Improving Text Complexity Measurement through the Reading Maturity Metric

A paper presented at the annual meeting of the National Council on Measurement in Education
Vancouver, BC

Tom Landauer
Denny Way

April 2012

Abstract

One of the key requirements of the Common Core State Standards for Reading is that all students must be able to comprehend texts of steadily increasing complexity as they progress through school. As a result, proven methodologies that can be used to objectively quantify the complexity of reading texts are needed. This paper describes Pearson's text complexity measure, the *Reading Maturity Metric* (RMM), which utilizes a recently developed metric for tracing the growth of the meanings of words and passages, referred to as *Word Maturity* (WM). The purposes of this paper are to describe how WM has been incorporated into Pearson's text complexity measure, to present initial comparisons between this new measure of text complexity and traditional readability measures, and to address measurement issues in the development and use of text complexity measurements.

Keywords: text complexity, common core, latent semantic analysis

Improving Text Complexity Measurement through the Word Maturity Metric

Introduction

The common core state standards (CCSS) stress the urgency of confronting the long-widening gap between the complexity of texts that must be read in college and career and the diminishing sophistication of texts students have been reading in grades K-12 over the last several decades (CCSSO & NGA, 2010). Correspondingly the standards also cite research showing a decrease in students' ability to read independently over many past years. To address these challenges, the CCSS include requirements related to text complexity in their reading content standards as shown in the introduction of Appendix A of the standards:

> One of the key requirements of the Common Core State Standards for Reading is that all students must be able to comprehend texts of steadily increasing complexity as they progress through school. By the time they complete the core, students must be able to read and comprehend independently and proficiently the kinds of complex texts commonly found in college and careers (CCSSO & NGA, 2010, Appendix A, p. 2).

Attempts to correlate the difficulty of readings and statistical features of the text, such as words per sentence, go back to the 19th century (an overview of this history can be found in DuBay, 2004). In the second half of the 20th century, a series of readability indices were developed, such as Flesch–Kincaid (Kincaid et al., 1975), Coleman-Liau (Coleman et al., 1975), and Dale-Chall (Chall et al., 1995). These readability measures have relied primarily on two main variables--word frequency to represent vocabulary difficulty and sentence length as a surrogate for complexity arising from the organization of the text. Driven by critiques of these traditional

measures some new approaches appeared in the late 1960s of which, the Lexile® Framework (Stenner, 1996) is the best known. Though the Lexile scale is based on just the same two measured variables, average sentence length and word frequency (Stenner et al., 1988), it uses IRT to simultaneously assign word frequency and sentence difficulty. This greatly facilitated the matching student reading abilities to texts they can understand.

More recently, readability measures have begun to take advantage of more powerful new computational models and methods. One variety based on modern computational linguistics, counts and analyzes a much wider range of potentially important linguistic components. Another new approach simulates the ways that words and passages acquire and use word and passage meaning. Pearson's Reading Maturity Metric (RMM) combines the two.

The principle innovation of the RMM is the use of a recently developed metric for tracing the growth of the meanings of words and passages, referred to as *Word Maturity* (WM). WM is a computational model of the development of the meanings of individual words and paragraphs with increased expose to text (Landauer, Kireyev & Panaccione, 2011). It measures how knowledge of word and paragraph meanings evolve toward that of literate adults despite the spelled "word forms" remaining unchanged.

The purposes of this paper are to describe how WM has been incorporated into the RMM, to present initial comparisons between this new measure of text complexity and traditional readability measures, and to address measurement issues in the development and use of text complexity measurements.

Word Maturity and Text Complexity

WM is based on an application of Latent Semantic Analysis (LSA) (Deerwester et al., 1990; Landauer & Dumais, 1997)--which is also the basis for Pearson's technology for automated scoring of essays and other constructed-response item types. LSA is typically applied to a large language corpus and represents the words and passages (and/or paragraphs) used in the corpus as numerical vectors standing for points in a very high (e.g. 50-1,500) reduced dimensional "semantic space". LSA uses a matrix algebra technique called singular value decomposition (SVD), a method for decomposing a rectangular matrix into the product of three other matrices: 1) a matrix that describes the original row entities as vectors of derived orthonormal vector values, 2) a matrix that describes the original column entities in the same way; and 3) a diagonal matrix containing scaling values such that when the three components are matrix-multiplied, the original matrix is reconstructed. An important advantage of SVD as applied to a language corpus is that it results in vectors representing passages as the sum of the vectors representing each of the words within them, and the vector for each word is an average of the vectors for all of the passages in the corpus that include it.

