the working group as either informational, literary non-fiction or, literary narrative follow the percentages called for at each band level by the CCSS.

This explanation was modified from the introduction to Appendix B of the Common Core State Standards, which contains the excerpted texts used in this part of the research study. Poetry and drama selections were not used in this study. See 5.1.1.2 below for other exclusions.

5.1.1.2 Passages Removed for Analysis

Reason for Removal	Number Removed
Dramas	10
Duplicates	25
Intended for teacher to read aloud	9
Total Removed Passages	44
Total Remaining Passages	168

5.1.1.3 Missing Data

None.

5.1.1.4 Text Properties

Average Number of Words	475.5
Grade Levels	2–12
Text Difficulty Measure(s)	Grade Level
Subsets Examined	Informational vs. Narrative

5.1.1.5 Reference Measures

The reference measure was the Common Core grade band as established by expert instructors (See 5.1.1.1): Texts were classified into five grade bands: Grades 2–3, 4–5, 6–8, 9-10, and 11–12. Metrics were compared against the rank order of these five grade bands.

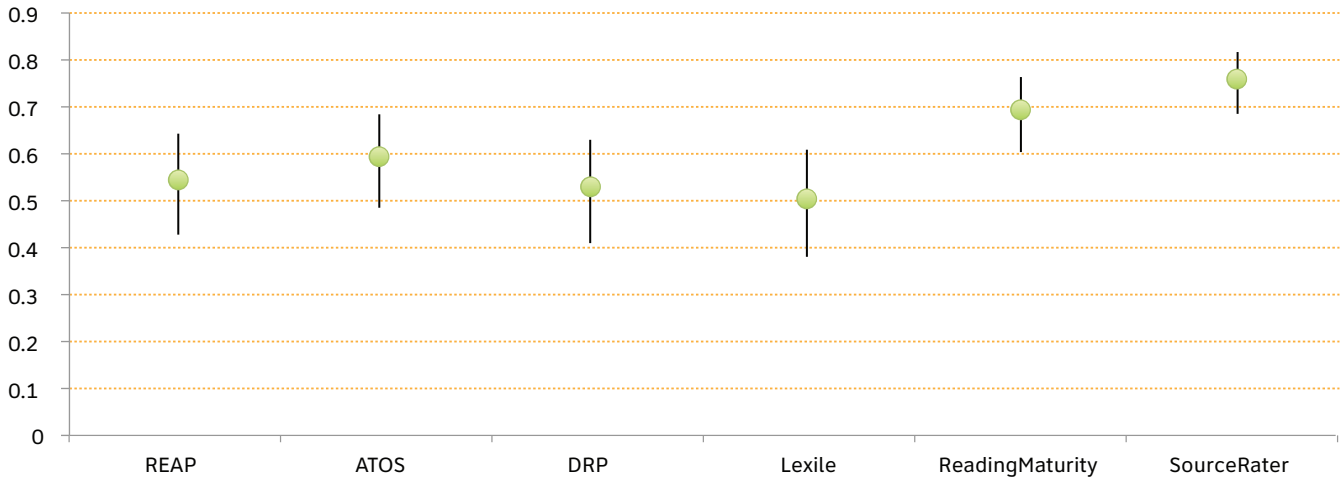
5.1.1.6 Notes / Caveats

Because these texts are clustered into grade bands of more than one grade, the sensitivity of the text difficulty construct is limited.

5.1.1.7 Results

Figure 5.1.1–1 shows the rank order correlation (*rho*) of the rank produced by each text difficulty metric with the text difficulty ranking assigned by the expert educators. Each correlation is shown with its 95% confidence interval. As a group, the metrics were moderately successful in predicting the expert ratings of difficulty level. SourceRater (*rho*=.76) and Reading Maturity (*rho*=.69) produced relatively high correlations compared with the other metrics, which showed *rho*s between .50 and .59. Note that the confidence interval for any given metric overlaps with the confidence interval for most of the others. However, the confidence interval for SourceRater overlaps only with that of Reading Maturity.

Figure 5.1.1–1: Common Core Exemplar Texts, Correlation with Grade Band (n=168)



95% Confidence Interval	REAP	ATOS	DRP	Lexile	Reading Maturity	SourceRater
Lower Limit	0.427	0.484	0.409	0.380	0.602	0.683
Rho	0.543	0.592	0.527	0.502	0.690	0.756
Upper Limit	0.641	0.682	0.628	0.607	0.761	0.814

5.1.2 Standardized State Test Passages

5.1.2.1 Initial Selection

Prior to the publication of the Common Core State Standards, a preliminary research project on the sources of text complexity was carried out using two of the measures (Coh-Metrix and Lexile) ultimately included in the present study. The results of that study are encapsulated in Appendix A of the Common Core State Standards.

A small team collected released state and national test passages, converted them to .txt format, and "scrubbed them" free of stray marks so they could be accurately read by the Coh-Metrix and MetaMetrics analyzer tools. This data set consisted of 1275 passages that had been used in a variety of state and national assessments and subsequently released. American College Testing also allowed use of a number of their unreleased passages for the preliminary study (those passages are not included in this study).

These collected passages, with the exception of the ACT passages, can be found at two open Google sites: Text Complexity Conversion Site and Text Conversion Project 2, where the passages are identified and housed.

These sites and their contents have recently been made public and are available for legitimate research purposes.

Identification of texts as belonging to informational, narrative, or mixed genre categories was done by educator judgment on a passage-by-passage basis. Where states identified their passages by a particular genre type, that identification was generally retained in our study (as "text type") after review and confirmation.

For this study, some of the passages used in the first round and stored on the Google sites were removed. Table 5.1.2.2 identifies the reasons for removal and how many passages of each category were removed.

5.1.2.2 Passages Removed for Analysis

Reason for Removal	Number Removed
Description of Passage (not passage)	3
Dramas	3
Duplicates	40
NAEP passages	24
Outline	1
Poem	1
Resumé	1
Science Assessments	7
Simulated Student Writing	5
Table	2
Not from grade-targeted test (mostly from NY Regents test)	505
Total Removed	592
Total Remaining	683

5.1.2.3 Missing Data

Metric	Number of Texts	Reason
SourceRater	399	Did not meet ETS criteria for valid grade level (see notes); classified by ETS as mixed genre

5.1.2.4 Text Properties

Average Number of Words	574.0
Grade Levels	3–11
Text Difficulty Measure(s)	Grade Level
Subsets Examined	Subset evaluated by ETS Grades 3–5 vs. Grades 6–8 v Grades 9–11 Informational v Narrative

5.1.2.5 Reference Measures

The reference measure is the grade level of the standardized test on which each passage appeared.

5.1.2.6 Notes / Caveats

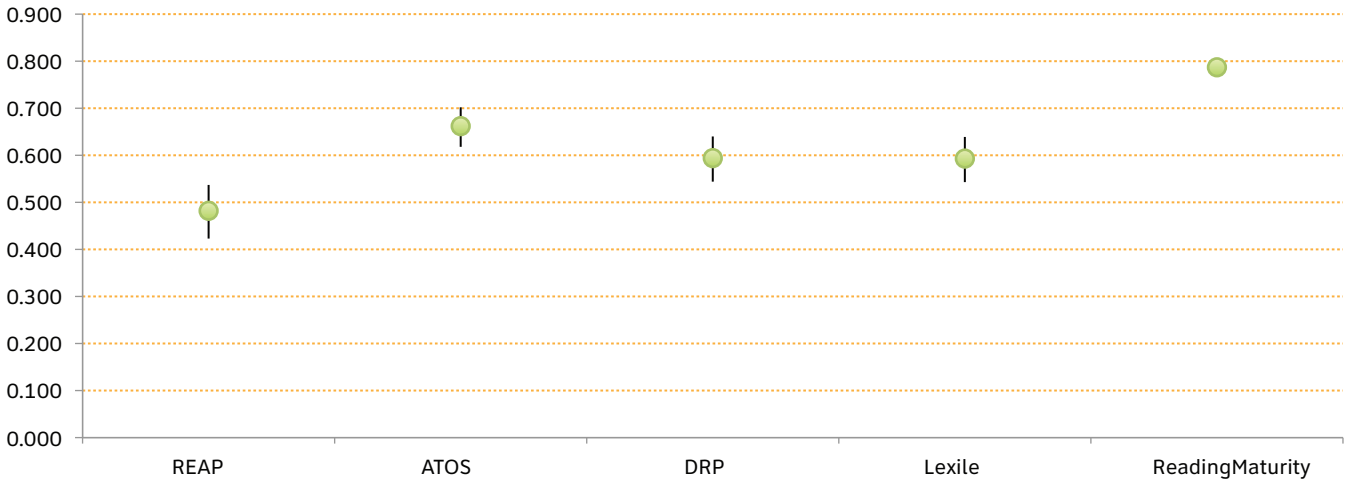
A standardized test for a given grade level contains texts that are relatively easy for the grade level, as well as texts that are relatively difficult for the grade level, and there is overlap in text difficulty from one grade level to the next. In addition, each state may use different standards for the difficulty of text used for testing at a given grade level. Therefore, although texts generally increase in difficulty from grade 3 to grade 12, a given text may not uniquely represent one specific grade. For example, not every 3rd grade text will be easier than every 4th grade text. This is in contrast to the Common Core exemplar texts, which were chosen to demonstrate increased text complexity at each band level.

ETS identified cases for which they expected the human-generated grade level to be less accurate, e.g. cases of short texts used as writing prompts or as practice test questions. SourceRater did not compute scores for these texts nor for texts that contain a mixture of narrative and informational text. (In contrast to the version of SourceRater available for this analysis, the current version of SourceRater can handle mixed genre texts.) An ETS-scrubbed version of each of the 285 texts that met the ETS criteria was distributed to each research group in order to have a comparison of all metrics, including SourceRater scores (See 4.2.5 for information about ETS text scrubbing.). New DRP and Reading Maturity scores were not provided for this subset of texts, so we computed scores for these 285 texts from the original full text set for those measures. (Questar determined independently that the ETS-scrubbed versions were identical to the DRP-scrubbed versions previously run through the DRP Analyzer for this subset of texts. See 4.2.3 for information about DRP text cleaning. Re-analysis therefore was not necessary.) We provide results for both the full set of texts and the subset of texts with SourceRater scores. Further subsets of the texts (split by text type and grade level) were taken from the full set, and, therefore, do not include SourceRater scores.

5.1.2.7 Results

Figure 5.1.2–1 shows the results for the full sample of texts, with Source Rater not represented, as explained above. Each rank order correlation (*rho*) is centered in its 95% confidence interval. As a group, the metrics were successful in predicting the state test grades. The Pearson Reading Maturity Metric produced text difficulty ranks that correlated *rho*=.79 with the grade level ranks of the state tests and showed no overlap of confidence interval with any other metric. At the low end, the confidence interval of REAP's .48 correlation also did not overlap with that of any other metric. The three intermediate metrics showed overlapping confidence intervals, with ATOS higher than Lexiles and DRP.

