Longitudinal Analysis of First Grade Texts Using the Reading Maturity Metric

Peter W. Foltz and Mark Rosenstein
Pearson

Abstract

The Pearson Reading Maturity Metric (RMM, readingmaturity.com) represents a new generation of automated text complexity measurement which incorporates modern approaches to natural language processing and artificial intelligence to assess broader and deeper levels of information in reading passages. The RMM uses computational linguistic variables to measure coherence, flow, word difficulty, the complexities of the grammar construction, sentence length, and semantic coherence in order to determine the overall difficulty and complexity of the reading passage. It also incorporates a computational language model to estimate how much language experience a student requires to be able to process the meaning of each word, sentence and paragraph in a text. The RMM provides predicted text complexity scores aligned to student grade levels as well to Common Core Grade bands as well as providing information detailing features that led to the level of text difficulty. The measure has been validated across a range of standardized data sets and shown to provide 20-30% more accurate measures of text level than more traditional measures.

In this study, first grade passages were analyzed to derive overall RMM scores for each passage as well as measures of individual features that contribute the RMM scores. These features included measures of length, coherence, rate of introduction of new words, diversity of vocabulary, and usualness of the flow of words. The study analyzes the kinds of features that have changed in the complexity of the passages over the past 50 years. The results indicate that generally, passage complexity, as measured by RMM, has increased from 1962 to 2013. Analysis of individual features shows that passages from earlier decades tend to be more coherent, but less like standard written English. Passages from later decades introduce new words at much higher rate and use words of greater difficulty.

Introduction

While automated measurement of text complexity has been widely used over the past 50 years, major changes in the technological approach have only really occurred recently. Traditional, and many currently used readability formulas rely on a few, simple superficial measures, such as the average number of words in sentences and how frequently words are encountered in general or in educational reading. This traditional approach is often sufficient to predict the relative reading difficult of entire books and articles. However, it is highly influenced by differences between narrative and non-fiction texts with large variability due to small changes in words and lengths of sentences. The traditional approach further misses almost half the complexity that occurs within and between sentences and paragraphs. This combination of insensitivity and variability can result in providing educations insufficient information to understand the sources of complexity in the text or how to remedy it and make it more comprehensible for the students.

The Reading Maturity Metric (RMM, http://readingmaturity.com) represents a new generation of automated text complexity measurement. It incorporates modern approaches to natural language processing and artificial intelligence to assess broader and deeper levels of information in reading passages. The RMM uses computational linguistic variables to measure aspects such as coherence, flow, word difficulty, complexities of the grammar construction, sentence length, and semantic coherence in order to determine the overall difficulty and complexity of the reading passage. It also incorporates a computational language model to estimate the language experience required for a student to be able to process the meaning of each word, sentence, and paragraph in a text (e.g., Landauer et al., 2011).

RMM Features

The RMM incorporates a range of features that measure aspects of complexity including the maturity of the words used, sentence and paragraph statistical complexity, within and between sentence coherence, use of punctuation, and ordering of the information. A number of these measures are based on natural language processing methods ranging from simple counts of features to statistical n-gram models based on probabilities of sequences of words in the English language

RMM further incorporates semantic information in its measures to analyze aspects such as how quickly the text flows from one topic to the next and how well the text holds together as whole. One key differentiator from other approaches is RMM's use of word maturity which is based on the Latent Semantic Analysis computational language model (Landauer, Foltz & Laham, 1998) and simulates the reading experience that is required to achieve a "mature" understanding of the meaning of each word, sentence, and paragraph in a text. The approach simulates what happens, on average, to the knowledge of every individual word and paragraph in a large corpus as a function of how much and what an average student has read. Unlike the traditional methods that treat the meaning of a word as static, the RMM "measures how the meanings of words themselves, and thus of passages, change for learners with increasing exposure to language" (Landauer, 2011, p. 3). Changes in the meanings of words as they mature have significant effects on the complexity of a text. For example, the maturity (e.g., a student's ability to grasp the meaning) of the word "capabilities" evolves slowly, despite exposure to the word in middle school and high school because it is not used in ways to provide specific contexts. In contrast, the meaning of the word "dog" develops a mature meaning very early on because its context-specific use. Computing the average maturity and highest maturity words that are used in a text provide measures of the degree of complexity of the words used in the text.

Based on an analysis of the features, the features are combined using machine learning approaches resulting in RMM values and the alignment of the RMM values to student grade levels (K-16) as well as to Common Core Grade bands. The RMM further provides information about words in the passage that may be more difficult for student comprehension in order to help text developers and teachers in choosing appropriate passages and instructional material around the passages. RMM scores are typically reported with guidelines and samples of appropriate texts that are at the same student independent reading level. It is currently made available as a free service to educators to analyze texts at readingmaturity.com.

Prior validation of RMM

Previous research with the RMM indicates a strong relationship between the complexity of reading passages and student performance on the test items associated with a passage. For example, the study examined the relationship between passage text complexity and the average Rasch difficulty values for passages on the vertically scaled Stanford Achievement Test, Ninth Edition (Stanford 9), ranging from grade 1 to grade 11. Figure 1 shows a bivariate plot, which indicates a very strong relationship (r = 0.84).

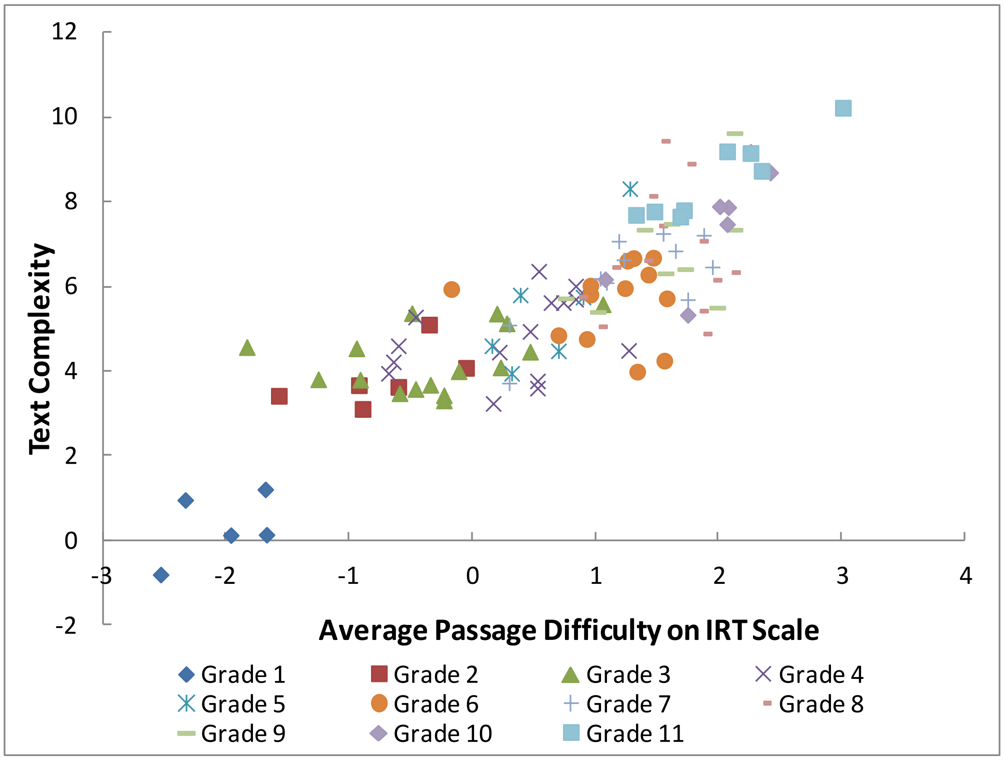

Figure 1. Relationship between Reading Maturity Metric scores And Stanford 9 Average Passage Difficulties

A recent study conducted by Nelson, Perfetti, Liben, and Liben (2011) evaluated several measures of text complexity and found strong performance by the RMM in comparison with other text complexity measures. For example, for 683 State Test texts, the rank order correlation of grade level estimates with grade bands established by expert instructors was 0.79 for RMM, compared to 0.48 to 0.66 for other traditional measures of text complexity.

Data

The data source consisted of the entire body of Scott Foresman first-grade core-reading texts, spanning seven editions, from the following years: 1962, 1971, 1983, 1993, 2000, 2007 and 2013. The content analyzed comprised 507 files. Individual year's files were divided into "levels" imposing a temporal ordering of presentation to students. Within each level the texts are sequenced, so it is possible to examine the entire progression of readings. Table 1 provides summary statistics for the content indicating the number of texts, number of levels, and some initial text metrics for each year. Since length measures tend to be right skewed (having a small set of larger readings), both mean and median are presented for average reading word count and sentence count. Given the preponderance of short words at this grade level, neither the mean nor median characters per word are indicative proxies for the hardest words.

Instead the 3rd quartile of characters per word is presented as a more telling measure of potential differences in word length.

year	n	levels	total words	3rd Qtr. char. per word	mean word count	median word count	mean sentence count	median sentence count
1962	114	5	19287	3.64	169.18	159.5	19.30	18.0
1971	42	4	8507	3.92	202.55	161.0	28.93	26.0
1983	107	5	30556	3.76	285.57	271.0	37.19	34.0
1993	64	6	9838	4.09	153.72	84.0	20.36	9.0
2000	66	5	13141	3.91	199.11	168.5	31.35	28.0
2007	60	5	11143	4.06	185.72	123.5	27.97	22.5
2013	54	5	9928	4.15	183.85	127.5	26.56	21.5

Table 1. Summary statistics of readings analyzed.