Landauer, Kireyev, and Panaccione (2011) introduced the WM metric for the development of reading comprehension-oriented word knowledge. Based on large and representative LSA text corpus, WM is obtained by cumulatively adding specific educational or naturally ordered samples of text paragraphs in quantities typical of student reading. The complete set of paragraphs is meant to be representative of the text encountered by literate adult readers. Figure 1 illustrates how the simulation traces the separate growth toward adult levels for every word of interest: for example all those used in a particular text. The order of the cumulative sets of paragraphs have usually been selected from books and readings whose overall difficulty has

previously been estimated in some way, here by Lexile ratings, However reasonably good results can be obtained even with random orders. An overview of the process for producing a WM scale, borrowed from descriptions in Landauer, Kireyev and Panaccione (2011) and Kireyev and Landauer (2011) is as follows:

1. Obtain a large and representative text corpus divided into paragraphs of text.

2. Use Singular Value Decomposition (SVD) to create an LSA "semantic space" for the entire corpus that contains at least all the words that will be encountered in the text, and as many more general and academic words and occurrences thereof as feasible.[1]

3. Create a series of subcorpora by adding portions of the total corpus in an order that approximates the order of text encounters by typical learners of the language, for example, here by using text at successive Lexile levels. Lexile measures use combinations of word frequencies and sentence lengths within paragraphs (note that these are causes of word learning, not results) to align paragraphs consistently with samples of human performance using cloze tests and IRT. (However, as in the above remark this ordering is not strictly necessary and may be somewhat unrealistic, and ordering in different ways produces only modest differences in results as illustrated later.

4. At each step, compute a new LSA semantic space for the cumulatively enlarged sample. This is the current space after simulated reading of all *n* paragraphs so far encountered.

[1] In LSA, as in nature, every word in a language can influence the meanings of any or all others to some extent, so the larger and more representative of the total relevant language the better.

5. Align the vectors of all the paragraphs now in the new space as well as possible with those of the same paragraphs in the larger full "adult" space. This employs the linear algebra method called Procrustes rotation, which is needed because SVD is unique only up to rotation and adding new paragraphs of text requires simultaneous least squares alignment of the new and old data.

6. Measure the cosine distance of each word vector in the current cumulative space to the vector it has in the full "adult" space.

7. If there are *m* cumulative spaces, the WM curve displays the cosine to the adult meaning at each step, showing the evolving status of a word (or paragraph) by how close its vectors are to that of a simulated adult reader in units of equal amount of simulated total reading [2]. Individual word development can then be traced as a function of the amount of simulated reading. Interpolation between scale points provides a continuous measure.

Figure 1, which previously appeared in Landauer, Kireyev & Pannacione (2011) and Landauer (2011), illustrates the WM trajectories for five different words: dog, electoral, primate, productivity, and turkey. The x-axis in this plot represents the cumulative number of paragraphs added to the corpus of text at each step and the y-axis represents the WM expressed as the cosine distance between the vector of the word in the semantic space of each cumulative sample and the vector of the word in the "adult space". The starting paragraph sub-corpus of 10,000 paragraphs

[2] This means total of all reading, not of just reading the word or passage being measured. The reason is that learning the meaning of new and maturing words depends on what other words have previously been encountered and in what contexts as well as on occurrence of the word itself. This is an important theoretical and empirical principle whose veracity is established by the success of WM itself as well as its close agreement with classical IRT based measures of vocabulary growth over a variety of sampling situations and methods.

is based on an estimate of exposure to ambient and written language before school entry, and the full corpus of 85,000 paragraphs in the example represents a calibration point at about the amount of text that has been read by an average college student (Landauer, Kireyev, & Pannacione, 2011). As would be expected, each word follows a different trajectory as the corpus of text increases. The word "dog" is nearly as mature in a corpus of 10,000 paragraphs as it is in the adult space. In comparison, the word "productivity" is relatively unknown until 50,000 paragraphs have been encountered. The other words follow different trajectories that for the most part (but not necessarily, because inserting new or conflicting meanings can temporarily decrease similarity to overall adult status) increase monotonically with increasing numbers of paragraphs.