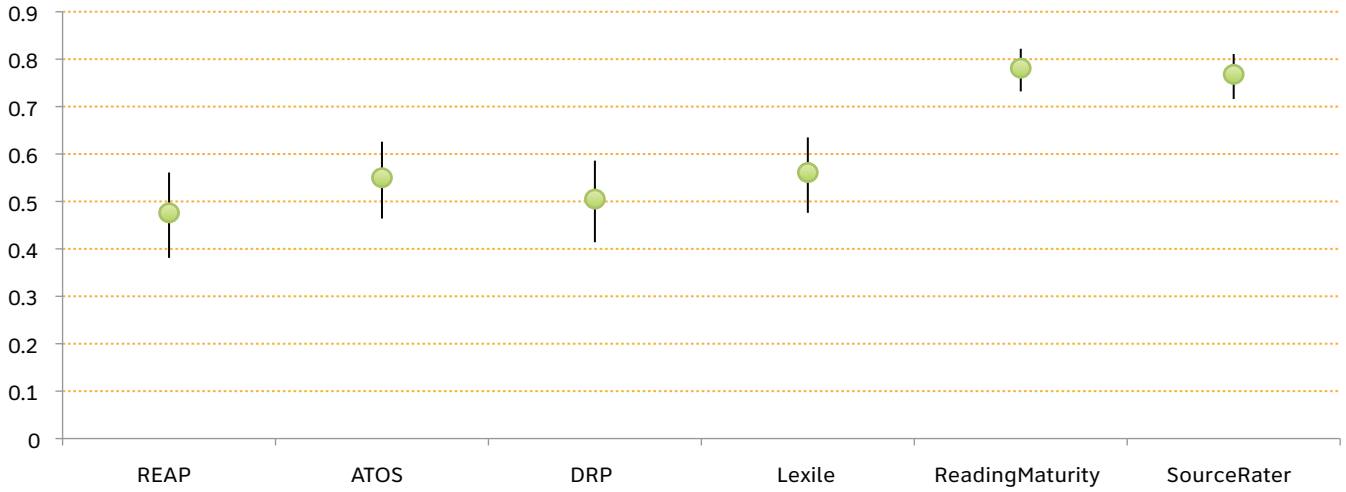
Figure 5.1.2–1: State Test Passages, Correlation with Grade Level (n=683)



95% Confidence Interval	REAP	ATOS	DRP	Lexile	Reading Maturity
Lower Limit	0.423	0.618	0.544	0.543	0.757
Rho	0.482	0.662	0.594	0.593	0.787
Upper Limit	0.537	0.702	0.640	0.639	0.813

The results for the subset of 285 passages scored by all metrics, including SourceRater, are shown in Figure 5.1.2–2. With this subset of texts, all the correlations are noticeably lower than the correlations obtained with the full sample, except for Reading Maturity, which, along with SourceRater produced higher correlations than the other metrics.

Figure 5.1.2–2: State Test Passages, ETS Subset, Correlation with Grade Level (n=285)



95% Confidence Interval	REAP	ATOS	DRP	Lexile	Reading Maturity	SourceRater
Lower Limit	0.381	0.464	0.414	0.476	0.732	0.716
Rho	0.476	0.550	0.505	0.561	0.781	0.768
Upper Limit	0.561	0.626	0.586	0.635	0.822	0.811

5.1.3 Stanford Achievement Test (SAT-9)

5.1.3.1 Initial Selection

Forty-seven passages from the Stanford Achievement Test (Pearson), Ninth Edition, Form S and 63 passages from Form T were distributed, totaling 110 passages.

5.1.3.2 Passages Removed for Analysis

Reason for Removal	Number Removed
Missing Data	12
Total Removed	12
Total Remaining	98

5.1.3.3 Missing Data

Metric	Number of Texts	Reason
SourceRater	12	Flagged for invalid end-of-sentence markers
Lexile	2	Non-prose text
REAP	All	No permissions (see notes)
DRP	1	Non-prose text

5.1.3.4 Text Properties

Average Number of Words	327.4
Grade Levels	1–11
Text Difficulty Measure(s)	Grade Level, Mean Rasch scores
Subsets Examined	Grades 1–5 vs. Grades 6-11

5.1.3.5 Reference Measures

Reference measures were the grade level of the test on which each passage appeared and the mean Rasch score of all question items pertaining to each text. Rasch scores model the probability that a given item is answered correctly as a function of both student skill and item difficulty. Two scores are generated from the Rasch model: a measure of student skill, based on the difficulty of items the student answered correctly (or incorrectly), and a measure of item difficulty, based on the skill of the students who answered the item correctly (or incorrectly). Model fitting involves iteratively adjusting these two scores until the model estimates stabilize.

5.1.3.6 Notes / Caveats

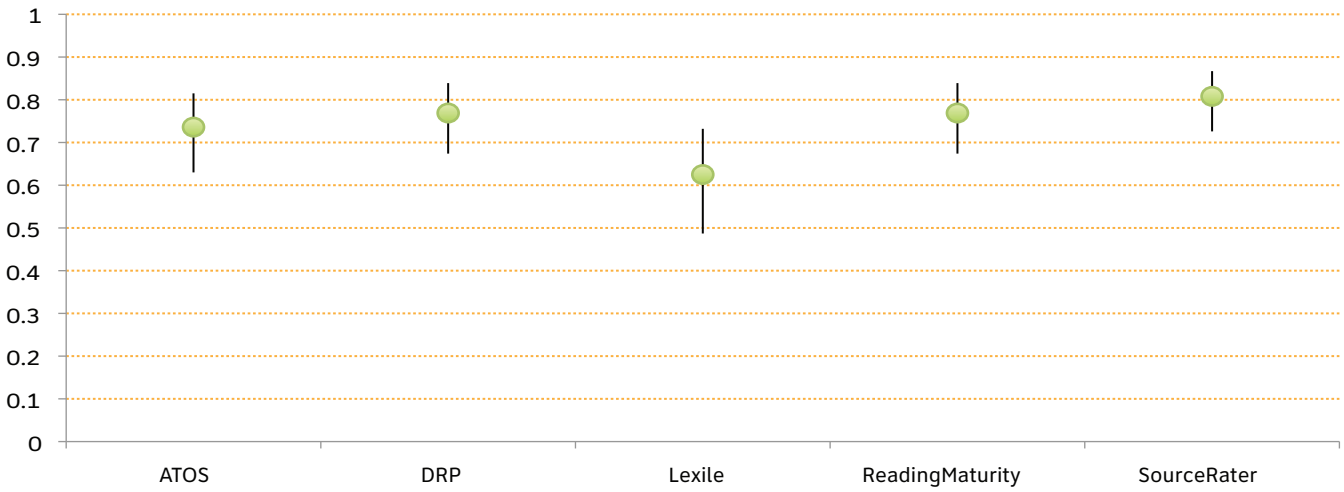
The mean Rasch score across all comprehension questions for a text depends not only on text difficulty, but also question difficulty. There is no assurance that each passage is followed by a set of equally difficult comprehension questions. However, mean Rasch scores provide a finer-grain measure of text difficulty than grade level and are based on student comprehension performance as opposed to human judgment of text difficulty.

Not all research groups received access to the SAT-9 texts because this required a legal agreement between institutions. For this reason, REAP scores are not available.

5.1.3.7 Results

The five metrics for which data were available were successful in predicting the grade level rankings of the SAT-9. (See Figure 5.1.3–1). The 95% confidence intervals all overlapped.

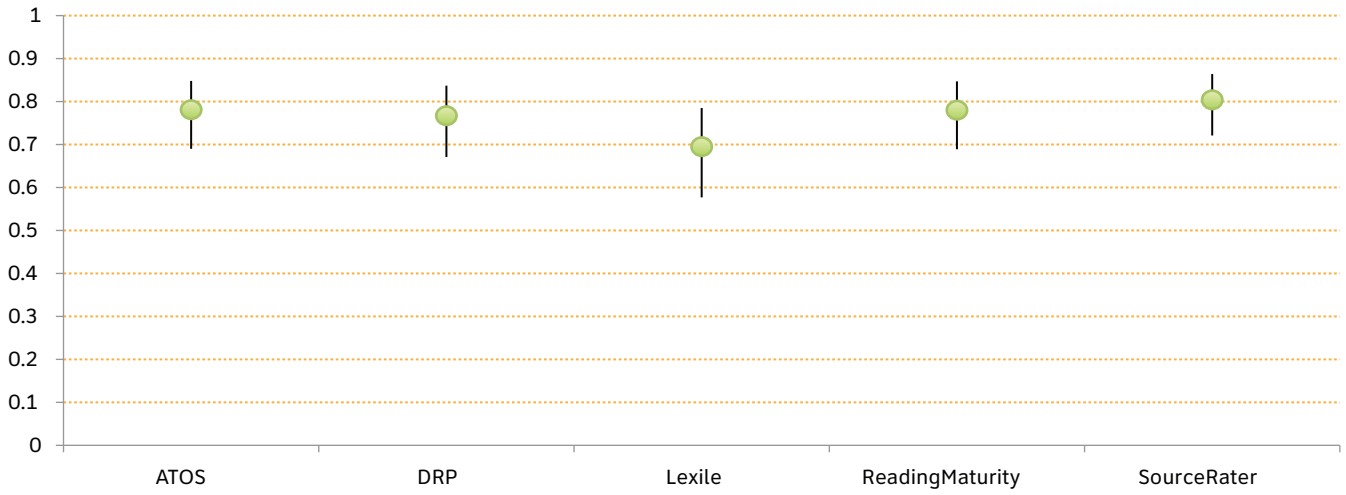
Figure 5.1.3–1: SAT-9, Correlation with Grade Level (n=98)



95% Confidence Interval	ATOS	DRP	Lexile	Reading Maturity	SourceRater
Lower Limit	0.630	0.674	0.487	0.674	0.726
Rho	0.736	0.769	0.625	0.769	0.808
Upper Limit	0.815	0.839	0.732	0.839	0.867

The correlations with Rasch scores are shown in Figure 5.1.3–2. The five metrics as a group were very successful in predicting the rank orders of the Rasch scores (*rhos* .7 to .8). The 95% confidence intervals overlapped completely.

Figure 5.1.3–2: SAT-9, Correlation with Rasch Scores (n=98)



95% Confidence Interval	ATOS	DRP	Lexile	Reading Maturity	SourceRater
Lower Limit	0.690	0.671	0.577	0.689	0.721
Rho	0.781	0.767	0.695	0.780	0.804
Upper Limit	0.848	0.837	0.785	0.847	0.864

5.1.4 Gates-MacGinitie Reading Test

5.1.4.1 Initial Selection

Ninety-seven passages from the Gates-MacGinitie Reading Test (Riverside Publishing) Form S were distributed. These consist of the reading comprehension passages for levels (grades) 1, 2, 3, 4, 5, 6, 7-9, 10-12, and AR (adult reading). No other components of the tests were used aside from the reading comprehension passages.

5.1.4.2 Passages Removed for Analysis

None.

5.1.4.3 Missing Data

None.

5.1.4.4 Text Properties

Average Number of Words	103.3
Grade Levels	1-Adult Reader
Text Difficulty Measure(s)	Grade Level, Mean Rasch scores
Subsets Examined	Grades 1–5 vs. Grades 6-adult

5.1.4.5 Reference Measures

The reference measures were the grade level of the test on which a passage appeared and the mean Rasch score for all question items pertaining to each text. Rasch scores model the probability that a given item is answered correctly as a function of both student skill and item difficulty. Two scores are generated from the Rasch model: a measure of student skill, based on the difficulty of items the student answered correctly (or incorrectly), and a measure of item difficulty, based on the skill of the students who answered the item correctly (or incorrectly). Model fitting involves iteratively adjusting these two scores until the model estimates stabilize.

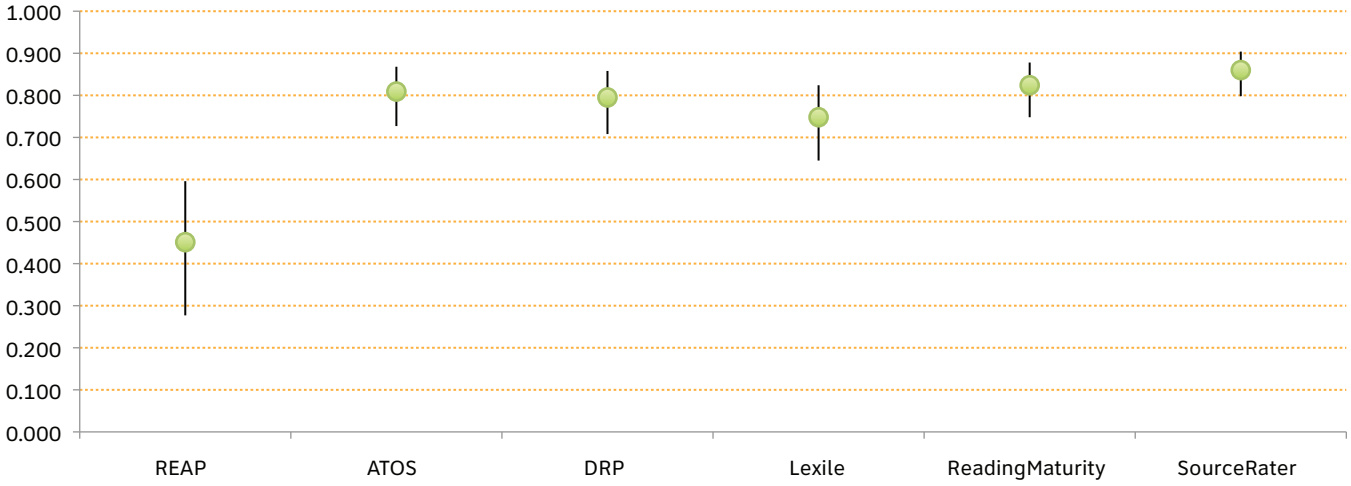
5.1.4.6 Notes / Caveats

Passages for the Gates-MacGinitie test were often very short, especially in the lower grades, which is a property known to make text difficulty estimates less reliable. In addition, question items in grades 1–2 follow each sentence rather than the entire text and consist of picture choices representing the meaning of the sentence.