Table 1 provides a starting point for trends that will be discussed in substantially more detail in the analysis section, but a couple of points are worth noting. The number of readings per year varies by almost a factor of 3 with a minimum of 42 in 1971 to the most in 1962 of 114. In 1993, the mean length of the readings was double the median, both in terms of number of words and number of sentences indicating a wide disparity in readings' lengths and the median reading length in 1993 is substantially below the median reading length of other years. Finally, there is potentially a trend of increasing word length, at least at the 3rd quartile over the series of readings.

A few text features have prerequisites that some of the readings did not meet, for instance sentence-to-sentence coherence requires a minimum of two sentences. The validation methods for RMM require at least 20 words and 2 sentences, which meant that it was not possible to compute RMM values for 14 readings. In addition RMM has additional validations that indicate when a text is statistically very different from the set of texts RMM was trained on, indicating caution is required in interpreting the RMM value. Of the level 1 readings, 33 were flagged as unusual with respect to the training texts and had unacceptably high confidence intervals and so were eliminated from RMM analysis. There were also 22 texts that had RMM scores flagged with advisories (indicating some level of deviance from the training set, so less reliable) that were included in the RMM analysis.

A number of issues resulting from the transcription of the data are detailed in Appendix 1. Since we did not have the original printed resources, we only attempted mechanical cleanup (such as unifying the character encodings to ASCII) that were unlikely to add structure that may not have been in the original source materials. In addition 39 texts appeared in more than one year, which complicates the analysis especially between years 2007 and 2013 where more than half the documents overlap the two years. Table 2 indicates the extent of this issue.

n	years
1	1971 & 1993
1	1983 & 2007
3	1993 & 2000
1	1993 & 2007
1	1993 & 2013
1	2000 & 2013
31	2007 & 2013

Table 2. Content appearing in more that one year.

Analysis

Analysis approach

The goal of the analysis was to analyze the degree to which first-grade texts have changed in complexity over the past 50 years. In addition, it sought to determine dimensions or textual features have these changes occurred. The results could then provide some understanding of the underlying pedagogical principles that may have gone into the design and choice of the texts. The analysis examines both RMM scores as a whole, but also includes other features, some of which comprise components of the overall RMM score.

We observed that for many attributes of the text, the variability between levels in a single year can exceed the between year variability, so our analysis for the most part includes the levels within a year. We begin the analysis with length, in this case word count to introduce the methodology. We then present RMM to give an overall view of the complexity of the readings, and then examine some components of RMM to explore in more depth differences in levels and across years.

Overall Length

The overall length of a reading provides an indication of the total amount of reading expected of the students and sheds some light on the potential strategies in designing texts of different levels of difficulty for first graders. In order to illustrate the variability of the readings, Figure 1 shows the word counts of all the readings by year broken out by level. The y-axis is log base 10 of word count with the largest text (almost certainly an outlier as described in Appendix 1) is excluded. The curve in each panel indicates a locally weighted regression fit through the data.

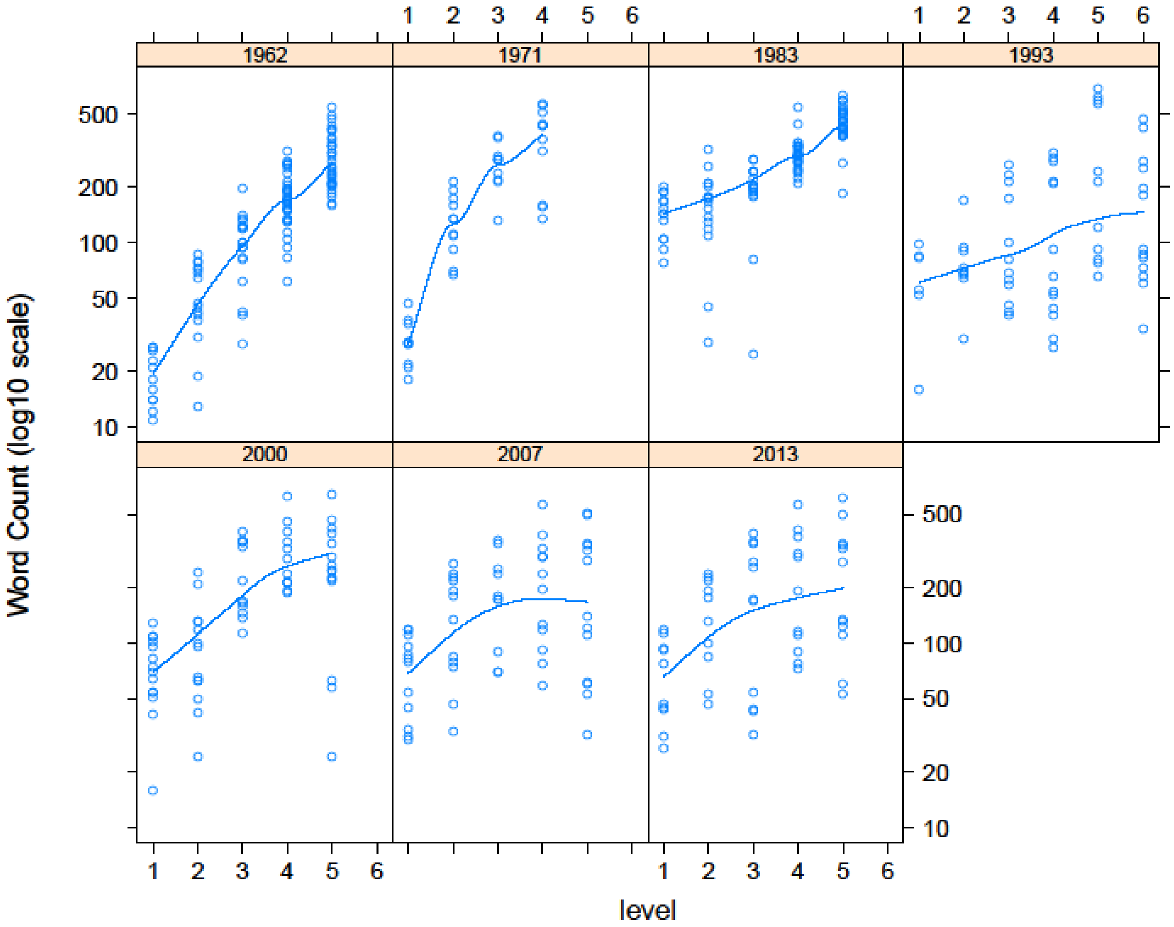

Figure 1. Word Count (log scale) by level per year.

Clearly the discipline over word count changed over the years. In 1962 there is a clear pattern of increasing word count by level with smaller spreads at lower levels. At the lowest level, text sizes were quite small, which carried over to 1971, but after 1971, the use of the very shortest readings (below about 20 words) essentially disappears. In 1983, the lengths at each level are tightly clustered with the length at the lowest level much higher than any other year, but an overall increase in length by level is still evident. For 1993, 2007 and 2013 there is great variability in length at each level, with modest increase by level often reaching an asymptote of about 200 words at the higher levels (though again note the extreme variability). Finally, 2000 shows a steady increase between levels with as tight variability as 1962.

To more easily compare length across years, the plot in Figure 2 abstracts each level with its median length and color codes each year's trajectory over levels. The y-axis is median word count, and the x-axis represents level. The slope information is slightly more difficult to compare by year due to years containing different number of levels, but the overall patterns are still reasonably clear.