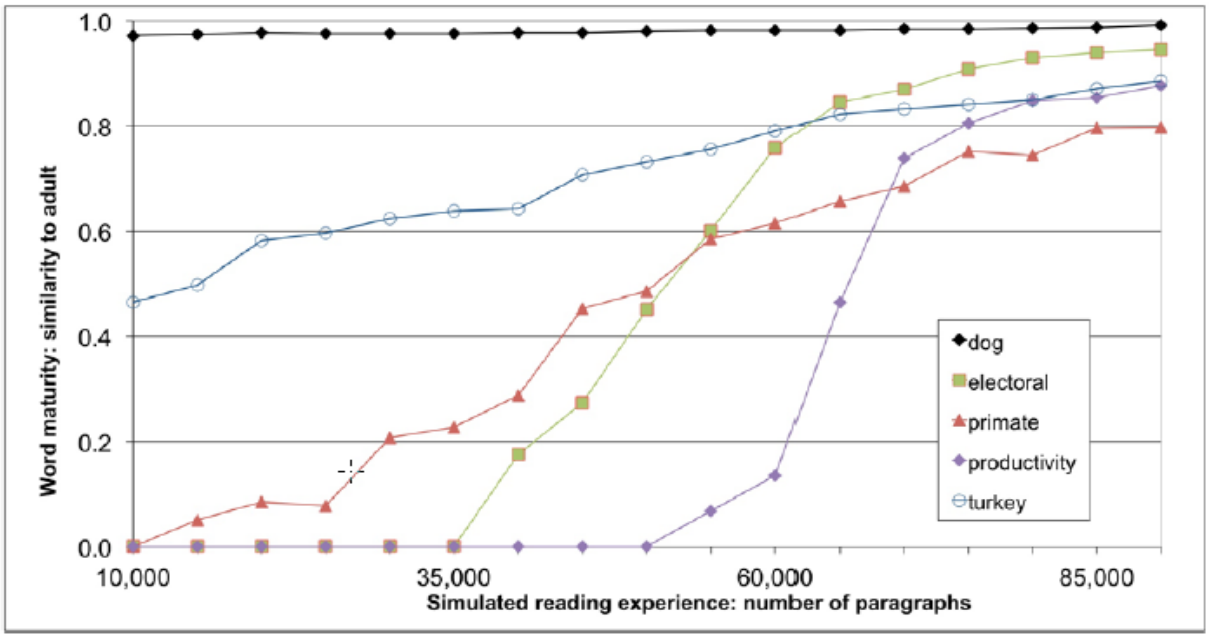

Figure 1. Examples of Word Maturity (WM) Trajectories for Five Words

Role of Word Maturity in Text Complexity

As illustrated in Figure 1, the trajectories of each word in the corpus of paragraphs can be represented as a vector of WM values and stored in an indexed database for subsequent use. To

incorporate WM into a text complexity measure, it is necessary to summarize the vectors in some fashion. The approach adopted in recent analyses is to determine the size of the corpus (in terms of cumulative paragraphs) that corresponded to a WM similarity value of 0.65. For example, in Figure 1 the word “turkey” reaches the word similarity value of 0.65 its threshold value (on the y-axis) when about 40,000 paragraphs have been encountered. In comparison, the word “productivity” does not reach its threshold value until about 68,000 paragraphs have been read. Once obtained, these values are re-scaled for each word in the corpus to a metric that ranges from zero to one. The rescaling is such that words that become mature quickly (such as “dog”) receive values close to zero and words that become mature relatively late (e.g., “productivity” or “primate”) receive values close to one. We refer to these values as “time to maturity” (TTM) and note that the higher the value the larger the exposed subcorpus of text needs to be before the word stabilizes to its meaning in the adult LSA semantic space.