5.1.4.7 Results

As can be seen in Figure 5.1.4–1, the metrics, with the exception of REAP, were very successful in predicting the grade level of Gates-MacGinitie passages. The 95% confidence intervals of SourceRater, Reading Maturity, Lexile, DRP, and ATOS overlapped, with *rho*s between .75 and .86.

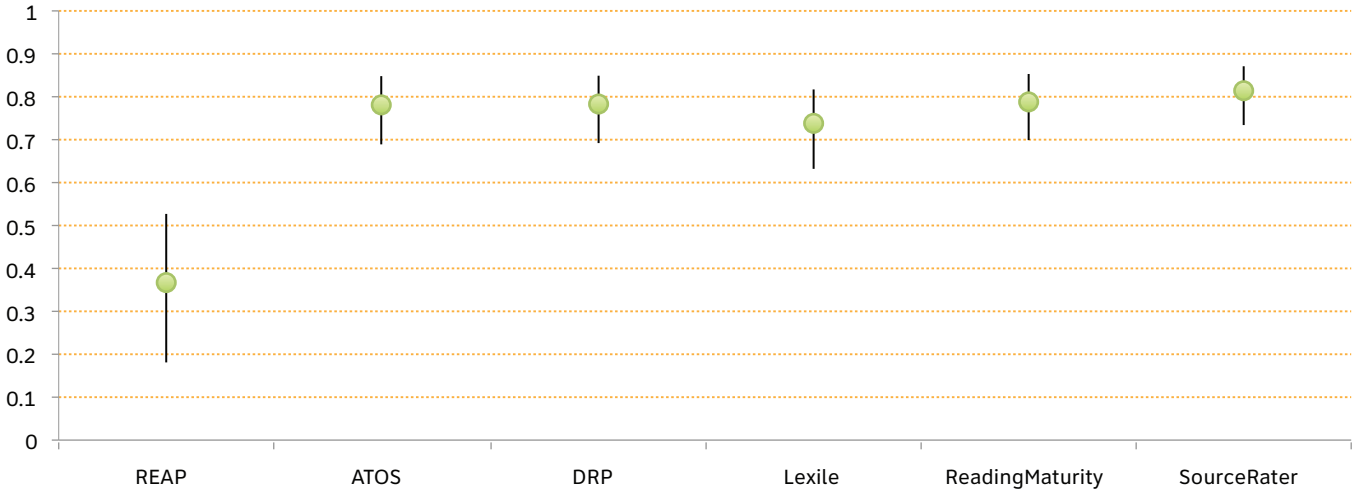
Figure 5.1.4–1: Gates-MacGinitie, Correlation with Grade Level (n=97)



95% Confidence Interval	REAP	ATOS	DRP	Lexile	Reading Maturity	SourceRater
Lower Limit	0.277	0.727	0.708	0.645	0.748	0.798
Rho	0.451	0.809	0.795	0.748	0.824	0.860
Upper Limit	0.596	0.868	0.858	0.824	0.878	0.904

A similar impressive result was obtained for Rasch scores, displayed in Figure 5.1.4–2. ATOS, DRP, Lexile, Reading Maturity, and SourceRater produced rank order correlations between .74 and .81 with overlapping 95% confidence intervals. REAP, again, was an outlier.

Figure 5.1.4–2: Gates-MacGinitie, Correlation with Rasch Scores (n=97)



95% Confidence Interval	REAP	ATOS	DRP	Lexile	Reading Maturity	SourceRater
Lower Limit	0.181	0.689	0.692	0.632	0.699	0.734
Rho	0.367	0.781	0.783	0.738	0.788	0.814
Upper Limit	0.527	0.848	0.849	0.817	0.853	0.871

5.1.5 MetaMetrics Oasis Passages

5.1.5.1 Initial Selection

Three hundred seventy-two passages from the MetaMetrics Oasis platform were distributed. The Oasis platform allows students to practice reading texts targeted to their reading level, while collecting responses to computer-generated multiple-choice fill-in-the-blank (cloze) items for each passage as the students read. The 372 informational passages comprise texts read by at least 50 different students with at least 1000 computer-generated items, allowing a stable empirical text complexity estimate.

5.1.5.2 Passages Removed for Analysis

None.

5.1.5.3 Missing Data

Metric	Number of Texts	Reason
DRP	101	< 125 words
SourceRater	All	ETS decision based on Oasis student/text sampling procedures

5.1.5.4 Text Properties

Average Number of Words	373.1
Grade Levels	N/A
Text Difficulty Measure(s)	Empirical complexity estimate based on cloze items
Subsets Examined	DRP subset (>125 words)

5.1.5.5 Reference Measures

The reference measure was empirical Lexile scores, as determined through modeling performance on cloze items as a function of student skill and text difficulty. Fitting the model begins with an estimate of text complexity based on the Lexile measure for the text. Student skill is then estimated based on the accuracy on a subset of texts that vary in difficulty around the Lexile score. Text difficulty is then re-estimated based on the skill of the readers who answered items correctly. This iterative process continues until the model estimates stabilize.

5.1.5.6 Notes / Caveats

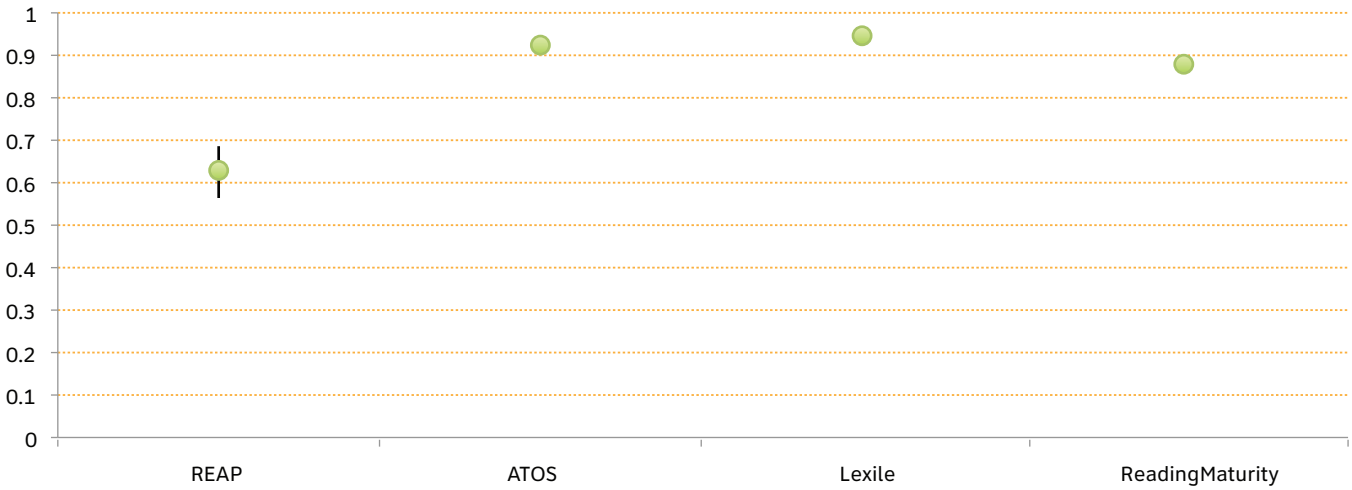
ETS did not provide scores on the grounds that the sampling procedure for the Oasis passages does not meet its standards for assessing text difficulty.

DRP measures were not provided for texts with fewer than 125 words. We provide results for the full set of 372 texts and also for the subset that includes DRP scores.

5.1.5.7 Results

Figure 5.1.5–1 shows the results for the four metrics that were applied to the Oasis passages. Lexile (.95), ATOS (.92), and Reading Maturity (.88) produced rank order correlations that were the highest observed for any of the text sets. The 95% confidence intervals were very short and overlapped only for Lexile and ATOS.

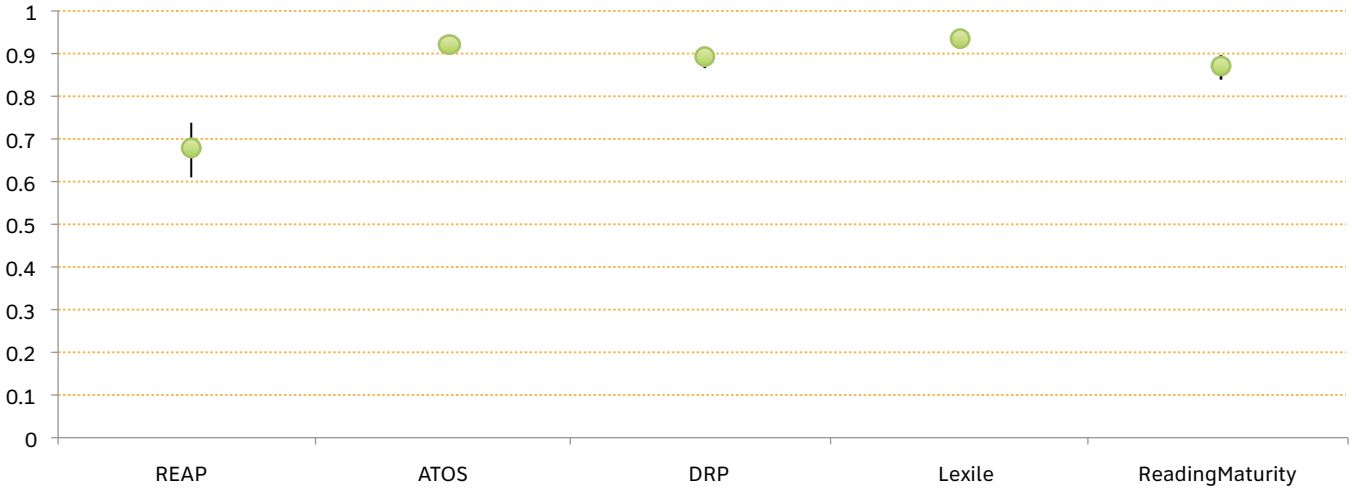
Figure 5.1.5–1: Oasis passages, Correlation with empirical Lexile (n=372)



95% Confidence Interval	REAP	ATOS	Lexile	Reading Maturity
Lower Limit	0.564	0.908	0.911	0.854
Rho	0.629	0.924	0.946	0.879
Upper Limit	0.686	0.937	0.939	0.900

For passages long enough for the DRP metric to be used, the picture is much the same, with DRP (*rho* =.89) joining Lexile (.95), ATOS (.92) and Reading Maturity (.88) as very high performers. These results are shown in Figure 5.1.5-2. The 95% confidence intervals overlapped for Lexile and ATOS.