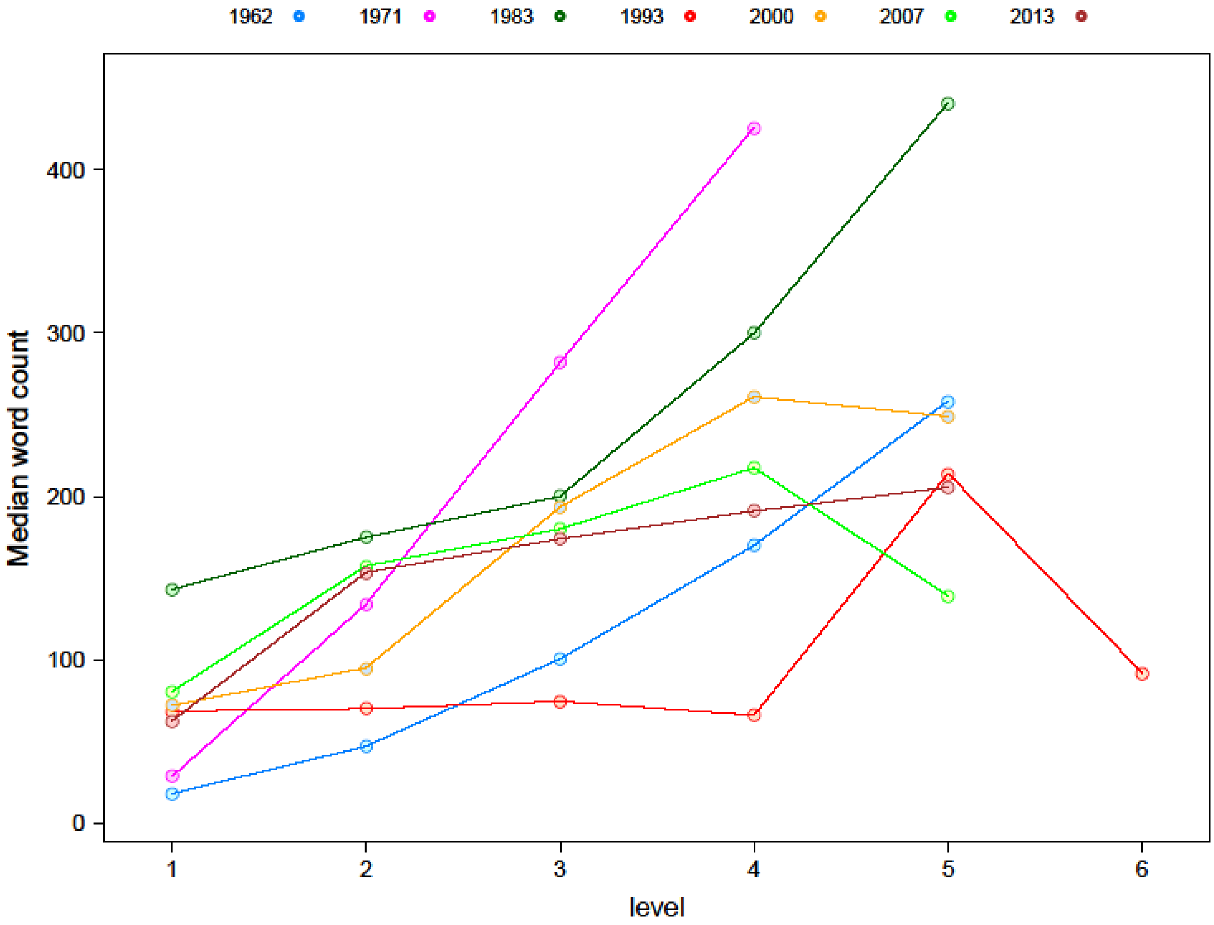

Figure 2. Reading length (median word count) by level, color coded by year.

In terms of target beginning and ending lengths, 1963 and 1971 begin with shorter readings, and 1983 with longer readings and the other years cluster at a median of about 80 words. The target ending reading lengths are more dispersed with 1971 and 1983 readings nearly twice as long as the median of the next set of years clustered around a median of 200 words, with 1993 and 2007 dropping in their last level. The change in control over length over levels is fairly easy to compare for 1963, 1971 and 1983. In all of these years, reading length increases with level, and likely at an increasing rate. 2013 increases, but at a very slow rate after leaving the first level. 1993 is interesting in that it is flat for the first 4 level, jumps up and then jumps back down. If the length of a reading is one proxy used in designing a reading suite, we see that different strategies have been applied over the years. There is mostly agreement that length should increase throughout the year, but starting and ending lengths as well as trajectories have varied. We next look at the paths of the Pearson Reading Maturity Metric for these readings.

RMM

Reading Maturity Metric (RMM) scores provide an overall characterization of the level of a text. The Figure 3 shows the median RMM and Q3 RMM by level for the readings. As mentioned in the Data section, for 47 readings we were unable to derive reliable RMM values. As in previous plots, the years are color coded, and in addition a grey rectangle has been added to indicate the Common Core K-1 grade level boundary with 2-3 grade level.

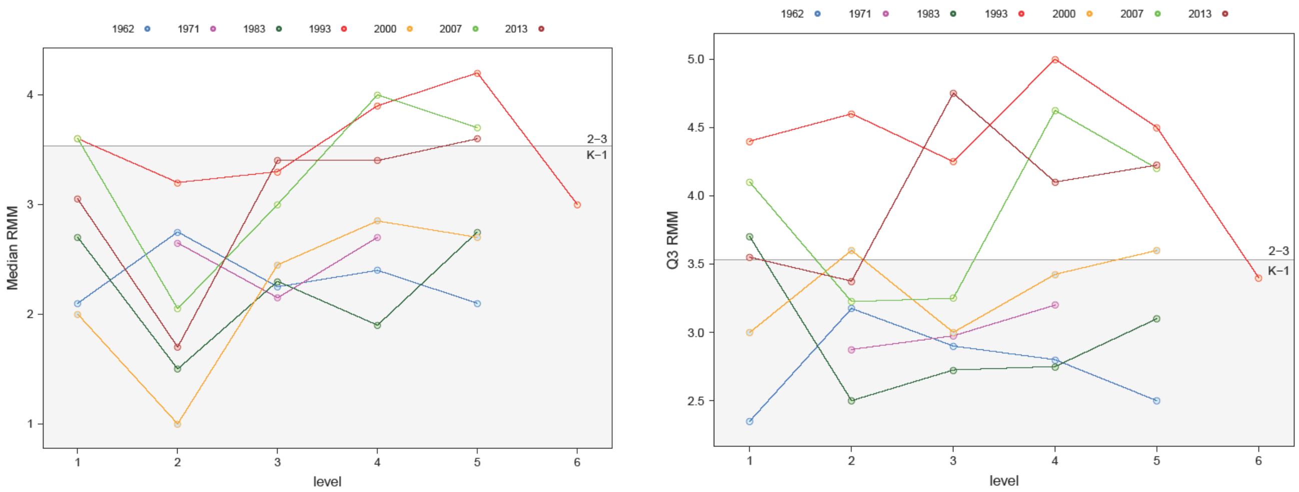

Figure 3. Reading Maturity Metric by level, color coded by year. Left panel: Median RMM; Lower panel: 3rd Quarter RMM.

In the left panel, for the years 1993, 2007 and 2013 for some of the later levels, median readings enter into the CCSS 2-3 grade band. Examining the 3rd Quarter RMM in the right panel, we see that the top quarter of readings for these later years are almost entirely in the 2-3 CCSS grade band in distinction to the other years, where the readings remain in the K-1 CCSS grade band. This result provides evidence that, at least for the upper quarter of readings, these later years were more difficult.

In order to understand what may lead to the observed differences in RMM scores, we now examine some components of RMM and other useful language features to better understand the source of this change.

Vocabulary

An important aspect of text complexity is vocabulary choice. We examine four measures related to vocabulary, Time to Maturity, Google Frequency, Type-Token ratio and the rate of introduction of new words. Time to Maturity (TTM) is a vocabulary measure based on Word Maturity (Landauer, Kireyev, &

Panaccione, 2011) that has been shown to characterize vocabulary difficulty well by reflecting when students obtain proficiency with words (Biemiller, et al., 2014). Figure 4 shows 3rd Quarter TTM by level with years color coded in the left panel, and for comparison the median TTM is shown in the right panel. Median TTM is flat for all years indicating that the average word in these readings was part of the average person's vocabulary before exposure to text. We see strikingly different TTM profiles among the years with respect to 3rd quarter words. While for 1962 and 1983 the third quarter TTM is essentially zero, meaning that the words would be those that students would typically know before exposure to text, while 1971 shows a steady increase from 0 TTM to about 0.05. In 2013 the difficulty of the vocabulary increases at each level and after the second level is above all the other years. So there is substantial contrast between the number of words readers may find unfamiliar between 1962 and 2013.

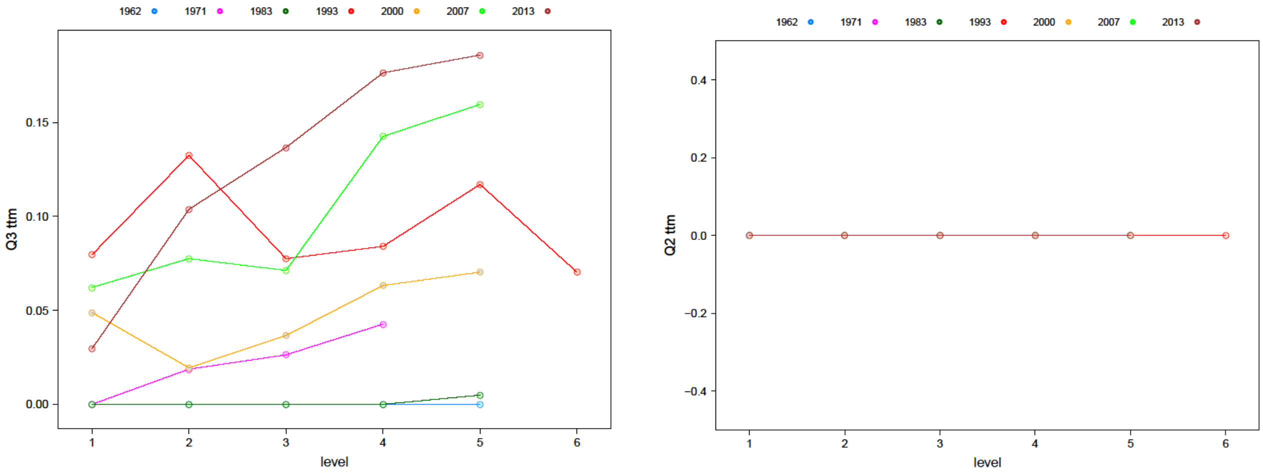

Figure 4. Time to Maturity by level, color code by year. Left panel: 3rd Quarter TTM; Right panel: Median TTM.