Figure 2 illustrates how the TTM values can differ across different passages taken from Appendix B of the Common Core State Standards (CCSS) for English language arts and literacy (CCSSO & NGA, 2010). The distribution of TTM values is shown for the passages “The Lighthouse Family: The Storm” by Cynthia Rylant (left panel, p. 41 from Appendix B) and “Address to Students at Moscow State University” by Ronald Reagan (right panel, p. 128 from Appendix B). These passages were classified at grade 2.5 and grade 7, respectively. As would be expected from the grade level classifications, the distribution of TTMs for Reagan’s address are less skewed and have a higher mean TTM (0.47) as compared with the less complex “Lighthouse” passage (mean TTM = 0.28).

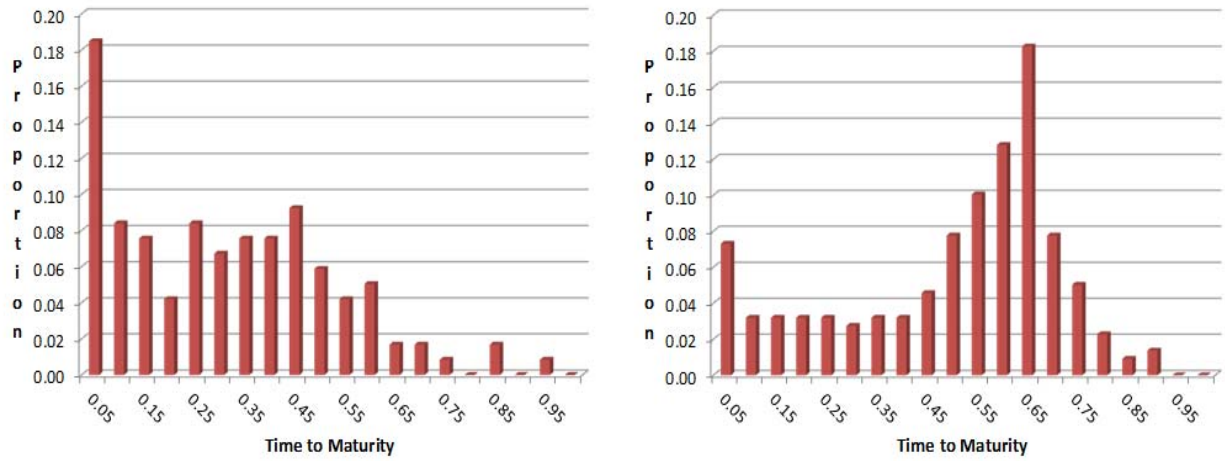

Figure 2. Example of distribution of TTM Values for Two Common Core Reading Passages

The Pearson Reading Maturity Metric Components

The scoring process for the RMM divides into two components: A component devoted entirely to the WM measurement, and a component devoted to linguistic description and analyses, plus results of their combination.

1. **The WM Component**. To obtain a text complexity rating, the TTMs of each of the words in a text are simply averaged, with the exception that the *n* words with the highest TTMs are given extra weight to account for the possibility that a few extremely rare words that don't follow the general distribution might skew the results. A very important aspect of the WM component for this application is that the units of measurement are simply and always word-forms. Misspelled words having no complexity measure are ignored in the averaging.

2. **The Linguistic Component**. Additional passage-level variables reflecting both standard linguistic features (e.g., average sentence length, average word length) and newer computational variables (e.g., n-gram probabilities, semantic coherence) were also included in the text complexity model (Landauer, 2011). The relative importance of the features was calibrated on reading passages from two sources: Pearson's Summary Street product and a

collection of publicly available state readings (described below). The calibrations used the assigned grade levels of the passages as the dependent variable (in some cases these were midpoints of grade ranges and so were non-integer). Jackknife-based cross-validation procedures were used in evaluating the model.

In some analyses, other features were found to improve correlations with one or more complexity outcome measures. In general, although WM accounts for a substantial portion of the total variation in and accuracy of the RMM, the linguistic component does serve to increase the predictive power of the measure. In addition, the use of WM in the RMM offers some additional practical advantages. For one, it tends to offset the need for human editing and correcting of errors in writing, spelling and syntax prior to model application.