Figure 5.1.5-2: Oasis passages, Passages with > 125 words only, Correlation with empirical Lexile (n=271)



95% Confidence Interval	REAP	ATOS	DRP	Lexile	Reading Maturity
Lower Limit	0.610	0.901	0.866	0.919	0.839
Rho	0.679	0.921	0.893	0.935	0.871
Upper Limit	0.738	0.937	0.914	0.948	0.897

5.2 Results by Text Type

5.2.1 Informational vs. Narrative Texts

The Common Core Exemplar Texts and state test passages were subdivided according to a text's status as informational or narrative. Identification of text types as informational, narrative, or mixed genre was determined by educator judgments on a passage-by-passage basis. Where states identified their passages by a particular genre type, that identification was generally retained in our study after review and confirmation.

Across the two text sets, the trend was that each metric was better correlated with grade level for the informational texts than for the narrative texts (see Figures 5.2.1–1 & 5.2.1–2). However, for the state test passages, Reading Maturity performed equally well and produced higher correlations on both types.

Figure 5.2.1–1: Common Core Exemplar Texts, Correlation with Grade Band, Narrative (n=65) vs. Informational (n=103)

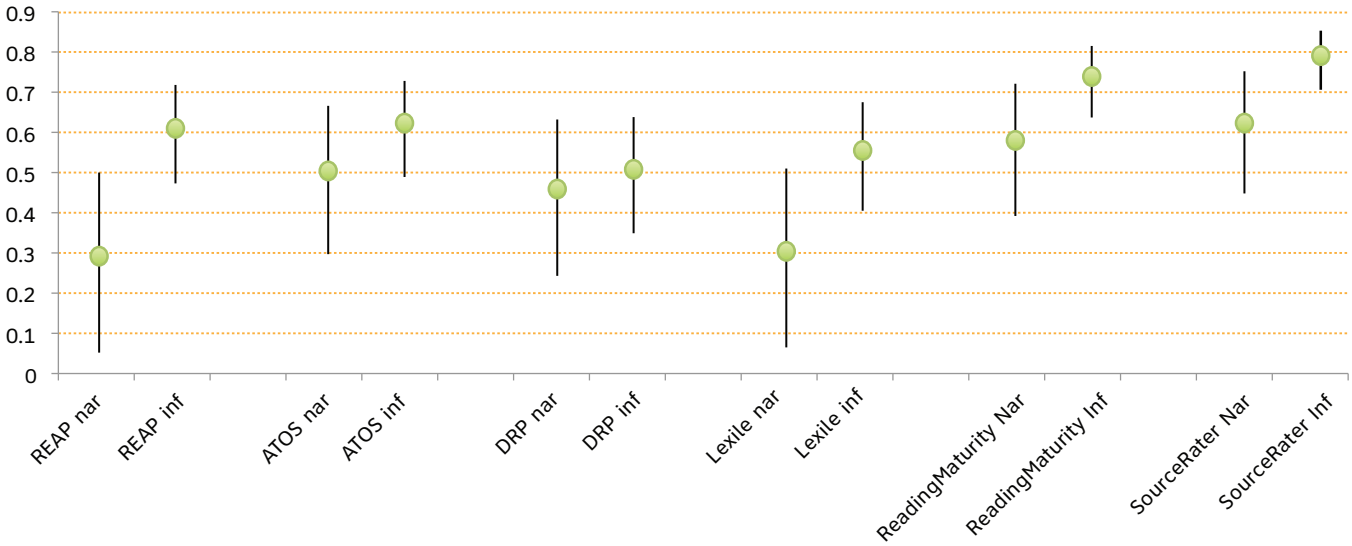


Figure 5.2.1–2: State Test Passages, Correlation with Grade Level, Narrative (n=275) vs. Informational (n=401)

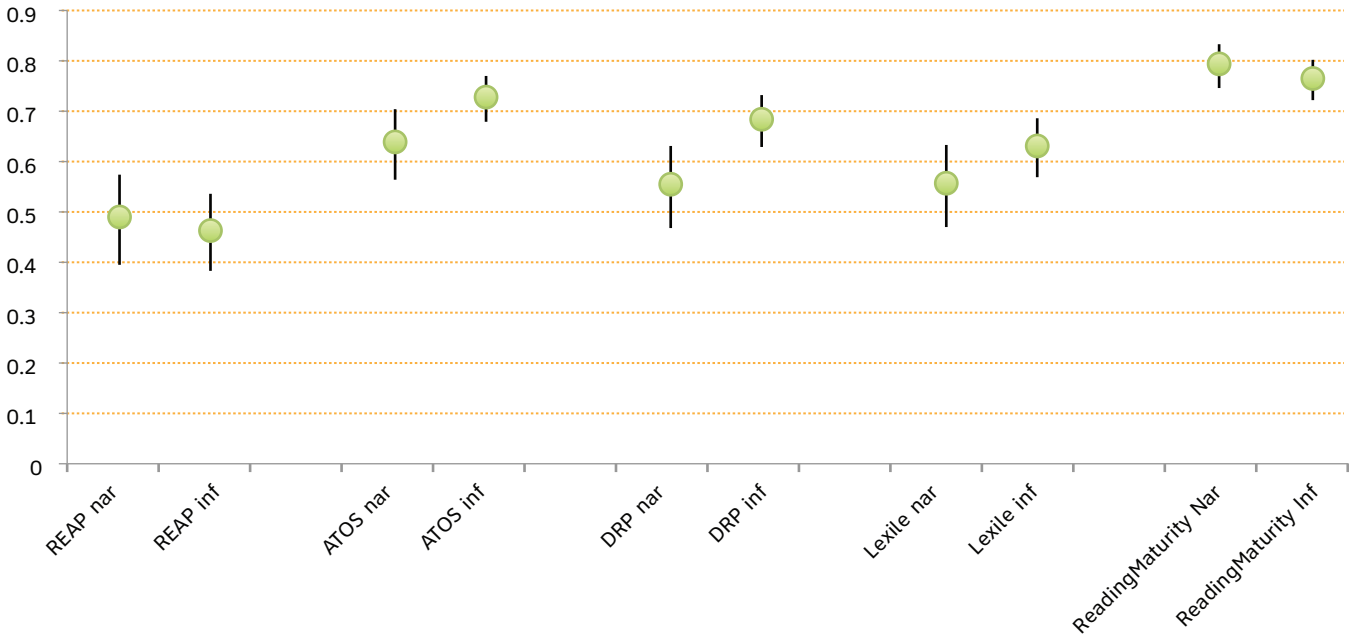


Figure 5.2.1–3 shows the mean value for each metric across expert-rated grade levels (the Common Core exemplar texts) separately for informational and narrative texts. (The Y axis is the average value of the metric at that grade level, rather than a correlation.) Generally, the complexity estimates for the two text types tended to diverge at the 6–8 grade band. Estimates of the complexity of narrative texts showed little increase from grade band 6–8 to band 9–10. However, all metrics showed some increase from grade band 9-10 to 11–CCR, and the SourceRater increase was especially large. These data are generated from small sample sizes.

Figure 5.2.1–3: Exemplar Text Metric Means by Text Type and Grade Level,
Informational n at each grade band = 9, 20, 19, 28, 28
Narrative n at each grade band = 11, 10, 23, 10, 10

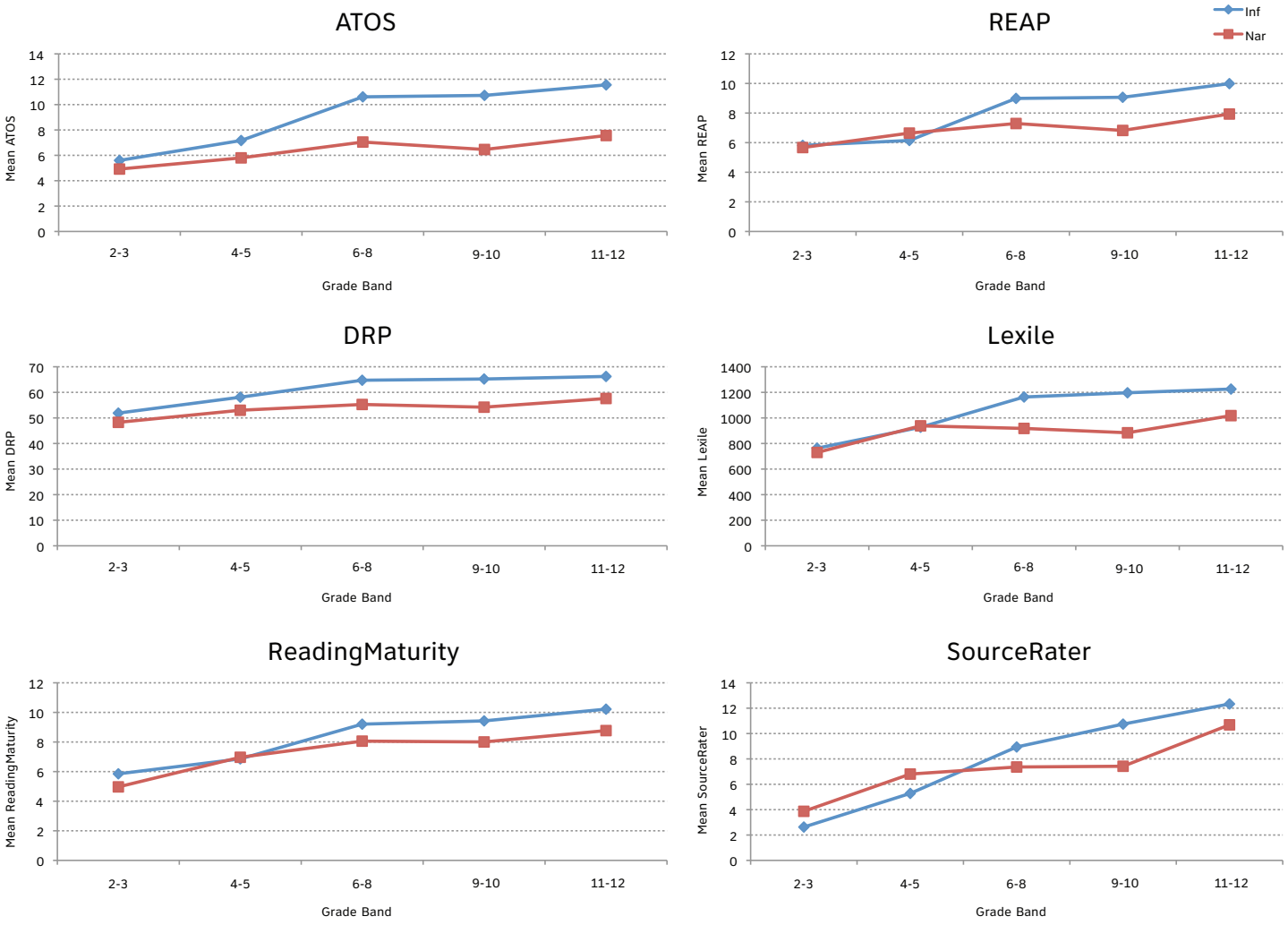
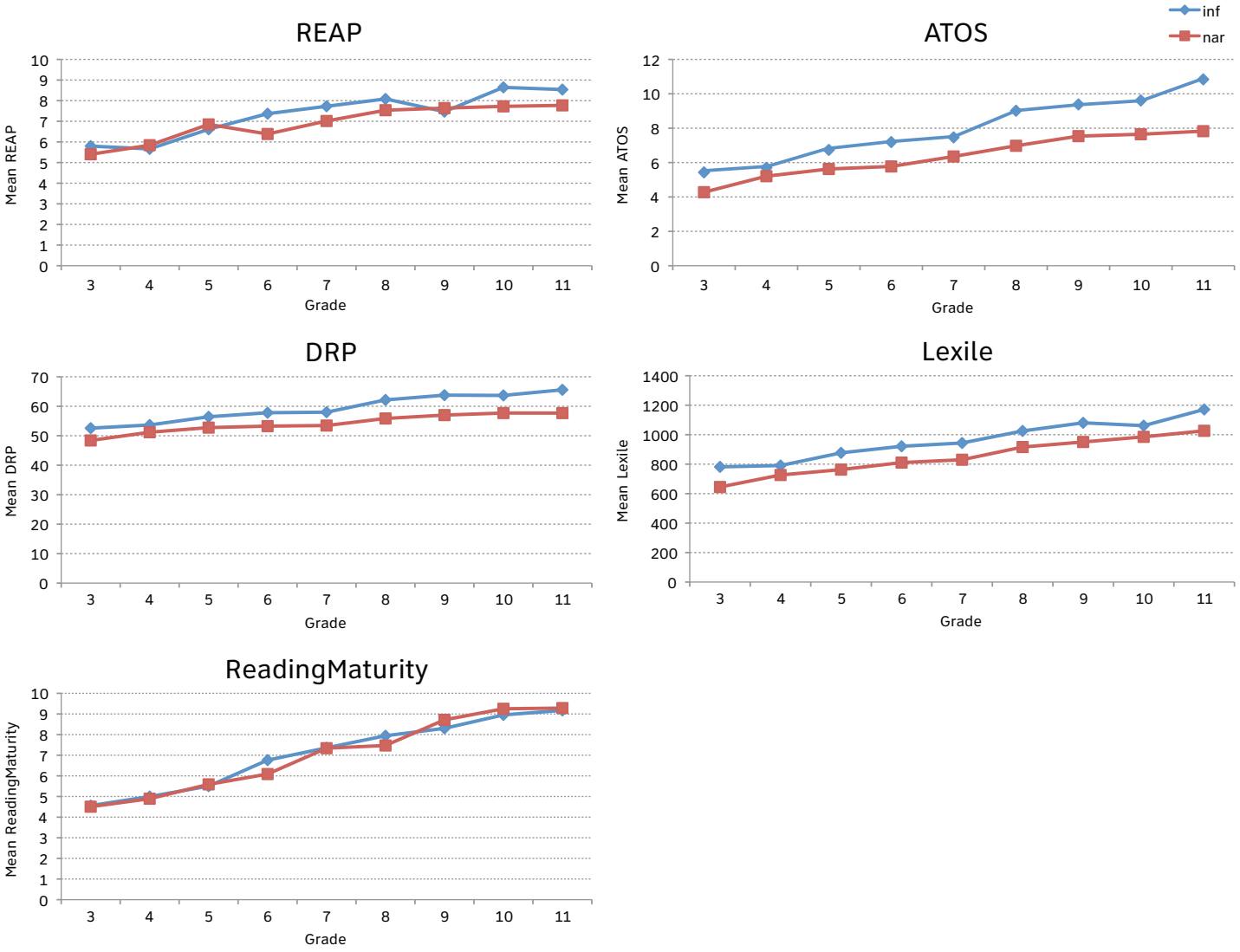


