METAMETRICS RESEARCH BRIEF

Empirical Lexile Measures for Words

Steve Lattanzio, Jeff Elmore and A. Jackson Stenner, Ph.D.

August 2016

OBJECTIVE

MetaMetrics® has historically provided measures of text complexity for books, articles, and other texts using The Lexile® Framework for Reading. A text is made up of many words, which raises the question: can a single word have a Lexile® measure? Quantifying word difficulty on the Lexile scale would provide a variety of educational benefits. Example utilities could include identifying challenging words in a text and more precise word selection for both human- and machine-generated assessment items. This research brief describes recent efforts to find empirical measures of the complexity (difficulty) of words.

The research-based solution described in this brief met several requirements selected to increase applicability. First, empirical word measures are connected to the Lexile scale. That is, the word measures should have some meaningful relationship to the text measures and not be arbitrarily or independently scaled. Second, whereas an equivalent text complexity and student ability measure yields an expected 75% comprehension rate within the Rasch model as used within the Lexile Framework (Rasch, 1960), an equivalent text complexity, word difficulty, and student ability measure should also be anchored to a 75% comprehension rate in a modified Rasch model. Finally, when determining the empirical complexity of words, circularity should be avoided such that the word measure is determined independently of a student measure that resulted from that student's performance on that particular word.

Data from EdSphere® (Hanlon et al, 2015), a personalized online learning platform was used in this analysis. The Lexile empirical word measures are used to generate targeted vocabulary lists in the Lexile "PowerV®" Word Selector (www.Lexile.com/powerv) and have also been used to develop a corpus-based measure for a more expansive set of words (Elmore, 2016).

METHODS

Participants:
This study includes historical EdSphere data gathered from the inception of EdSphere's predecessor, Learning Oasis™ (Hanlon, Swartz, Stenner, Burdick, & Burdick, 2012) – June 2007 – through September 2014. During this period, there were approximately 6.7 million valid items administered which constitute an item-level encounter. These items were spread over 110,204 unique articles (ranging from 0L to 2920L), 17,966 unique users (ranging from grade 3 through 12), and 59,740 unique words. Note that these numbers differ slightly from a previous study using the same data (Lattanzio, 2015) as the criteria for what constitutes a valid encounter was modified to better reflect the current research initiative.

Procedure:
For 5.7 million (out of the 6.7 million) item-level encounters, an estimate of student ability was assigned using the "Simple" method (Lattanzio, 2015). The Simple method uses the student's performances on temporally adjacent article encounters to estimate reader ability during a particular article encounter, yielding an estimate that is independent of the encounter of interest. Not all item-level encounters could be assigned an ability estimate due to encounters at the beginning or end of a student's record lacking temporally adjacent encounters. Each item-level encounter also has a text complexity measure (the theoretical Lexile measure of the article containing the item), a known word that was clozed, and a response record (right or wrong), which will be denoted as y.

The traditional Rasch model used with the Lexile Framework takes the form (with parameters in logits)

$$p = \frac{e^{\theta-\beta+1.1}}{1+e^{\theta-\beta+1.1}} \tag{1}$$

where p is the probability the item-level encounter is answered correctly, θ is the student's ability estimate, β is the theoretical text complexity of the article, and adding 1.1 logits places the targeting at 75% instead of 50% (a specification of The Lexile Framework for Reading). Note that one logit equals 180L on the Lexile scale.

The standard protocol used by MetaMetrics when individual auto-generated cloze item difficulties are unknown is to use the Ensemble Rasch Model (ERM) (Lattanzio, Burdick, & Stenner, 2012), which treats the item difficulty of an auto-generated cloze item as a stochastic variable (i.e. as if the item difficulty came from an ensemble of possible item difficulties). This ensemble is approximated as a normal distribution with a mean item difficulty equal to the text complexity and a standard deviation of 132L (Sanford, 2006). The ERM has the form

$$p = E\left\{\frac{e^{\theta-(\beta_a+\varepsilon)+1.1}}{1+e^{\theta-(\beta_a+\varepsilon)+1.1}}\right\}, \varepsilon \sim N(0,\sigma^2) \tag{2}$$

where E is the expected value operator, β_a is the theoretical text complexity of the article, and σ = 132L. There is no analytic expression for

Equation (2) and so it is often calculated through numerical integration.