Results of Completed Data Analyses

Initial applications of the RMM were reported in Landauer (2011). In that study, the calibrated text complexity model described above was applied to passages from various reading and English language arts assessments. The following data sets were analyzed:

- **Summary Street**: Randomly chosen whole texts from a large set of readings used in Pearson K-12 tutorial reading passages, the vast majority of which are informative text, all used in Pearson's WriteToLearn® reading comprehension and iterative summarization formative tutorial exercises (N = 800).
- **State Readings**: A collection of publicly-available diverse texts used as items in No Child Left Behind (NCLB) reading comprehension tests in 27 state and two national assessments (N = 1221).

- **SAT9**: Texts used in the well established reading comprehension subtest of the Stanford Achievement Test, Version 9 (N = 110).
- **Common Core**: Example texts (N = 243) appearing in Appendix B of the Common Core State Standards for English language arts and literacy (CCSSO & NGA, 2010). These texts were selected by the authors of the standards to exemplify the level of complexity and quality that the Standards require all students to engage with across a range of grades.
- **NAEP**: Passages from the National Assessment of Educational Progress administered at grades 4, 8, and 12 (N=40).

For each of these data sets, the WMM was used to assign grade levels to the texts analyzed. These assignments were compared to the actual grade levels of the passages. In addition, we assigned grade levels to each passage based on two well-known readability formulas, Flesch–Kincaid (Kincaid et al., 1975) and Coleman-Liau (Coleman et al., 1975). Results are summarized in Table 1, which includes summary statistics for the grade level assignments and for the predicted grade levels based on the text complexity model and the two readability formulas. The table also includes the correlations between model-based grade level predictions and the grade levels assigned. It can be seen in Table 1 that the RMM grade levels were notably more highly correlated with the actual passage grade levels than the grade levels predicted by the readability formulas. It can be also seen that the mean RMM grade levels were within one-half of a grade level of the actual grade level mean for all of the data sets except the Common Core readings, where the RMM mean was 1.25 grade levels higher than the actual mean grade level. Compared with the RMM, the readability formulas generally (but not always) produced mean grade levels that were further from the actual mean grade levels. With one exception (Flesh-Kincaid grade

levels for the State Readings), the standard deviations of the grade levels predicted by both the RMM and readability formulas were lower than the actual grade level standard deviations.

Table 1. Summary Statistics for Assigned Grade Level of Passages and Grade Levels Based on the Reading Maturity Metric or Readability Measures

		Grade Level of Passages						Reading Maturity Metric				
Passage Set	**N**	**Mean**	**Std**	**Min**	**Max**	**Rxy**		**Mean**	**Std**	**Min**	**Max**	**Rxy**
Summary Street	800	6.86	2.49	3.00	10.50	N/A		7.14	2.05	3.00	11.86	0.881
State Readings	1,196	8.29	2.86	3.00	12.00	N/A		8.12	2.40	2.62	12.24	0.862
Stanford-9	111	6.01	2.86	1.00	11.00	N/A		5.62	2.04	-0.80	10.22	0.844
Common Core	243	6.48	3.59	0.50	11.50	N/A		7.73	2.27	-0.82	11.40	0.736
NAEP Grade 12	40	7.59	3.15	4.00	12.00	N/A		7.85	1.99	3.88	11.38	0.783
		Flesh-Kincaid						Coleman-Liau				
Passage Set	**N**	**Mean**	**Std**	**Min**	**Max**	**Rxy**		**Mean**	**Std**	**Min**	**Max**	**Rxy**
Summary Street	800	9.48	2.20	5.57	17.00	0.637		9.87	1.84	4.06	15.33	0.636
State Readings	1,196	9.06	3.10	4.00	17.00	0.591		8.75	2.71	1.95	18.51	0.598
Stanford-9	111	6.93	2.01	4.00	13.73	0.673		6.48	2.82	-0.26	12.26	0.732
Common Core	243	8.06	2.71	4.00	17.00	0.485		7.23	3.09	-3.82	13.78	0.508
NAEP Grade 12	40	7.57	2.09	4.00	13.10	0.286		7.43	2.32	2.43	11.92	0.217

Note: Rxy is between assigned passage grade level and RMM or readability measures. For common core passages, assigned grade levels were the midpoint of assigned by expert human raters (CCSSO & NGA, 2010, Appendix B).