Figure 5.2.1–4 shows the same kind of text type x grade comparison for the state tests. Unlike the comparison for the exemplar texts, estimates for narrative as well as informational texts increase nearly monotonically with increasing grade levels. These data are based on larger sample sizes than the exemplar texts. Estimates are moderately and uniformly higher for informational than narrative texts across grades, except for ATOS, which shows increasing differences in the later grades and Reading Maturity, which shows no difference at any grade level between the two types. Informally, it also appears that Reading Maturity shows a more constant increment (linear slope) across grade levels for both text types.

Figure 5.2.1–4: State Test Passage Metric Means by Text Type and Grade Level
Informational *n* at each grade level = 37, 44, 47, 46, 58, 80, 17, 34, 38
Narrative *n* at each grade level = 40, 39, 42, 31, 13, 55, 19, 17, 19



5.2.2 Elementary vs. Upper Grades

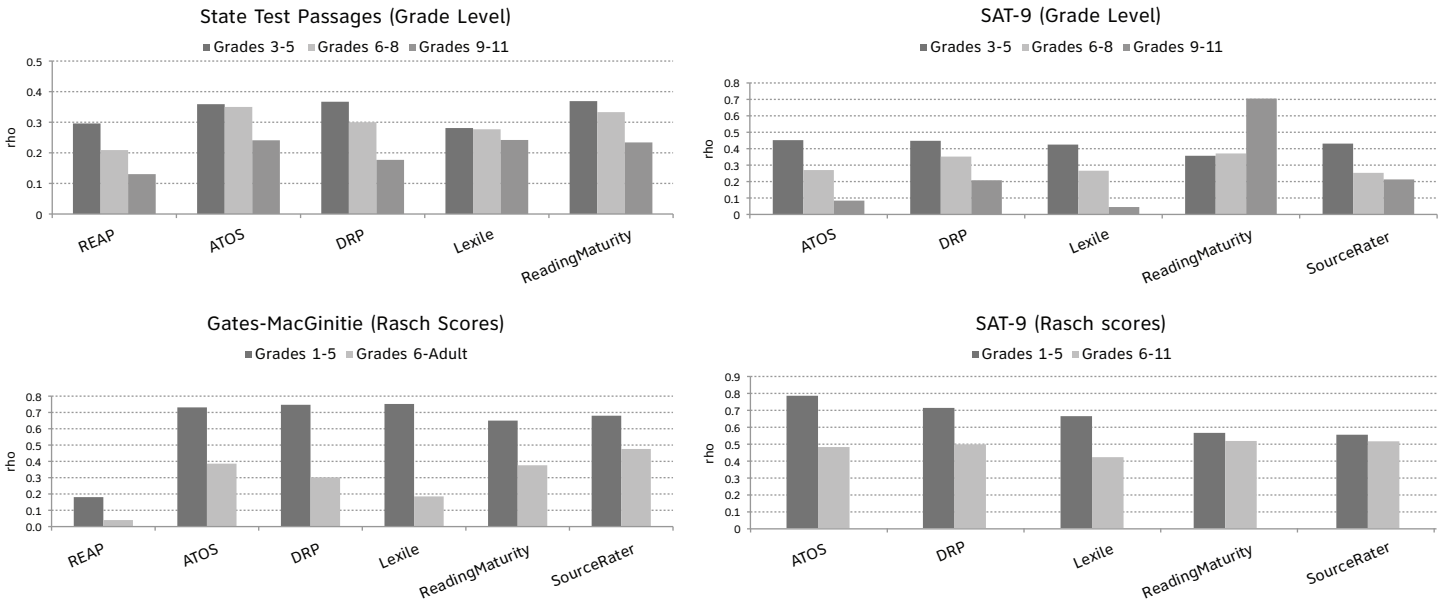
We compared the ability of each metric to discriminate among grades and student performance levels within broad grade bands. For text sets with grade level as the reference measure, we divided the texts into three equal groups of three grades (3–5, 6–8, 9–11) so that correlation coefficients would be comparable across the grade groupings. For the text sets with a continuous range of Rasch scores as the reference measure, we subdivided the scores into elementary grades (1–5) and upper grades (6+).

Of the four text sets that included grade as a reference measure, only two (the state test passages and the SAT-9) had texts that were leveled grade-by-grade for the full range of grades. Of the three text sets with student performance-based difficulty measures as the reference measure, only two (the Gates-MacGinitie and the SAT-9) also included grade level information that allowed us to form subgroups of data.

As shown in Figure 5.2.2–1, the metrics discriminate better among grades within lower grade bands than within higher grade bands. For example, among the state tests, discrimination is poorer among the three grades within the 9–11 grade band (i.e. grades 9, 10, and 11) than among grades within the lower bands of grades 3–5 and 6–8. This pattern is repeated for the SAT-9 grade level data, with the exception that the Pearson Reading Maturity Metric is more correlated with grade level in the 9–11 range than the lower grade ranges.

All metrics are more correlated with Rasch scores within grades 1–5 than within grades 6-adult for the Gates-MacGinitie test. This pattern repeats for ATOS and Lexile in the SAT-9 Rasch data.

Figure 5.2.2–1: Comparisons of within-grade-band correlations for lower and upper grade bands.

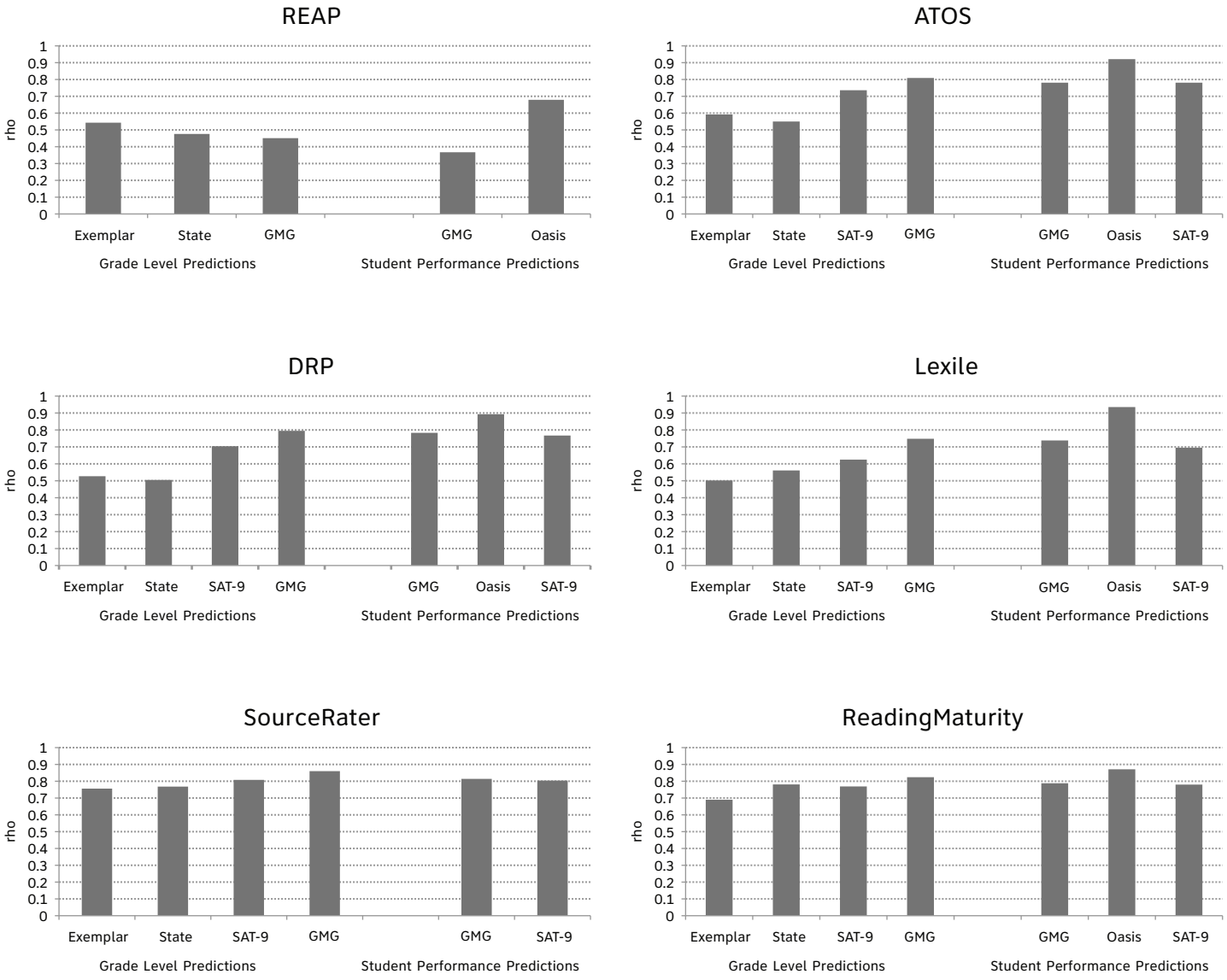


5.2.3 Grade Level vs. Student Performance Data

The data predicted by the metrics included a mix of grade levels and passage difficulty based on student performance data, and the metrics were generally successful at both kinds of predictions. Two sets of texts – the Gates-MacGinitie and the SAT-9 – could be compared on both grade level and performance. Figure 5.2.3–1 separates grade level reference measures from student performance reference measures.

As can be seen, ATOS, DRP, and Lexile showed overall better predictions for student performance than for grade levels. However, for the Gates-MacGinitie and the SAT-9, which included both student performance and grade levels, their predictions were similar. Their lower performance on grade level measures reflects lower predictions on the Common Core Exemplar texts and the state test passages. In contrast, SourceRater and Reading Maturity did well on both grade level and student performance measures. One implication is that measures that include more complexity indicators may be capturing some of what the human raters use (beyond vocabulary and sentence length) when they rate grade levels.