For a given word, this solution finds the empirical word difficulty by effectively treating the item difficulty as a linear combination of the text complexity and the word difficulty. In essence, this is adding a variable to explain the variance in item difficulty previously just treated stochastically in the ERM, resulting in a word-text Rasch model (WTRM) of the form

$$p = \frac{e^{\theta - c\beta_a - (1-c)\beta_w + 1.1}}{1 + e^{\theta - c\beta_a - (1-c)\beta_w + 1.1}} \quad (3)$$

where β_w is the difficulty of the word and c is a single parameter describing the relative weighting between β_a and β_w that make up the item difficulty estimate. Since θ and β_a are fixed (in order to retain the existing features of The Lexile Framework for Reading), the only remaining parameters to solve for are c and β_w. While each word has its own value for β_w, there is only a single value for c to be used in the model. For a given value of c, there exists a unique value of β_w for each word, found directly by solving for the β_w that makes the expected number of correct responses to items equal to the observed raw score for a particular clozed word. This means that c can be found by systematically sampling over various values (specifically, values between zero and one), fitting β_w for each word for each unique value of c considered, and seeing which c produces the best fit to the data.

The best fit is determined by minimizing the log-loss (Gneiting & Raftery, 2007) between the success probabilities p from Equation (3) and the actual outcomes on the item encounters y. The log-loss is calculated as

$$logloss = -\frac{1}{N}\sum_{i=1}^{N} y_i \ln(p_i) + (1 - y_i)\ln(1 - p_i) \quad (4)$$

where N is the number of items, y_i is the actual outcome of item i (zero or one), and p_i is the success probability of item i. The log-loss is a proper scoring function, which means that the minimum expected log-loss corresponds to the true set of probabilities. That is, a prediction model should provide its best and properly calibrated probability estimates to minimize the expected log-loss.

Measures:
Lexile measures (Stenner et al., 1988): a developmental scale that measures reader ability and text complexity on a common scale using semantic and syntactic features. Independent psychometric studies of the Lexile scale (Mesmer, 2007; White & Clement, 2001) indicate that it is a valid and reliable measure of reader ability and text complexity. This research brief describes also placing individual words on the Lexile scale.

ANALYSES

The performance of the WTRM in Equation (3) was compared to the ERM in Equation (2). This was done by comparing both the point-measure (a point-biserial between a logit value and the dichotomous right/wrong response) and the total log-loss. Furthermore, Pearson correlations are provided between the empirical word measures and existing word measures from other sources, namely Age-of-Acquisition (Kuperman, Stadthagen-Gonzalez, & Brysbaert, 2012), EDL (Taylor, Frackenpohl, & White, 1979), and PPVT™ (Dunn & Dunn, 2012). These correlations are corrected for the artifacts associated with the reliability of the empirical measures and the range restriction of the available words (Hunter & Schmidt, 2004).

RESULTS & DISCUSSION

The log-loss was minimized at a value of $c = 0.46$. The resulting distribution of empirical word difficulties is shown in Figure 1. Figure 2 shows the relationship between the exponent in the WTRM and the exponent in the ERM. Note that the stretching out of the WTRM exponent is due to the fact that the ensemble interpretation is removed and that the two exponents are not expected to line up along the 45-degree line.