Challenges and Opportunities in the Use of Text Complexity

Pearson's RMM appears promising for instruction and assessment applications related to the CCSS. The CCSS in English language arts stress the ability to read, understand, and analyze increasingly complex texts. Indeed, one of the goals of the CCSS is to "translate the broad (and, for the earliest grades, seemingly distant) aims of the [College and Career Readiness] standards into age- and attainment-appropriate terms. The Standards set requirements not only for English language arts (ELA) but also for literacy in history/social studies, science, and technical subjects" (CCSSO & NGA, 2010, p.3).

The primary challenge at this time is that the text complexity calibration is dependent upon the external grade level assignments for each passage that is part of the calibration samples. As previously mentioned, the current version of the text complexity model was calibrated using

samples of reading passages, texts from readings from Pearson's Summary Street product that are recommended for use at particular grade levels, and publicly-available passages used in state-level reading comprehension tests at particular grade levels. From the perspective of the CCSS, grade level assignments for existing texts may not reflect the rigor that is needed to prepare students for college and careers. This is essentially the conclusion drawn in Appendix A of the CCSS, which points out a "general, steady decline—over time, across grades, and substantiated by several sources—in the difficulty and likely also the sophistication of content of the texts students have been asked to read in school since 1962" (CCSSO & NGA, 2010, Appendix A, p.3).

The issues associated with calibrating the RMM given passages with human-assigned grade levels is akin to those faced with Pearson's Intelligent Essay Assessor (IEA; c.f., Landauer, Latham & Foltz, 2003). IEA computes the overall content similarity in a LSA space between a new essay and essays on the same topic that have been graded by humans, and subsequently determines the nearness of the new essays to human graded essays. Based on this proximity in semantic space to the human graded essays, IEA predicts what grade a human would have given to the new essay. In a similar fashion, the RMM uses the summary WM values and other variables for passages that have been assigned grade levels by humans to subsequently predict what grade level would be assigned to a new text passage. As with calibrating essays, the sample size (e.g., number of passages), the distribution of grade levels, and the reliability of the human grade level assignments will affect text complexity model calibration. We will continue to explore the effects of different text calibration samples as we refine the RMM.

Another issue in calibrating the RMM has to do with differences between literary and informational genres. Following the lead established in the NAEP Reading Framework (National

Assessment Governing Board, 2010), the CCSS stress a need for balancing between the broad categories of literary and informational tests:

> The K–5 standards include expectations for reading, writing, speaking, listening, and language applicable to a range of subjects, including but not limited to ELA. The grades 6–12 standards are divided into two sections, one for ELA and the other for history/social studies, science, and technical subjects. This division reflects the unique, time-honored place of ELA teachers in developing students' literacy skills while at the same time recognizing that teachers in other areas must have a role in this development as well.
>
> Part of the motivation behind the interdisciplinary approach to literacy promulgated by the Standards is extensive research establishing the need for college and career ready students to be proficient in reading complex informational text independently in a variety of content areas. Most of the required reading in college and workforce training programs is informational in structure and challenging in content; postsecondary education programs typically provide students with both a higher volume of such reading than is generally required in K–12 schools and comparatively little scaffolding (CCSSO & NGA, 2010, p.4).

Sheehan, Kostin, Futagi, and Flor (2010) suggested that that many important linguistic features function differently within informational and literary texts. As a result, their text complexity research utilized two distinct prediction models: one optimized for application to literary texts and one optimized for application to informational texts. However, the RMM in most cases gave very similar result for informational and literary texts. Although the reasons for the difference have not been sufficiently analyzed as yet, a reasonable conjecture is that WM's exclusive

consideration of the growth of word meaning, rather than what one might call the stile of its presentation, focuses it on the text's information content in the information-theoretic sense, We might, then, consider which roles of text are needed for readiness what foci of higher education, rather than only its "reading difficulty."