Another measure of vocabulary difficulty is to use the relative frequency of terms in a large corpus as a surrogate for difficulty. Figure 5 show the median Q1 frequency taken from the Google Book Corpus to estimate the vocabulary difficulty. Under this measure, again we see that 1962 and 1983 have vocabulary that appears substantially more frequency (hence inferred easier) than the other years. The anomalous behavior of low frequency for 1962, level 1 is due to the frequency of proper nouns, which predominate in those readings.

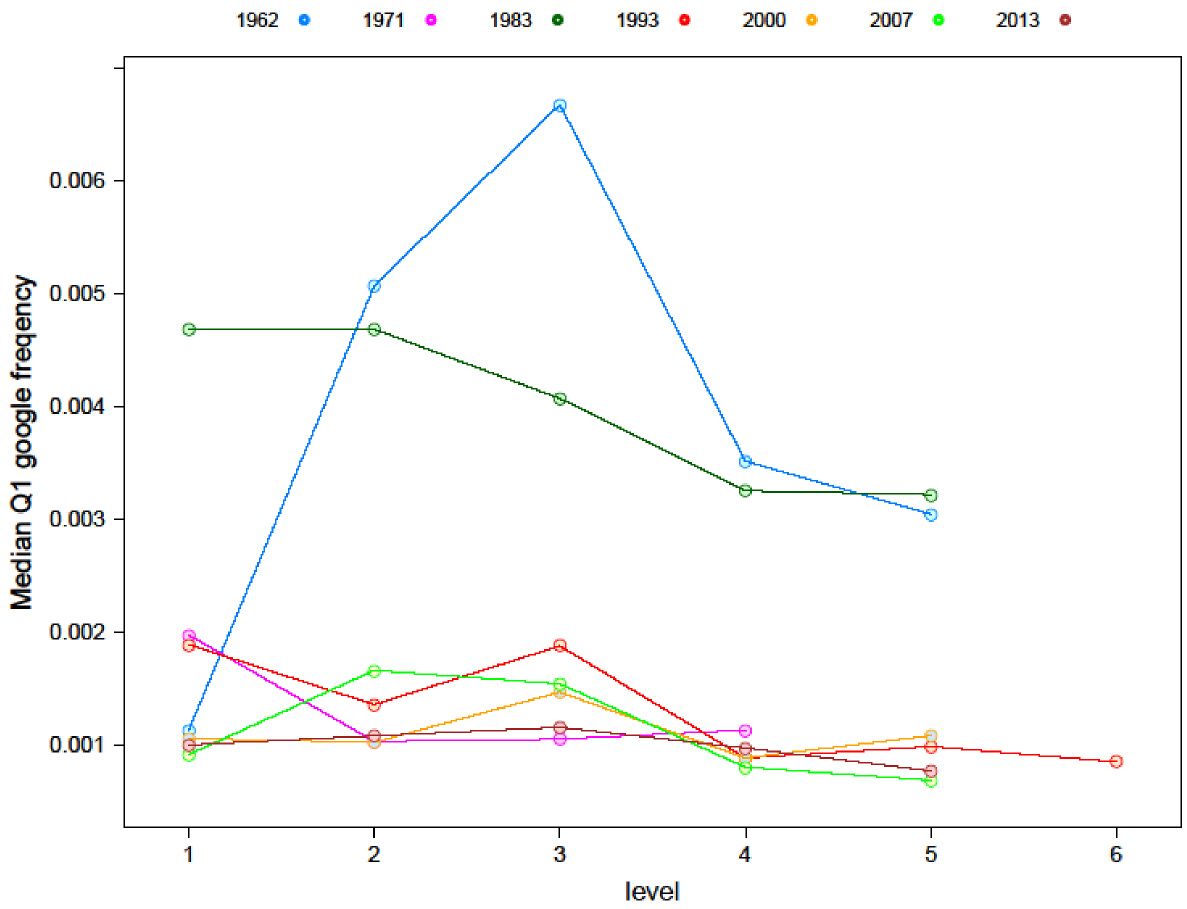

Figure 5. Median Q1 Google Frequency by level, color code by year.

The next measure, type-token ratio attempts to capture the variability of vocabulary use. The three earliest years (1962, 1971, 1983) have the lowest type token ratio indicating words are often reused, while later years tend to use a wider range of vocabulary. This indicates greater use of repetition in the earlier years as a writing style.

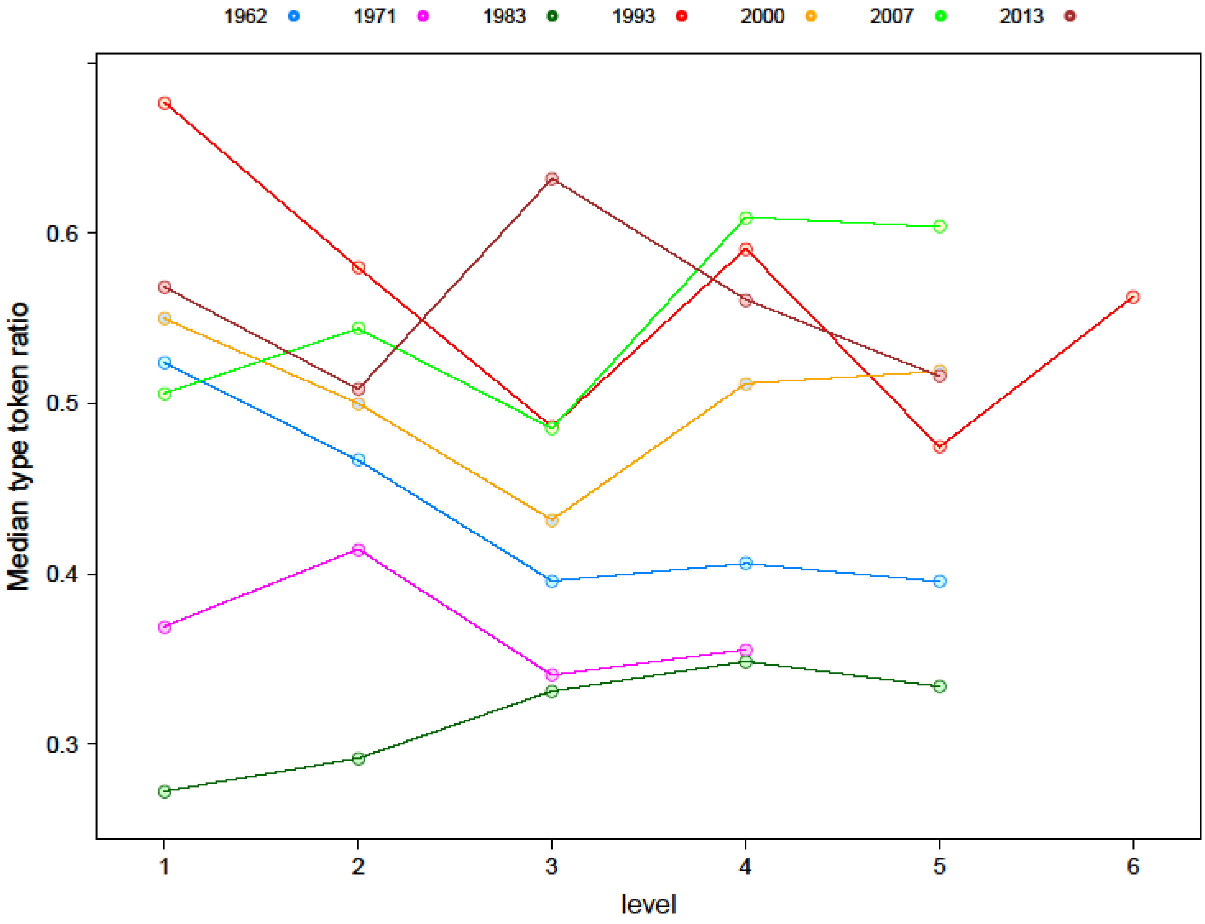

Figure 6. Median type-token ratio by level, color coded by year.

Repetition can also be examined in relation to how many new words are introduced into each text. Figure 7 displays the median unique words in each reading per level. This curve unlike the type-token ratio tends to follow length fairly closely. Of more interest is looking at the introduction of works along the text sequence which is shown in Figure 8.