Figure 5.2.3–1: Comparisons of correlations with grade level vs. student-performance based reference measures



5.3 Variability among the Metrics

Overall, the metrics were successful in predicting grade levels and Rasch scores of text difficulty based on student performance. However, there were some apparent differences among the metrics that we summarize here, highlighting comparisons in which 95% confidence intervals were non-overlapping. One notable pattern involves two groupings of metrics based on their shared use of related difficulty measures. One group includes the three metrics that rely primarily on word difficulty (word frequency or grade level) and sentence length (ATOS, DRP, and Lexile), while adding variables such as word length (ATOS, DRP) and within sentence punctuation (DRP). The second group (SourceRater and Pearson's Reading Maturity) also uses measures that reflect word frequency, word length and sentence length. However, they add a broader range of linguistic predictors, including text coherence (both Source Rater and Reading Maturity), word meaning features (Source Rater), syntactic measures (both), paragraph length (both), and text genre (Source Rater), among others.

The metrics within these two groups tend to pattern together in their correlations with reference measures. In almost all cases, ATOS, DRP, and Lexile were similar in their correlations with reference measures. Similarly, SourceRater and Reading Maturity were comparable in their correlations with reference measures.

Comparing these two groups, two observations emerge. First, the two groups showed comparably high correlations with a number of reference measures. Second, when there were differences between the two groups, they tended to favor the metrics in the second group (Reading Maturity and SourceRater). Only for the Oasis passages did ATOS, Lexile, and DRP (the first group) show higher correlations than Reading Maturity (there were no data for SourceRater for these passages as noted in section 5.1.5).

For example, SourceRater and Reading Maturity were more highly correlated with grade level of the Common Core exemplar texts and state test passages than were ATOS, Lexile, and DRP. For the state grade levels, Reading Maturity was more highly correlated with the grade levels than were ATOS, Lexile, and DRP, and this was also true for the subset of narrative state test texts. For the informational texts, Reading Maturity was correlated more highly with grade level than was Lexile (although not higher than ATOS or DRP).

For the Common Core Exemplar texts, Reading Maturity tended to show higher correlations with text grade levels, including the subset of informational texts, although the confidence intervals overlapped

with those of other measures in some comparisons. SourceRater showed higher correlations with the Common Core Exemplar text grade levels than any metrics of the first group. SourceRater showed higher correlations than either DRP or Lexile for the informational subset of these texts. For the subset of state test passages that ETS analyzed, SourceRater was more highly correlated with grade level than all the metrics of the first group.

In several cases, the REAP measure, compared with the other metrics, was less correlated with reference measures of text difficulty. This was true for correlations with grade levels on the state test passages, for the Gates-MacGinitie grade levels and Rasch scores, especially for grades 1–5, and for the Oasis observed Lexile scores. It should be noted, however, that the primary purpose of the REAP project is to assist instructors in searching for texts on the web that satisfy specific lexical constraints while matching individual students' interests. So, while grade level computations are a part of this matching process, it is not REAP's primary objective. Given that REAP uses measures similar to those used by other text tools (including word frequency, word length, and sentence length), it is likely that the difference in correlations comes from less extensive norming to outside measures compared to the other metrics.

5.4 Coh-Metrix

The Coh-Metrix Text Easability Assessor gauges texts along five dimensions, which the developers characterize as follows:

Narrativity: The degree to which a text is story-like. It relies on indicators such as word familiarity, use of pronouns, the ratio of verbs to nouns, and many other countable factors that are characteristic of stories more than informational texts.

Referential cohesion: The degree of co-reference (word overlap and pronouns) across the text.

Syntactic simplicity: How short and familiar the syntactic structures used in the text are. Texts with shorter sentences and clauses and more familiar structures will have high scores on the syntactic simplicity dimension.

Word concreteness: The relative numbers of concrete (perceptible in reality), imageable (evocative of a mental image), and "meaningful" (associated with the meanings of other words) words in the text.

Deep cohesion: The degree to which causal, temporal, and logical connectives are present in the text.

Figure 5.4–1 shows Spearman's *rho* correlation of each of these dimensions with grade level, and Figure 5.4–2 plots each dimension against grade level for each text set. When texts were grouped in grade bands, the data point was plotted in the middle of the grade-band (for example, the mean narrativity for the Common Core exemplar grades 2–3 is plotted at grade 2.5).

We should note that the correlations between single dimensions and grade level have not taken into account the shared variance of the dimensions, which means attributions about specific dimensions contributing uniquely to grade levels is unwarranted. With this caveat in mind, we can nevertheless see that syntactic simplicity was the dimension most correlated with grade level, with higher graded text having more complex syntax. For most of the text sets, referential cohesion and narrativity were also correlated with grade level, with more cohesive texts and more narrative texts in the younger grades (this correlation approaches statistical significance for the SAT-9). However, narrativity was not reliably correlated with grade level for the Common Core exemplar texts. Figure 5.4–2 reveals that the Common Core exemplar texts tend to maintain a slightly higher degree of narrativity in the upper grades and a slightly lower degree of narrativity in the lower grades compared with other text sets, resulting in a more constant degree of narrativity across the grades.

Word concreteness was reliably correlated with grade level only for the Common Core exemplar texts and the state test passages. Figure 5.4–2 shows that word concreteness was lower overall for these text sets in comparison to the others, and that the Common Core exemplar texts steadily increase in abstractness (decrease in concreteness) as grade band increases.

Deep cohesion was reliably, but only weakly correlated with the state test passages, with more deep cohesion (more connectives) in the upper grades.

Figure 5.4–1: Correlation of Coh-Metrix dimensions with grade level

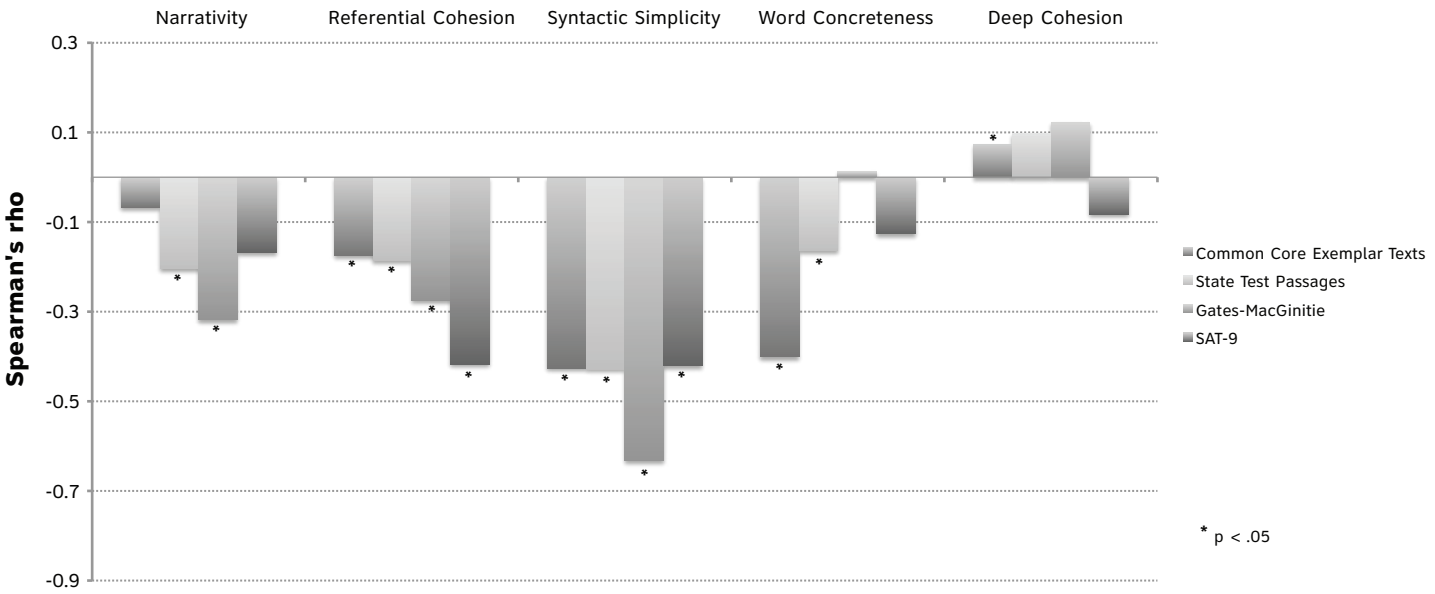


Figure 5.4–2: Coh-Metrix dimension percentiles by Grade Level

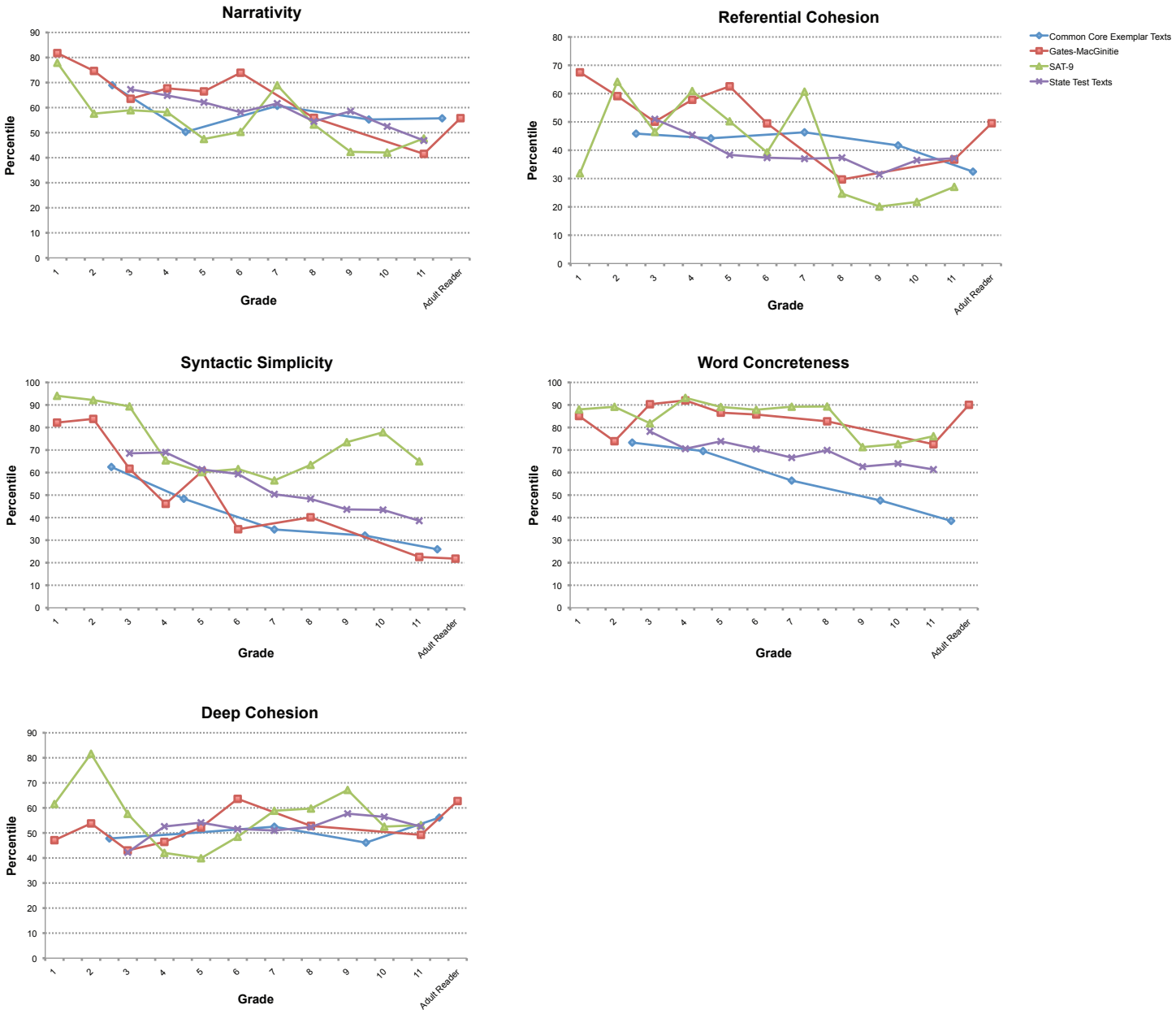
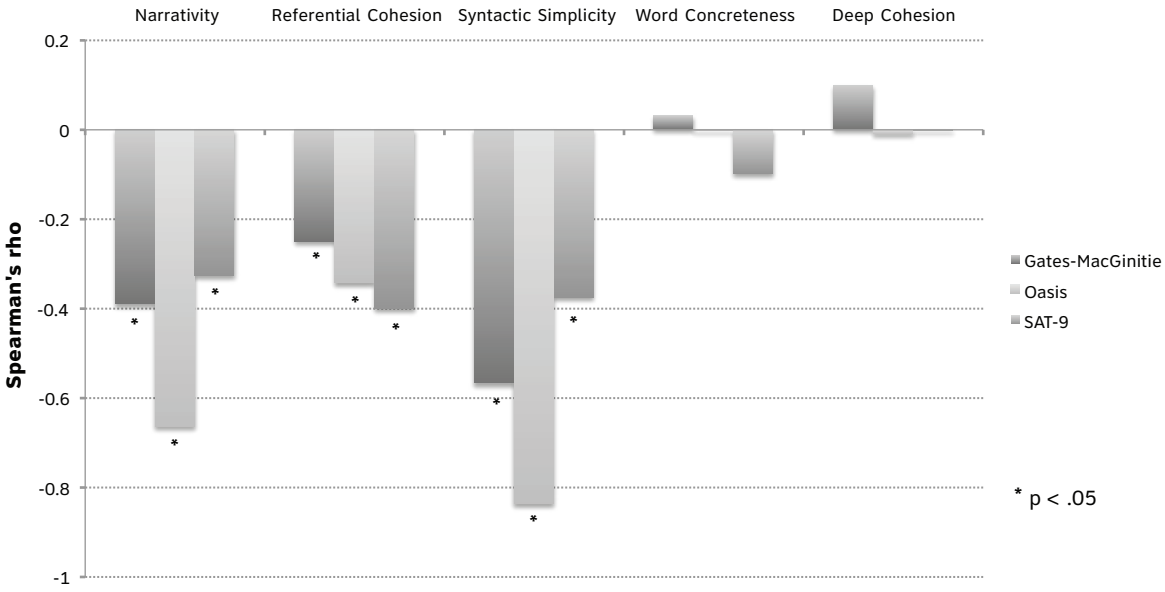