There are a number of other applications of text complexity and word maturity that can be anticipated. One is in selecting texts in an instructional setting, where assignment can be personalized to the individual's reading level. Another is in teaching and measuring vocabulary, through which words can be selected and presented to individuals based on their WM trajectories. Still another application of text complexity would be to look at the flow and order in which information is conveyed in texts, and how that affects comprehension. This strand of research would seem particularly important with math and science texts.

Summary

The purpose of this paper was to summarize and expand the work reported in Landauer (2011) on Pearson's RMM and in Landauer, Kireyev, and Panaccione's (2011) foundational research on WM. We elaborated on the role played by WM in the RMM, provided some examples of how WM is distributed in text passages, and presented additional details of analyses originally reported in Landauer (2011) related to the validity of the approach.

As outlined in the CCSS for ELA and literacy, quantitative evaluations of text are an important component of measuring text complexity. By considering WM and other structural features of text that go beyond traditional readability measures, the WMM incorporates a broader modeling of word knowledge that, based on our initial research, provides a superior prediction of criterion grade level classifications for a wide variety of K-12 readings. While the methodology of WM

and text complexity has grown out of computational linguistics and cognitive psychology, it is hoped that this paper will help to make the approach more understandable to those working in educational assessment.

References

Coleman, M., & Liau, T. L. (1975). A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, 60, 283-284.

Council of Chief State School Officers [CCSSO] & National Governors Association [NGA] Center for Best Practices. (2010). *Common core state standards for English language arts & literacy in history/social studies, science, and technical subjects*. Available at: http://www.corestandards.org.

Deerwester, S., Dumais, S., Furnas, G.W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6), 391-407.

DuBay, W.H. (2004). *The Principles of Readability*. Available at: http://www.impact-information.com/impactinfo/readability02.pdf.

Graesser, A.C., McNamara, D.S., Louwerse, M.M., & Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, and Computers, 36*, 193-202.

Graesser, A.C., McNamara, D.S., & Kulikowich, J.M. (2011). Coh-Metrix: Providing multilevel analyses of text characteristics. *Educational Researcher*, 40(5), 223-234.

Kincaid, J. P., Fishburne, R. P., Rogers, R. L., & Chissom, B. S. (1975). *Derivation of New Readability Formulas (Automated Readability Index, Fog Count, and Flesch Reading Ease formula) for Navy Enlisted Personnel*. Research Branch Report, 8-75. Chief of Naval Technical Training: Naval Air Station Memphis.

Landauer, T.K. (2011). *Pearson's text complexity measure*. Iowa City, IA: Pearson White Paper. Available at: http://www.pearsonassessments.com/textcomplexity.

Landauer, T. K., Kireyev, K., & Panaccione, C. (2011). Word Maturity: A New Metric for Word Knowledge. *Scientific Studies of Reading*, 15(1), 92-108.

Landauer, T.K., Laham, D., & Foltz, P.W. (2003). Automatic essay assessment. *Assessment in Education: Principles, Policy & Practice*, 10(3), 295-308.

Landauer, T. K., & Dumais, S. (1997). A solution to Plato's problem: The Latent Semantic Analysis theory of the acquisition, induction, and representation of knowledge. *Psychological Review*, 104, 211-240.

National Assessment Governing Board (2010, September). *Reading framework for the 2011 National Assessment of Educational Progress*. U.S. Department of Education. Available at: http://www.nagb.org/publications/frameworks/reading-2011-framework.pdf.

Sheehan, K. M., Kostin, I. Futagi, Y., & Flor, M. (2010). *Generating automated text complexity classifications that are aligned with targeted text complexity standards* (ETS Research Report RR-10-28). Princeton, NJ: ETS.

Stenner, A.J., Horabin, D., Smith, R & Smith, R. (1988). *The Lexile Framework*, Durham, NC: Metametrics, Inc.

Stenner, A. J. (1996). Measuring Reading Comprehension with the Lexile Framework. *Fourth North American Conference on Adolescent / Adult Literacy*. Washington D.C.