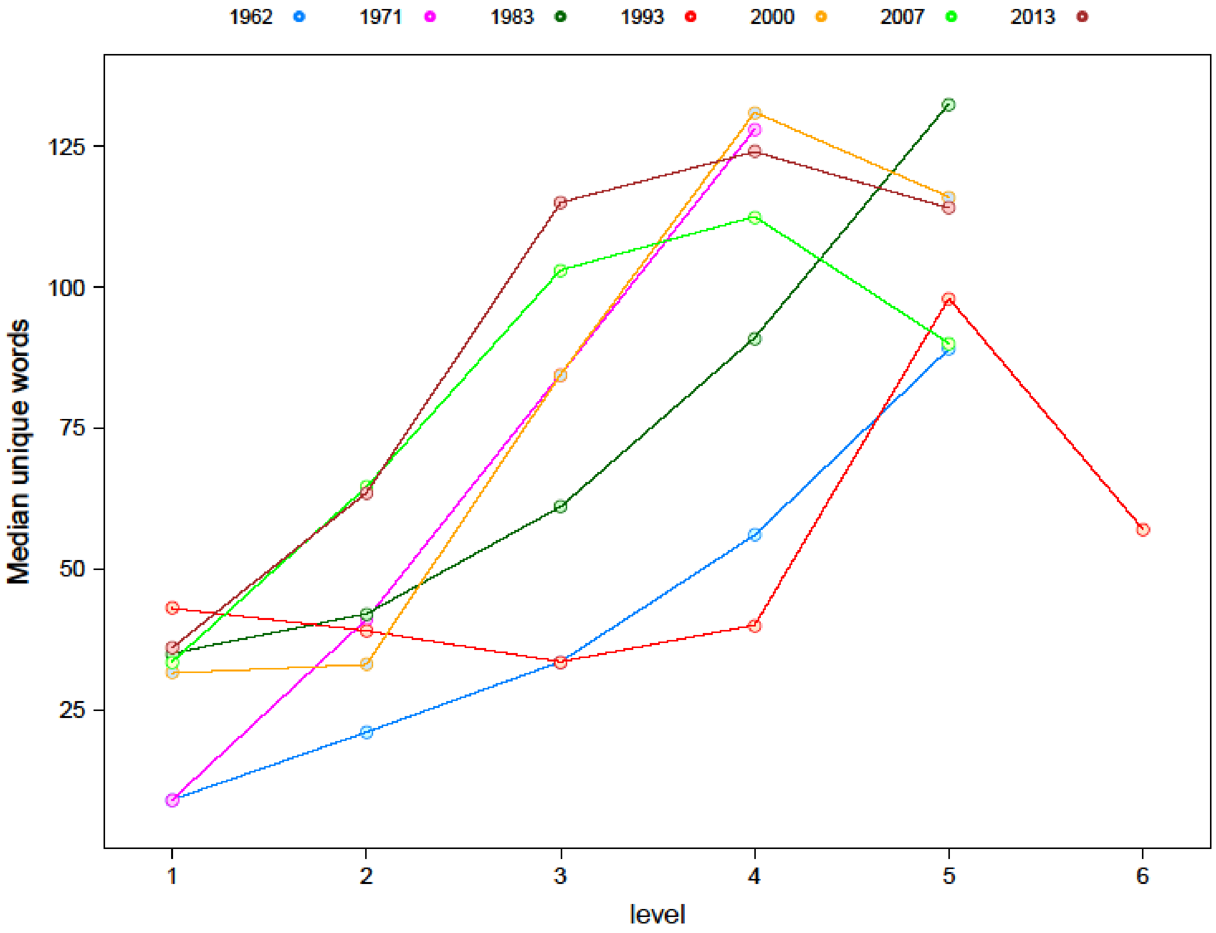

Figure 7. Median unique words by level, color code by year.

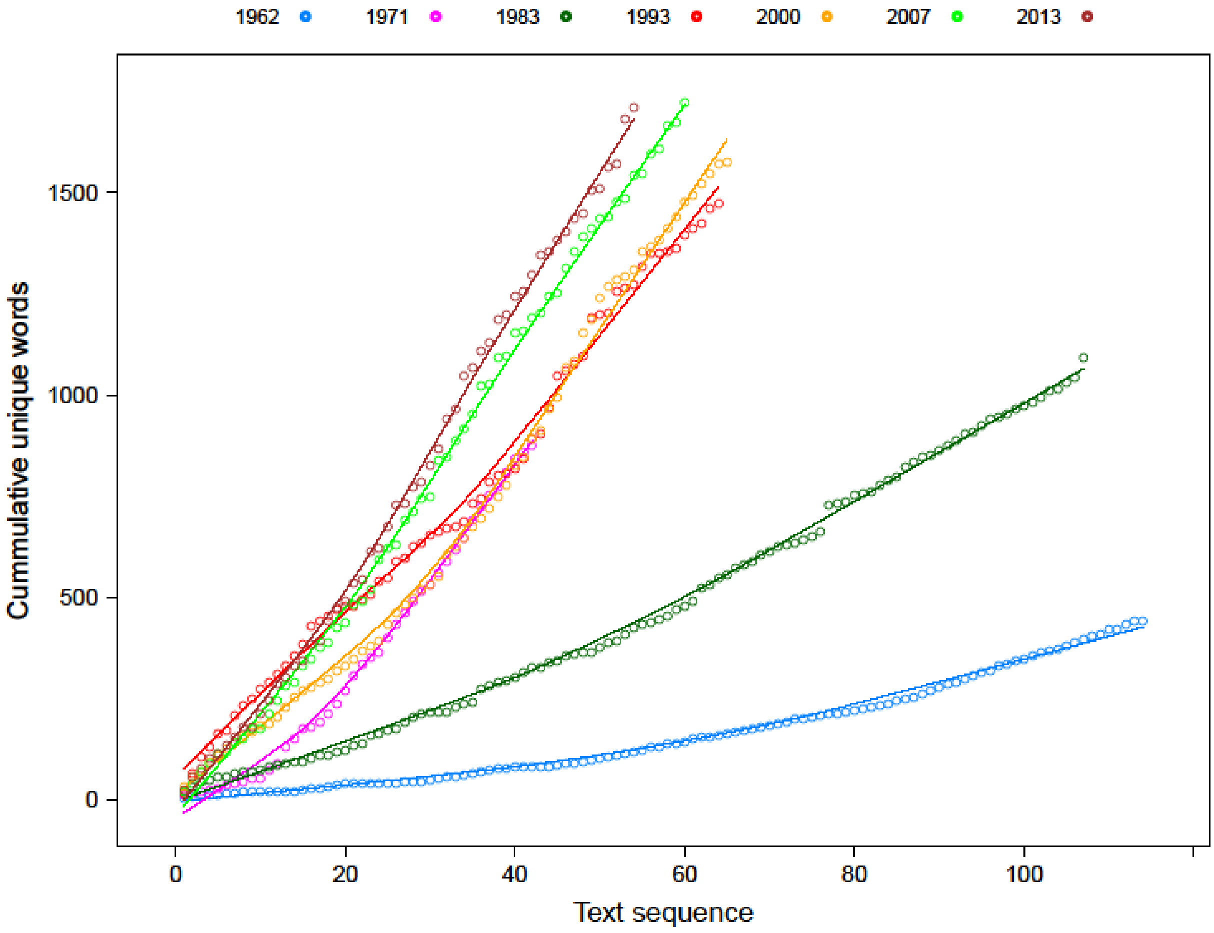

Figure 8. Cumulative unique words by level, color code by year.

Figure 8 is the first time we have displayed the readings in the full text sequence instead of by level. It shows the accumulation of vocabulary with each additional reading. In this figure and the next, which shows the number of new words for each reading we can potentially infer the pedagogical strategy, at least as reflected in the analysis of how the texts has changed between the years. In 1962 fewer than 500 words were introduced in over 100 texts, while by 2013 approximately 1700 words were introduced in approximately 50 texts. There is a clear pattern that in later years more words are introduced more quickly. Figure 9 provides a different visualization of this same pattern, by looking at the rate of introduction versus the cumulative introduced vocabulary.