Figure 5.4–3 shows that all three text sets with student performance-based reference measures generally show the same pattern. Difficulty is predicted most strongly by syntactic simplicity, followed by narrativity (which was especially highly correlated with performance on the Oasis passages), followed by referential cohesion. Word concreteness and deep cohesion were not reliably correlated with performance-based measures of text difficulty.

Figure 5.4–3: Correlation of Coh-Metrix dimensions with student performance-based text difficulty measures



6 CONCLUSIONS

All of the metrics were reliably, and often highly, correlated with grade level and student performance-based measures of text difficulty across a variety of text sets, and across a variety of reference measures. Some differences among the correlations were also observed. In most cases, informational texts were more highly correlated with the metrics than were the narrative texts. Further examination showed that for the Common Core exemplar texts, the metrics' estimates of narrative text difficulty tended to flatten from grade band 4–5 to grade band 9–10, whereas the estimates of informational text difficulty flattened from grade band 6–8 to grade band 9–10. The interpretation of this finding is that the expert educators who chose the exemplar texts for the Common Core Standards distinguished between these middle grade bands based on something other than what is currently measured and weighted heavily in the metrics we evaluated. It could be that subject matter and themes, for example, lead educators to select which texts are appropriate for 6th through 8th grade. Of note is that SourceRater estimates of informational text difficulty did not flatten for this dataset. This implies that for informational texts, SourceRater was able to quantify sources of variation between texts in each grade band that are correlated with how experts classify texts into grade bands.

For the state test passages, the metrics estimated the narrative texts as less difficult than the informational texts across most grade levels, and the correlations with grade level were also lower for the narrative texts compared to the informational texts. The Pearson Reading Maturity Metric was the only measure that did not show this difference, with equally high correlations for both types of text. (SourceRater was not tested.)

The metrics showed the highest correlations with the grade levels of more uniform sets of standardized test passages (i.e. the SAT-9 and Gates-MacGinitie grade levels). The consistency and careful norming that are expected from a single standardized test probably explains the higher correlation with grade levels from these two individual tests. In contrast, the state test passages, which produced lower correlations, included a mix of standardized tests across a variety of state standards. Correlations were also generally somewhat lower for the Common Core exemplar texts, suggesting that the experts' selection of the exemplar texts relied on judgment factors beyond the text features measured by the metrics. However, the metrics that used a broader range of variables

did better on the Common Core texts than those that rely mainly on word difficulty and sentence length. The lack of some third reference point (the elusive "gold standard") means we cannot privilege either expert ratings or the text difficulty measures. Student performances on these texts, which of course we do not have, might correlate more with one or the other. Nevertheless, the fact that the two metrics that used a broader range of linguistic and text variables did well on the exemplar texts (correlations around *rho*=.7) suggests that a substantial portion of subjective expert judgments can be captured by an extended range of objective text measures.

Differences in which aspects of the texts were correlated with grade level may also explain the varied performance in predicting grade level across reference text sets. The Coh-Metrix analysis revealed that the Gates-MacGinitie and SAT-9 were both correlated with the same Coh-Metrix dimensions: narrativity, referential cohesion, and syntactic simplicity. Narrativity and syntactic simplicity correlate with the word difficulty and sentence length variables that are used by all metrics, which may help explain why these two reference measures were the best predicted grade level measures across all the metrics.

In the Coh-Metrix analyses, the Common Core Exemplar text grade levels were not reliably correlated with narrativity, but were reliably correlated with word concreteness, and the state test passage grade levels were reliably correlated with all of the dimensions. This may help explain why the metrics using a broader range of linguistic and text variables were more highly correlated with grade level for the Common Core exemplar texts and the state test passages than were metrics using mainly word difficulty and sentence length. For example, the use of word meaning features (e.g. concreteness), despite being correlated with word frequency, may capture additional features of text difficulty that affect expert judgment and student performance.

All of the metrics were highly correlated with text difficulty measures based on student performance, including performance on both cloze test items and multiple-choice comprehension questions. Each of the performance-based difficulty measures was correlated with the same three Coh-Metrix dimensions; narrativity, referential cohesion, and syntactic simplicity, which were the same three features most correlated with the Gates-MacGinitie and SAT-9 grade levels.

In addition, the pattern was generally that the metrics were better able to predict grade level and comprehension performance in the lower grades compared with the upper grades. This may reflect increased variance among factors determining grade levels and especially Rasch scores in the upper grades. At the upper grades, one expects more non-

systematic (individual student) differences in knowledge and also in the sources of information needed to answer questions (greater variance in knowledge-based information and inferences).

An important conclusion is that the metrics as a whole, despite some clear differences among them, performed reasonably well across a variety of reference measures. This robustness is encouraging for the possibilities of wide application for most of the measures. The results also confirm the continued viability of traditional components of readability (word difficulty and sentence length) for assessing text difficulty, especially when the assessment of difficulty includes standardized tests and student performance on these tests. Even metrics using a broader range of measures include the word difficulty and sentence level measures that are basic to readability measurement. However, the measurement components these broader-range metrics add allow some gains in predictions. Indeed, the broader-range metrics showed particularly robust performance, with correlations of *rho*=.80 for most reference measures and the lowest correlation for a full text set at .69.

The question of whether objectively measured text complexity predicts student performance was answered in the affirmative. Indeed the metrics were at their best when they were predicting measures that included student performance. However, this predictive value tended to decrease for higher grades, where unmeasured factors play an increasing role. To what extent the features that make text complex include additional features – beyond syntax and vocabulary – that make text difficult remains to be determined. Similarly, more work is needed to understand how the features that make texts difficult for readers change with grade levels. More research, with larger sets of student performance data and text samples at the upper ranges, must be a near term priority for the field.

We close with a reminder that the results of this study, and hence our conclusions, are limited by the sets of reference data the study was able to obtain. Considering the vast universe of texts and student performances on them, this is more than the usual caveat about the limits of any study.

7 EDUCATIONAL IMPLICATIONS

The potential for applying text difficulty measures to education is much the point of this research. These applications are many, and there is more to consider in a comprehensive accounting than we can provide here. Instead, we want to highlight a few that deserve further discussion.

The first implication is that the success of text difficulty metrics in predicting reference measures of text grade levels and student performance means that a wide range of applications consistent with the goals of the Common Core Standards can be supported by text difficulty tools. This applies both to tools based on the variables of traditional readability formulae (word difficulty and sentence length) and those with broader indicators of text complexity. It applies also to the work of school-based curriculum text selection, publishers' attempts to meet curriculum standards, and reading assessment.

An especially interesting application is the use of text difficulty tools in reducing the large gap that currently exists between typical high school texts and college texts (Appendix A of Common Core Standards). How to close this gap, in effect to recalibrate text difficulty across grade levels, is an issue to address carefully, but a systematic approach will certainly include the application of text difficulty measures of the kind we have studied. The variability that the metrics showed in differentiating among the higher grades (8–12) is a factor in how a given metric can be recalibrated to close the gap. A common scale, based on this study and including the metrics examined here, has been published and is included as Appendix C. Based on this common scale, Appendix F shows the trajectory towards college and career readiness for each metric along with the metric's mean score on a selection of career-oriented and first year college texts for comparison.

Another implication is that the variety of applications benefits from the variety of tools. For example, whereas tools based on readability variables will serve a variety of practical purposes, educators charged with curriculum design might be interested in gaining a finer grain view of the properties of a set of texts. A text complexity measure such as Coh-Metrix, provides, rather than a single measure of text difficulty, information on specific dimensions that distinguish among texts and, thus, would be useful for this purpose.

Beyond these clear practical implications are some that are subtler. The broader-variable text metrics, which were nearly always as accurate

and sometimes more accurate than metrics based primarily on word difficulty and sentence length, provide useful measures of the degree to which a text has features of deep cohesion that support integration across sentences. When these features are absent, the reader may be called on to do more text work on his or her own in order to achieve a coherent understanding. It is possible, based on the Coh-Metrix measures, that the absence of deep cohesion features has little impact on student performance. (By contrast, superficial or referential cohesion did matter.) Why this is the case is a matter for more research. It is possible that deep cohesion does matter more in texts with technical science content for example, where tracking causal relations is critical. However, there are other possibilities: one is that word meaning and syntactic complexity, whose importance is confirmed across all the metrics in this study, are more powerful factors in student performance than deep cohesion.

Students with a sufficient lexicon and ability to comprehend syntax may have the capacity to make the connections needed to comprehend the text without explicit text connectors. If this analysis is correct, it may suggest the value of more practice with texts containing more complex syntax, and it reinforces the call for more and better vocabulary instruction, which apparently has continued, long after observations on this problem by Durkin (1979), to be a small and unsystematic part of literacy instruction (Scott & Nagey, 1997; Blachowicz & Fisher, 2000; Biemiller, 2001).

However, there is another perspective to be considered. It could be that some of the features of text that do make a difference if they are truly absent are just not "absent enough" in well-written texts. Curriculum and test designers, compared with a random, average writer, may be more careful to make or choose texts sufficiently coherent, at the deep as well as the surface level. In effect, the variability in explicit text features that matter for cohesion might not be large enough for the importance of these features to be detected in a sample of well-written texts.