Point-measures for the ERM in Equation (2) were calculated as the Pearson correlation between θ - β_a and the dichotomous right/wrong responses y. Point measures for the WTRM in Equation (3) were calculated as the Pearson correlation between θ - $c\beta_a$ - $(1$ - $c)\beta_w$ and y. The point-measure for the ERM is 0.32 and the point-measure for the WTRM is 0.41. Correcting for the reliabilities of the student measures ($r_{student}$ = 0.55) and student and word measures ($r_{student\text{-}word}$ = 0.70) over the items taken yields point-measures of 0.43 for the ERM and 0.50 for the WTRM. The reliability of the combined person and item parameters was the only artifact considered. There is no correction for the fact that y is dichotomous in order to retain the interpretability of a point-measure. Note that measures were only included if they had an estimated standard error less than 300L. The choice for a cutoff is fairly inconsequential since including measures with higher standard error will lower the raw observed correlations, but will also produce lower reliabilities that yield higher corrected correlations. Note, that this treatment considers the theoretical Lexile measure of the article to have perfect reliability. Theoretical Lexile measures for articles in EdSphere have been shown to have near perfect reliability (Lattanzio, 2015).

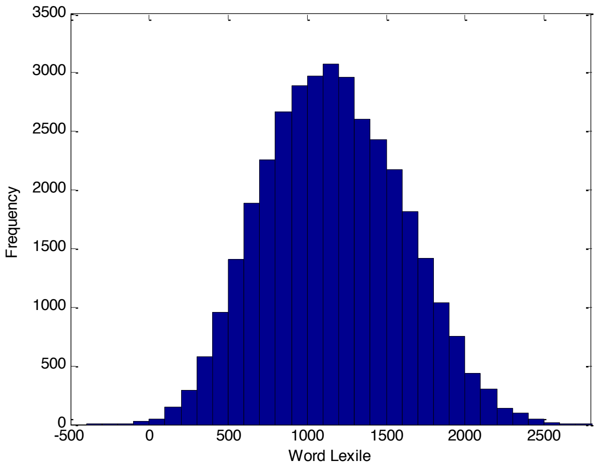

Figure 1. Distribution of empirical word measures.

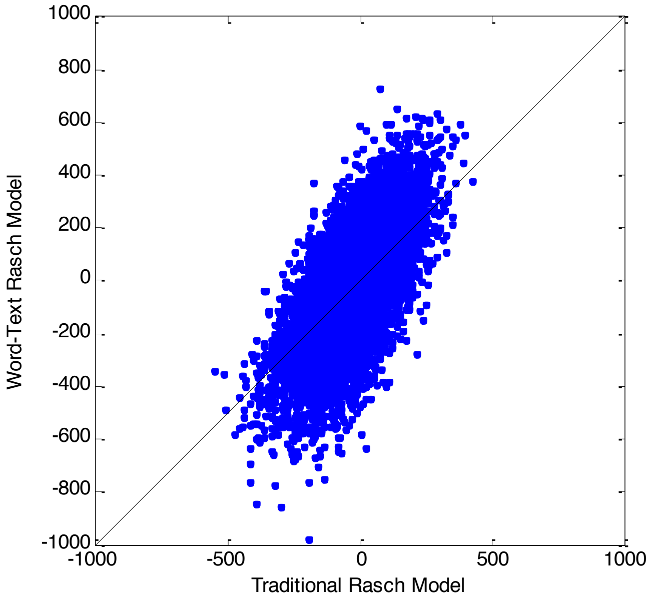

Figure 2. $\theta - c\beta_a - (1 - c)\beta_w$ vs. $\theta - \beta_a$

Table 1 shows the relationship between the empirical Lexile word difficulties and three other measures: Age of Acquisition, EDL, and PPVT. The table includes observed Pearson correlations (r_o), artifact corrected Pearson correlations (r_c), and Spearman's Rho correlations (ρ), along with the sample size for the sample shared words (n). Note that Pearson correlations involving empirical Lexile word difficulties are the only ones that are corrected for the estimated reliability of the empirical measures and range restriction as they are the only ones where that information is available for this study. Spearman's Rho correlations are also provided because the scales for the different measures are not necessarily collinear. Only words with estimated standard errors (for their empirical Lexile measures) less than 100L were included.