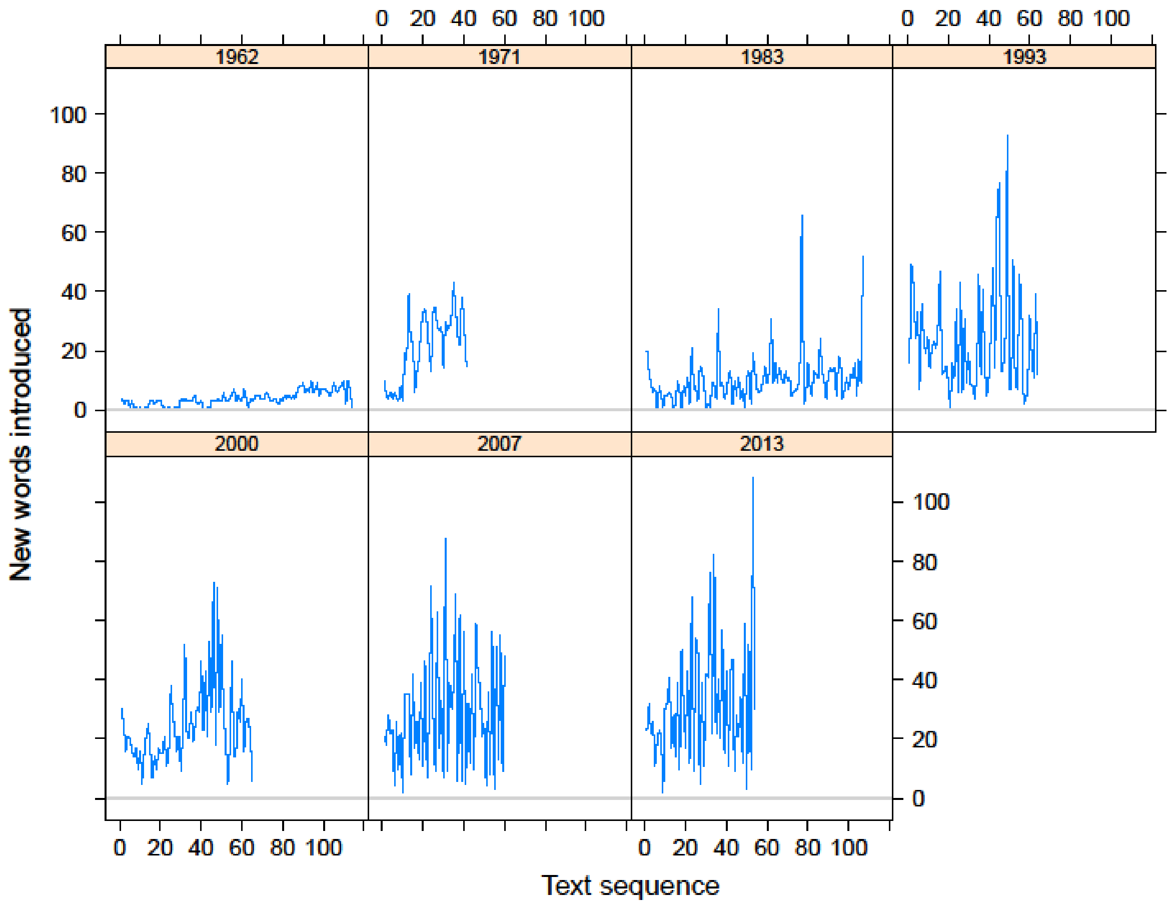

Figure 9. Introduction of new words along text sequence.

It is clear that in the later years more new words are being introduced in each text sequence.

Coherence

One coherence measure computes the semantic overlap of one sentence of the text to the next. This measure indicates how quickly new ideas or information are introduced at the level of meaning rather than just at the word overlap level. (e.g., Foltz, Kintsch, & Landauer, 1998). For example, to compute the coherence, of the text "SF71 Gr1.1.6 The red apples" (1971 level 1 example), we compute the LSA cosine similarity of each sentence to the next:

Text	Coherence to previous sentence
The Red Apples	NA
The horse saw the red apples.	0.37
The pig saw the red apples.	0.63
The duck saw the red apples.	0.95
The pig got on the horse.	0.08
The duck got on the pig.	0.48
The duck got the apples.	0.86

Median: 0.56, Mean: 0.56 (note assumes title is a sentence)
Table 3. Sentence to sentence coherence for example text.

Figure 10 shows the median sentence to sentence coherence for the texts by year and grade level. Overall, the earlier years show greater levels of coherence.

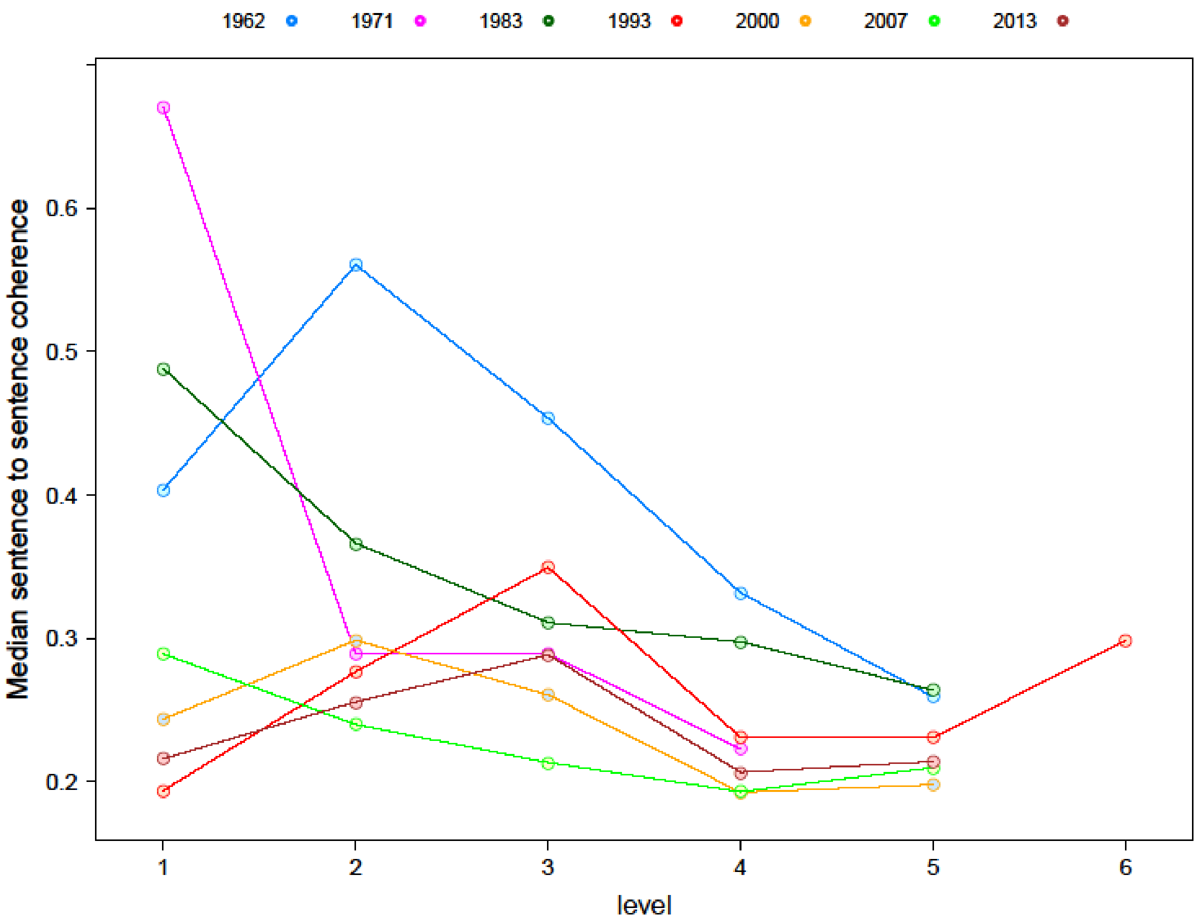

Figure 10. Median sentence to sentence coherence by level, color coded by year.

A related question is the degree of variability of content between texts along the instructional sequence. For example, if texts in a progression cover similar topics, then it would be expected that a student can use more of what was learned from the prior text to help with interpreting the next text. The following plot shows the semantic overlap (cosine similarity between successive pairs of readings), with greater values indicating more overlap from one text to the next in the sequence.

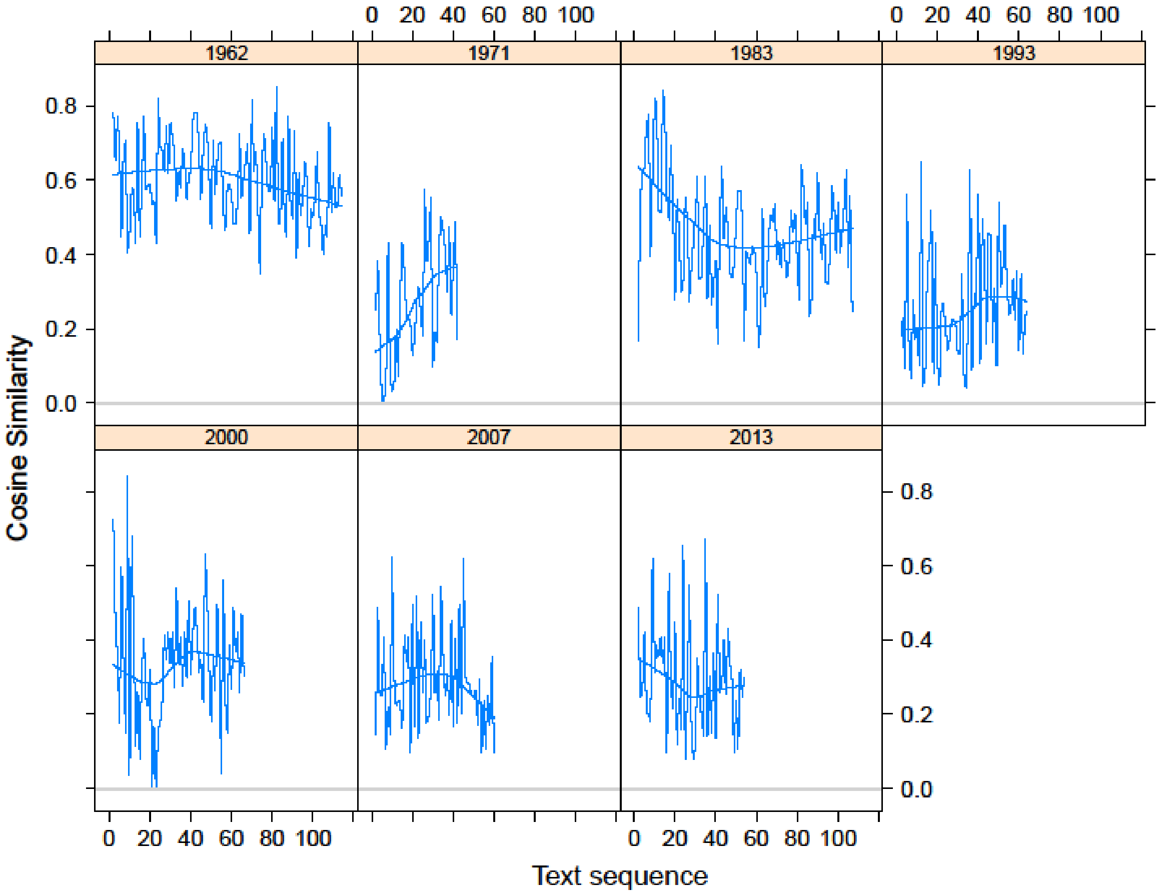

Figure 11. Cosine similarity of readings along text sequence by year.