Another implication of this work is the contrast between narrative and informational texts. The greater success of metrics that primarily use word difficulty and sentence length in predicting reference measures for informational texts than narrative texts suggest that text types are important in considering application of text difficulty tools. The ability of some of the tools (those with a broader range of variables) to do well on both text types is one of the more interesting outcomes of this study. It is not surprising that measures based primarily on word difficulty and sentence length capture properties of narrative texts imperfectly. It may be surprising that these measures when combined with other linguistic

and text factors do rather well. The subjective judgment of which works of fiction are appropriate for which grade levels is a complex issue, and one might assume that no quantitative measures could approximate these judgments. That was not the case here. The evidence suggests some alignment between measures beyond word difficulty and sentence length and judgments of grade level.

We conclude our brief treatment of implications by pointing out that the success of the quantitative measures provided by the metrics we studied does not mean there is no role for qualitative analysis. It is rightly valued in the Common Core State Standards. There are genres, notably poetry and drama, whose "difficulty" involves factors that are not readily measured by most, perhaps all, of the metrics considered in this study. Their placement in the curriculum, at least for now, must be done by human judgment and use of qualitative rubrics. The selection of texts for specific grade levels, as opposed to broader grade bands, might well benefit from systematic use of qualitative rubrics. This possibility, as well as numerous other practical, important questions requires further consideration and, where possible, some real research.

REFERENCES

- ACT, Inc. (2006). Reading between the lines: What the ACT reveals about college readiness in reading. Iowa City, IA: Author.
- ACT, Inc. (2009). The condition of college readiness 2009. Iowa City, IA: Author.
- Biemiller, A., & Slonim, N. (2001). Estimating root word vocabulary growth in normative and advantaged populations: Evidence for a common sequence of vocabulary acquisition. *Journal of Educational Psychology*, 93, 498–520.
- Blachowicz, C. L. Z., & Fisher, P. (2000). Vocabulary instruction. In M. L. Kamil, P. B. Mosenthal, P. D. Pearson, & R. Barr (Eds.), Handbook of reading research: Vol. 3 (pp. 503–523). Mahwah, NJ: Lawrence Erlbaum.
- Caruso, J.C., & Cliff, N. (1997). Empirical size, coverage, and power of confidence intervals for Spearman's *rho*., *Educational and Psychological Measurement, 57*(4), 637–654.
- Council of Chief State School Officers. (2010, June 2). The standards. Retrieved from http://corestandards.org/the-standards
- Durkin, D. (1979). What classroom observations reveal about reading comprehension. *Reading Research Quarterly*, *14,* 518–544.
- Scott, J.A. and Nagy, W.E. (1997). Understanding the definitions of unfamiliar verbs. *Reading Research Quarterly, 32,* 184–200.
- Heilman, M., Collins-Thompson, K., Callan, J., & Eskenazi, M. (2006). Classroom success of an intelligent tutoring system for lexical practice and reading comprehension. *Proceedings of the Ninth International Conference on Spoken Language Processing.* Pittsburgh, PA.
- Sheehan, K.M., Kostin, I., Futagi, Y., & Flor, M. (2010, December). Generating automated text complexity classifications that are aligned with targeted text complexity standards. (Publication No. RR-10–28), Princeton, NJ: Educational Testing Service.

APPENDIX A

Full Results Table: Spearman's *rho*

	n	REAP	ATOS	DRP	Lexile	RM	SR
CC Exemplar, All	168	0.543	0.592	0.527	0.502	0.690	0.756
CC Exemplar, Informational	103	0.610	0.623	0.508	0.555	0.739	0.791
CC Exemplar, Narrative	65	0.292	0.504	0.459	0.304	0.580	0.623
State Tests, All	683	0.482	0.662	0.594	0.593	0.787	
State Tests, ETS Subset	285	0.476	0.550	0.505	0.561	0.781	0.768
State Tests, Grades 3–5	254	0.296	0.359	0.367	0.281	0.369	
State Tests, Grades 6–8	285	0.209	0.350	0.300	0.277	0.333	
State Tests, Grades 9–11	144	0.130	0.241	0.177	0.242	0.234	
State Tests, Informational	401	0.463	0.728	0.684	0.631	0.765	0.781
State Tests, Narrative	275	0.490	0.639	0.555	0.557	0.794	0.756
GMG Grade Level, All	97	0.451	0.809	0.795	0.748	0.824	0.860
GMG Rasch, All	97	0.367	0.781	0.783	0.738	0.788	0.814
GMG Rasch, Grades 1–5	53	0.181	0.731	0.747	0.752	0.650	0.680
GMG Rasch, Grades 6-adult	44	0.040	0.386	0.302	0.185	0.376	0.476
SAT-9 Rasch, All	98		0.781	0.767	0.695	0.780	0.804
SAT-9 Rasch, Grades 1–5	41		0.784	0.712	0.663	0.564	0.553
SAT-9 Rasch, Grades 6-11	57		0.480	0.496	0.420	0.516	0.514
SAT-9 Grade, All	98		0.736	0.769	0.625	0.769	0.808
SAT-9 Grade, Grades 3–5	34		0.452	0.448	0.425	0.357	0.431
SAT-9 Grade, Grades 6–8	38		0.270	0.352	0.266	0.371	0.253
SAT-9 Grade, Grades 9–11	19		0.084	0.208	0.045	0.705	0.213
Oasis Empirical, All	372	0.629	0.924		0.946	0.879	
Oasis Empirical, ≥125 Words	271	0.679	0.921	0.893	0.935	0.871	

CC = Common Core; GMG = Gates-MacGinitie; SAT = Stanford Achievement Test;
n =number of texts in the sample; RM = Reading Maturity; SR = SourceRater

APPENDIX B
Full Results Table: Pearson's *r*

	n	REAP	ATOS	DRP	Lexile	RM	SR
CC Exemplar, All	168	0.537	0.571	0.515	0.504	0.700	0.744
CC Exemplar, Informational	103	0.630	0.631	0.527	0.606	0.755	0.797
CC Exemplar, Narrative	65	0.298	0.495	0.494	0.290	0.593	0.619
State Tests, All	683	0.463	0.651	0.585	0.589	0.783	
State Tests, ETS Subset	285	0.469	0.567	0.518	0.573	0.790	0.777
State Tests, Grades 3–5	254	0.267	0.394	0.371	0.302	0.355	
State Tests, Grades 6–8	285	0.198	0.354	0.284	0.262	0.353	
State Tests, Grades 9–11	144	0.158	0.249	0.164	0.224	0.215	
State Tests, Informational	401	0.459	0.707	0.662	0.621	0.763	0.791
State Tests, Narrative	275	0.459	0.611	0.511	0.554	0.802	0.764
GMG Grade Level, All	97	0.441	0.774	0.751	0.698	0.789	0.819
GMG Rasch, All	97	0.372	0.776	0.769	0.748	0.766	0.823
GMG Rasch, Grades 1–5	53	0.175	0.721	0.729	0.736	0.647	0.646
GMG Rasch, Grades 6-adult	44	0.044	0.411	0.331	0.198	0.415	0.512
SAT-9 Rasch, All	98		0.791	0.775	0.723	0.774	0.804
SAT-9 Rasch, Grades 1–5	41		0.727	0.678	0.637	0.610	0.552
SAT-9 Rasch, Grades 6-11	57		0.546	0.543	0.420	0.585	0.607
SAT-9 Grade, All	98		0.701	0.696	0.606	0.765	0.796
SAT-9 Grade, Grades 3–5	34		0.488	0.501	0.412	0.381	0.443
SAT-9 Grade, Grades 6–8	38		0.316	0.371	0.245	0.420	0.318
SAT-9 Grade, Grades 9–11	19		0.186	0.216	0.078	0.655	0.259
Oasis Empirical, All	372	0.621	0.918		0.949	0.875	
Oasis Empirical, ≥125 Words	271	0.678	0.940	0.922	0.961	0.895	

APPENDIX C

Common Scale for Band Level Text Difficulty Ranges

Common Scale for Band	Text Analyzer Tools					
	ATOS	DRP	FK	Lexile	SR	RM
2nd–3rd	2.75–5.14	42–54	1.98–5.34	420–820	0.05–2.48	3.53–6.13
4th–5th	4.97–7.03	52–60	4.51–7.73	740–1010	0.84–5.75	5.42–7.92
6th–8th	7.00–9.98	57–67	6.51–10.34	925–1185	4.11–10.66	7.04–9.57
9th–10th	9.67–12.01	62–72	8.32–12.12	1050–1335	9.02–13.93	8.41–10.81
11th–CCR	11.20–14.10	67–74	10.34–14.20	1185–1385	12.30–14.50	9.57–12.00

Key:

ATOS ATOS® (Renaissance Learning)
DRP Degrees of Reading Power® (Questar Assessment, Inc.)
FK Flesch Kincaid® (public domain, no mass analyzer tool available)
Lexile Lexile Framework® (MetaMetrics)
SR Source Rater© (Educational Testing Service)
RM Pearson Reading Maturity Metric© (Pearson Education)

Measures not in concordance table:
REAP (Carnegie Mellon University)
Coh-Metrix (University of Memphis)

APPENDIX D

Common measures for sample CCSS Exemplars, Career, Citizenship and College Texts

(Band) Title of Text	ATOS	DRP	Lexile	REAP	RM	SR
Sample Titles from Appendix B						
(2-3) Bat Loves the Night	5.0	53	760	4.9	5.6	1.0
(2-3) Cricket in Times Square (read aloud)	4.0	47	530	6.8	6.2	—
(4-5) A History of US: The First Americans, Prehistory to 1600	7.3	57	760	5.1	6.8	6.6
(4-5) Horses	5.6	56	910	3.6	5.7	2.2
(6-8) Cathedral: The Story of Its Construction	10.7	65	1120	11.4	9.1	5.8
(6-8) A Short Walk Through the Pyramids and Through the World of Art	9.1	61	1150	8.1	9.1	9.1
(6-8) The Dark is Rising	6.5	57	980	8.0	8.1	8.5
(6-8) The Tell-Tale Heart	6.7	56	640	11.2	9.4	10.3
(9-10) Gettysburg Address	8.7	62	1220	10.7	10.1	10.9
(9-10) I Have a Dream Speech 1963	9.1	61	1190	5.8	10.2	10.0
(9-10) The Gift of the Magi	6.5	55	880	7.4	9.4	8.5
(9-10) The Odyssey	8.5	60	1210	4.9	9.5	8.4
(11–12) Jane Eyre	9.2	64	1060	7.6	10.7	8.4
(11–12) The Declaration of Independence	15.1	71	1450	9.9	10.8	15.4
(11–12) The Great Gatsby	9.0	66	1490	8.5	10.3	13.9
College and Career Ready: Sample Career Documents						
Florida Real Estate Commission Newsletter	11.7	73	1270	12.0	—	11.5
Integrated Pest Managements for Home Apple Growers	10.5	67	1270	6.9	—	8.8
Learn About the United States: Quick Civics Lessons for the Naturalization Test	9.7	64	990	7.9	—	10.4
College and Career Ready: Sample First Year College Texts						
Media & Culture	13.9	74	1369	10.0	10.9	12.9
Microeconomics	12.7	68	1284	11.1	10.2	11.3
Understanding the Bible	14.9	74	1501	10.7	12.3	14.8

APPENDIX E
How to access the measures

1. **ATOS analyzer: Renaissance Learning**
   www.renlearn.com/ar/overview/atos/

2. **Coh-Metrix Easability Tool. University of Memphis**
   Beta site available at: http://141.225.42.101/cohmetrixgates/Home.aspx?Login=1

3. **Degrees of Reading Power: DRP Analyzer—Questar Assessment, Inc.**
   www.questarai.com (Contact info@questarai.com or 1-845-277-1600
   with requests for DRP Text Analysis Services).

4. **Lexiles—Metrametrics**
   www.lexile.com/analyzer/

5. **Pearson Reading Maturity—Pearson Knowledge Technologies**
   Beta site available at: www.readingmaturity.com

6. **REAP—Carnegie Mellon University**
   www.reap.cs.cmu.edu/

7. **Source Rater Educators Testing Service**
   Beta site available at: http://naeptba.ets.org/SourceRater3/

APPENDIX F

Trajectories of all Measures to College and Career Readiness