In 1962 there is substantial overlap (semantic relatedness) between readings (cosine around 0.6). In 1971 there is much lower overlap, but it increases as the sequence progressives. In 2000 in the lower levels there is substantial variability among content, switching topics frequently and 1993, 2000, 2007 and 2013 have an overall overlap of about 0.3, but there is substantial variability in content. This shows that more and different types of topics were introduced in the later years.

Informativeness and English-like measures

Informativeness is a measure of the degree of information value of words. It was measured by the median vector length of words (e.g. see Landauer, Foltz & Laham 1998). Despite the smaller vocabulary, we see that in terms of informativeness, 1962 starts very low in informativeness, but is only exceeded by 1983 by year end. In all years the informativeness increases by level initially, but levels off in a number of cases.

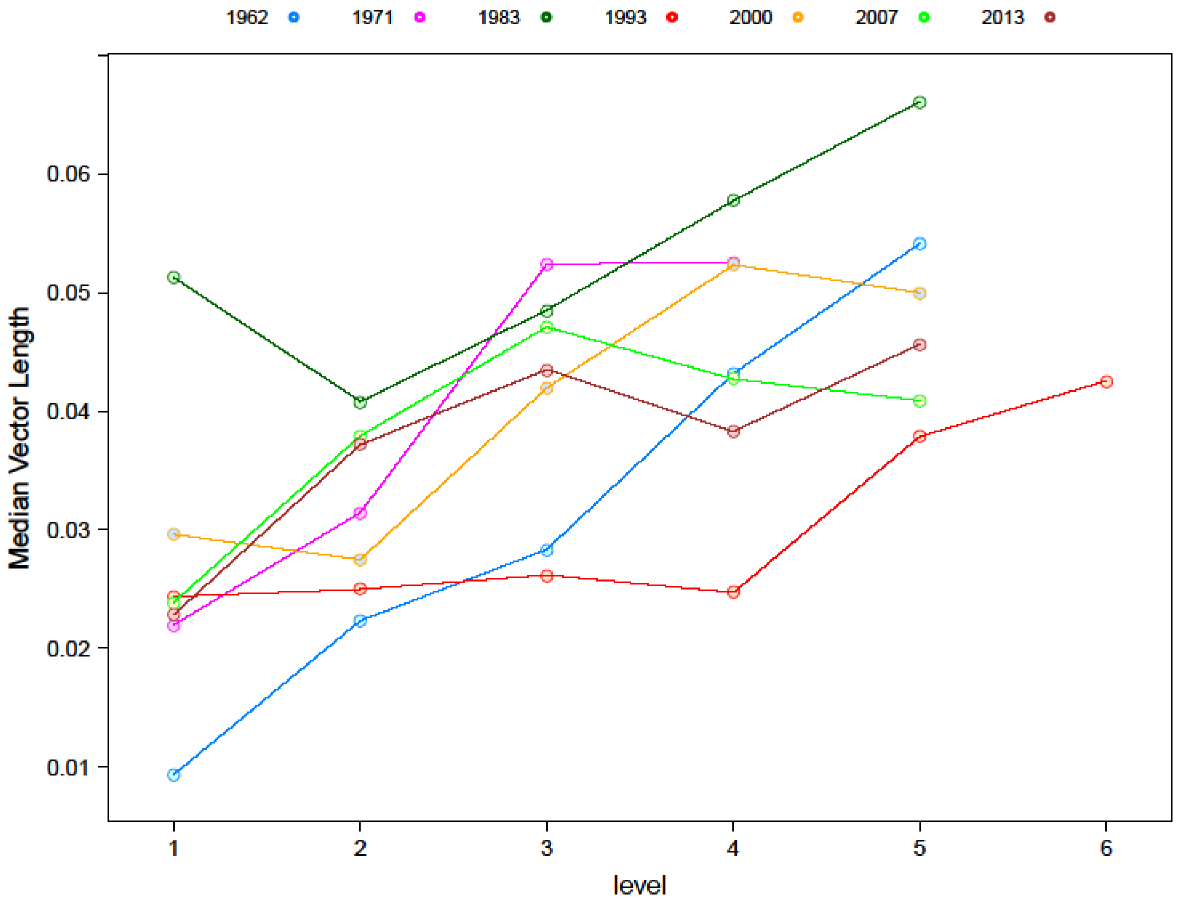

Figure 12. Median vector length by level, color coded by year.

Entropy provides a measure of how similar these texts are to the frequency patterns of bigrams in a large corpus. Figure 12 shows the median 2-gram entropy of texts by level and year. Despite the high coherence shown in the sentence to sentence coherence graph, 1962 is quite unlike the corpus text (unusual English constructions). Another interesting feature of this plot is that within each year, the bigram entropy tends to be fairly stable.

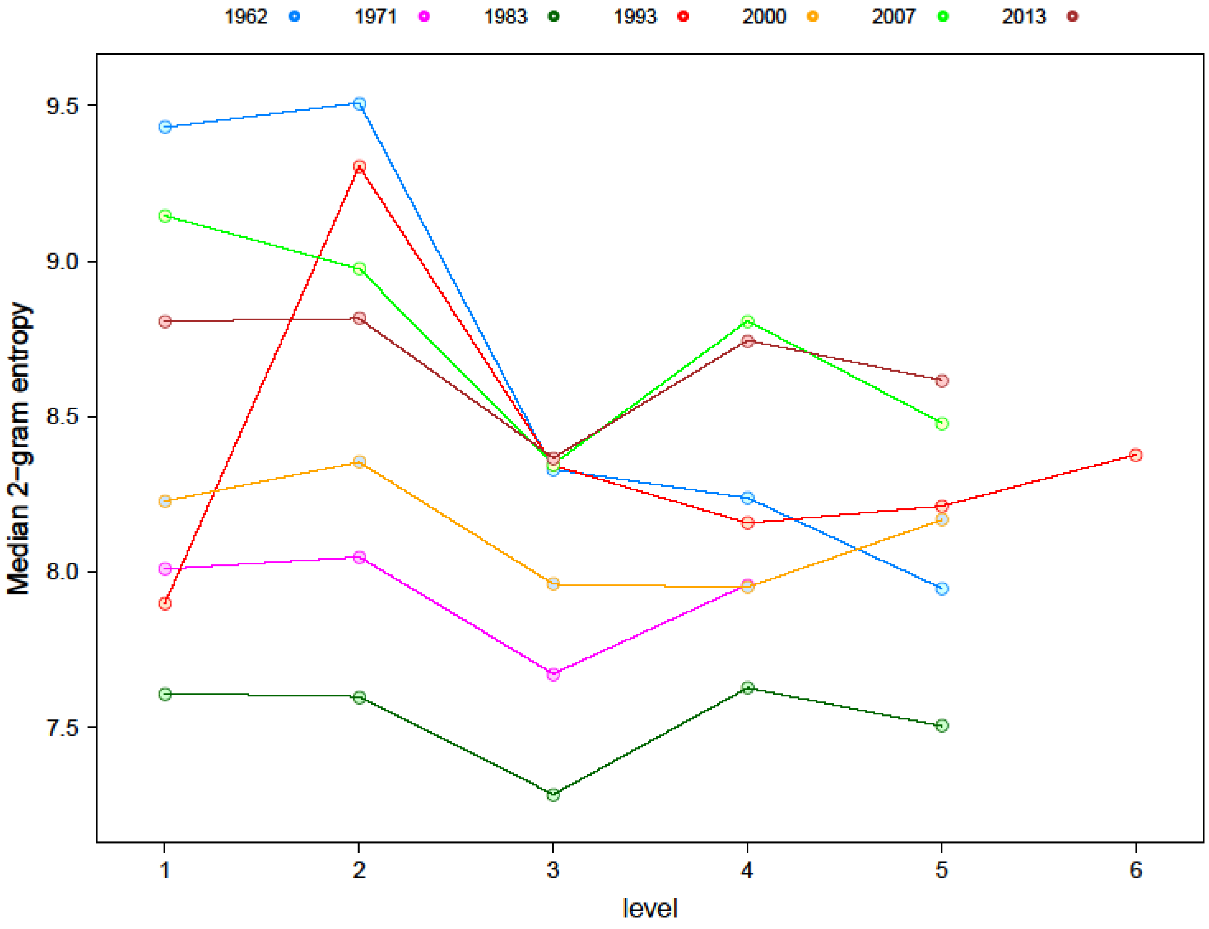

Figure 13. Median 2-gram entropy by level, color coded by year.

Discussion

In this study, First grade passages across 50 years were analyzed to derive overall RMM scores for each passage as well as measures of individual features that contribute the RMM scores. These features included measures of length, coherence, rate of introduction of new words, diversity of vocabulary, and usualness of the flow of words.

A number of issues arise when attempting to apply automated methods to analyze first grade readings contrasted with analysis of readings from upper primary grades and above. For many of these readings use of the illustrations is critical. For instance for reading "sf62-gr1.1.3-puff.txt" from 1962:

Look, Dick.
Help Sally.
Puff! Puff!

Look, Dick.
Look, Jane.
Look here.

What Puff is can only be told by the illustrations and also possibly from overlapping content amongst the items at this level. This text also illustrates the potentially difficult issue of evaluating the impact of proper nouns on text difficulty. Specifically as the demographics of the student population changes (see for instance Aud, Fox & KewalRamani, 2010) the familiarity of names is unlikely to remain constant.

While noting the limitations of automated approaches and this data set, the results indicate that generally, passage complexity, as measured by RMM, has increased from 1962 to 2013. For example, the RMM top quarter of readings tended into being categorized as CCSS grade band 2-3 in more recent years, while at the K-1 level in earlier years. Generally, texts of greater length were used in later decades. Analysis of individual features shows that passages from earlier decades tend to be more coherent, but less like standard written English. Passages from later decades introduced new words at much higher rate (30 words/text in 2013 vs. 4 words/text for 1962) and use words of greater difficulty.

The results provide empirical evidence that the complexity of the 1st grade passages have changed over the past 50 years. These changes reflect changes pedagogical strategies that are expressed through the choice of texts used and likely in how they were written or edited. Thus, the fact that the texts become more complex, in terms of rate of introduction of new words, lower repetition, diversity of topics and overall complexity indicates that 1st grade reading in the present day does show that the expectations of what constitutes successful reading acquisition at the first grade level have increased.

Acknowledgements

The authors would like to acknowledge the help of Megan Ross and Tamara Milbourn.

References

Aaron, I.E., Artley, A.S., Goodman, K.S., Huck, C.S., Jenkins, W.A., Manning, J.C. Monroe, M., Pyle, W.J., Robinson, H. M., Schiller, A., Smith, M.B., Sullivan, L.M., Weintraub, S., & Wepman, J.M. (1971). *Scott Foresman Reading Systems.* Glenview, IL: Scott, Foresman and company.

Aaron, I.E., Jackson, D., Riggs, C., Smith, R.G., & Tierney, R.J. (1981). *Scott, Foresman Reading*. Glenview, IL: Scott, Foresman, and Company.

Allington, R.L., Askew, B.J., Blachowicz, C., Butler, A., Cole, J., Edwards, P.A., Gonzalez, G.A., Harris, V.J., Hutchinson, S.M., Morrow, L.M., Sebesta, S.L., Sulzby, E., & Tierney, R.J. (1993). *Celebrate Reading.* Glenview, IL: Scott Foresman.

Afflerbach, P., Beers, J.W., Blachowicz, C.L.Z., Boyd, C.D., Diffily, D., Gaunty-Porter, D., Harris, V., Leu, D.J., McClanahan, S.D., Monson, D.L., Pérez, B., Sebesta, S.L., & Wixson, K.K. (2000). *Scott Foresman Reading.* Glenview, IL: Scott Foresman.

Afflerbach, P., Blachowicz, C.L.Z., Boyd, C.D., Cheyney, W., Juel, C., Kame'enui, E., Leu, D.J., Paratore, J.R., Pearson, P.D., Sebesta, S.L., Simmons, D., Vaughn, S., Watts-Taffe, S., & Wixson, K.K. (2007). *Reading Street.* Glenview, IL: Scott Foresman.

Afflerbach, P., Blachowicz, C.L.Z., Boyd, C.D., Izquierdo, E., Juel, C., Kame'enui, E., Leu, D., Paratore, J.R., Pearson, P.D., Sebesta, S., Simmons, D., Taffe, S.W., Tatum, A., Vaughn, S., Wixson, K.K. (2013). *Reading Street (Common Core*). Glenview, IL: Scott Foresman.

Aud, S., Fox, M., & KewalRamani, A. (2010). Status and trends in the education of racial and ethnic groups (NCES 2010-015). National Center for Education Statistics. Washington, DC: U.S. Government Printing Office.

Biemiller, A., Rosenstein, M., Sparks, R., Landauer, T. K, & Foltz, P. W. (2014). Models of Vocabulary Acquisition: Direct Tests and Text-Derived Simulations of Vocabulary Growth, *Scientific Studies of Reading*, 18(2), 130-154, DOI: 10.1080/10888438.2013.821992

Foltz, P. W., Kintsch, W,. & Landauer, T. K. (1998). The measurement of textual coherence with Latent Semantic Analysis. *Discourse Processes, 25*(2&3), 285-307.

Landauer, T. K, Foltz, P. W., & Laham, D. (1998). An introduction to Latent Semantic Analysis. *Discourse Processes*, *25*(2&3), 259-284.

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

Nelson, J**.**, Perfetti, C., Liben, D., & Liben, M. (2012). Measures of Text Difficulty: Testing their Predictive Value for Grade Levels and Student Performance. Student Achievement Partners*.* Available from http://www.ccsso.org/Documents/2012/Measures of Text Difficulty_final.2012.pdf

Robinson, H.M., Monroe, M., & Artley, A.S. (1962). *The New Basic Readers: Curriculum Foundation Series.* Chicago, IL: Scott, Foresman, and Company.

Appendix 1: Data Issues

We spot checked the text files and noted a number of issues, which are described in this Appendix. Without the original materials, we decided not to attempt to infer/guess the original intent and for the most part (except as noted) the analysis is based on the texts as we received them.

Character Encoding

The data were in a number of different and undocumented character encodings including ISO-8859 and UTF-8. We attempted to unify the encodings into ASCII, before proceeding with the analysis. Almost all the non-ASCII characters seemed to arise from various “smart” quotes. This unification allowed us, for instance, to automatically count can’t and can't as the same word.

Line Ending Punctuation and Titles

The following is an example from file “SF71 Gr1.1.4 The Boy.txt”.

```
The boy
The boy is eating the sandwich
The boy is eating cake
The boy is eating pie
The boy is eating cherries
Now the boy is in bed.
```

This example indicates two issues. The first is that without the original source material there is no simple, non-error-prone automated way to determine the presence or absence of a title. The second is that for unexplained reasons sentence ending punctuation is missing. Both of these issues can impact automated assessment.

Potentially duplicated text

The longest text file “sf07-gr1.3.7-frog-and-toad-together.txt” contains 880 words, which seemed a potential outlier as the next longest text contains less than 700 words. It appears that this text is potentially the concatenation of two texts. Given the wide range of text lengths, it is difficult to automatically detect this type of concatenation and it is not clear if this is a rare event or how widespread this issue is, but caution is required.

Sidebars, captions, and other text intrusions

Some items had what seemed like extraneous text, possibly captions.
For instance in file “Sf07-gr1.5.12-bens-great-ideas.txt”.

```
Ben's Great ideas.
Benjamin Franklin proved that lightning is electricity.
Franklin invented lightning rods. They are put on houses to prevent fires caused by lightning.
Chair with ladder, Bifocal glasses, Franklin stove.
```

It seems like the highlighted text is a caption. Similarly in file “sf07-g1.5.9-alexander-graham-bell.txt”, it appears a timeline is interspersed with the text.

> Alexander Graham Bell.
> Alexander Graham Bell was born in Scotland in 1847. His father was a famous teacher who taught people how to speak well.
> Scotland; time line, 1847, born. Alexander's mother was deaf. She was still able to teach him to play the piano. Alexander was good at music and science. The Bell family. Alexander is on the left. Alexander was interested in sound. He also liked to invent things. He built a machine that could speak. He also tried to make his dog talk. Alexander using an early invention. In 1871, Alexander moved to Boston. During the day, he taught deaf students how to speak. At night, he did experiments with sound. Classroom at a school for the deaf. 1981: teaches deaf students in Boston. Alexander wanted to learn more about electricity. In 1874, he met Tom Watson. Tom knew how electricity worked. They began to work together. 1874: begins work with Tom Watson. Electricity can make sound travel through wires. Alexander stopped teaching. He did experiments day and night. He and Tom wanted to invent a machine that could send voices from one place to another. A model of Bell's first telephone. 1874: begins work with Tom Watson. On March 10, 1876, Alexander and Tom reached their goal. Alexander spoke to Tom through the first telephone. Alexander Graham Bell using his telephone. 1876: invents telephone with Tom Watson. Alexander and Tom made the telephone better. Soon it could send voices many miles. In 1915, they made the first telephone call across the United States. The first telephone call across the United States. 1915: makes first telephone call across the United States. Alexander spent his life inventing. He died in 1922. Alexander Graham Bell changed the way people communicate with one another.

It seems likely that the highlighted text is from captions and/or a timeline. In addition we noted a likely transcription error (darker highlight/red highlight). Notice also that here there is a likely title (the first line) with sentence ending punctuation. While most of the files do not seem to have these type of issues, it is worth noting that at least some of the content analyzed differs in potentially important ways from the content the students actually read.