The reliability was estimated using the word difficulty estimates and their estimated standard errors for the words shared between the measures being compared. The reliability of the empirical Lexile measures is 0.97, 0.98, and 0.97 for the comparisons with Age of Acquisition, EDL, and PPVT, respectively.

Range restriction is the ratio between the standard deviation of the sample and the standard deviation of an ideal sample. In this study, the ideal sample is a normal distribution of word difficulty measures where 95% of the word measures are between 0L and 2000L, which turns out to be a distribution with a standard deviation of 510L. This makes the range restriction for the correlations between empirical Lexile word measures and Age of Acquisition, EDL, and PPVT to be 0.77, 0.81, and 0.70, respectively.

Table 1

Correlations Between Different Word Measures

	Age of Acquisition	EDL	PPVT
Empirical Lexile	r_o=0.73, r_c=0.82, ρ=0.74 n=11,012	r_o =0.77, r_c=0.84, ρ=0.78 n =4,807	r_o =0.73, r_c=0.85, ρ=0.72 n=197
Age of Acquisition	-	r_o =0.84,ρ=0.84 n =8,750	r_o =0.86, ρ=0.87 n =379
EDL	-	-	r_o =0.76, ρ=0.77 n =219

REFERENCES

Dunn, L. M., & Dunn, D. M. (2012). *Peabody Picture Vocabulary Test: PPVT™-4*. Johannesburg: Pearson Education Inc.

Elmore, J., Lattanzio, S., Stenner, A., Sanford-Moore, E. (2016). Calculation of Lexile Word Measures using a Corpus-Based Model and Student Performance Data (MetaMetrics Research Brief). Durham, NC: MetaMetrics.

Gneiting, T., & Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477), 359-378.

Hanlon, S.T., Neuvirth, K., Tendulkar, J., Houchins, J., Swartz, C. S., & Stenner, A. J. Edsphere [Computer software]. Durham, NC: MetaMetrics.

Hanlon, S. T., Swartz, C. W., Stenner, A. J., Burdick, H., & Burdick, D. S. Learning Oasis [Computer software]. Durham, NC: MetaMetrics.

Hunter, J. E., & Schmidt, F. L. (Eds.). (2004). *Methods of meta-analysis: Correcting error and bias in research findings*. Sage.

Kuperman, V., Stadthagen-Gonzalez, H., & Brysbaert, M. (2012). Age-of-acquisition ratings for 30,000 English words. *Behavior Research Methods*, 44(4), 978-990.

Lattanzio, S. M., Burdick, D. S., & Stenner, A. J. (2012). Ensemble Rasch models (MetaMetrics White Paper). Durham, NC: MetaMetrics.

Lattanzio, S. M. (2015). Empirical Lexile measures for EdSphere (MetaMetrics Research Brief). Durham, NC: MetaMetrics.

Mesmer, H. (2007). *Tools for matching readers to text: Research based practices*. Guilford Publications, Inc.

Rasch, G. (1960), *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danish Institute for Educational Research.

Stenner, A. J., Horablin, I., Smith, D. R., & Smith, M. (1988). The Lexile Framework. Durham, NC: MetaMetrics.

Sanford-Moore, E. E. (2006). PASeries Reading—Technical manual. Durham, NC: MetaMetrics.

Taylor, S. E., Frackenpohl, H., & White, C. E. (1979). *EDL core vocabularies in reading, mathematics, science, and social studies*. New York: McGraw-Hill.

White, S., & Clement, J. (2001). Assessing the Lexile Framework: Results of a Panel Meeting. Working Paper No. 2001-08. *National Center for Education Statistics.*

For more information, visit www.MetaMetricsInc.com.

MetaMetrics® is focused on improving education for students of all ages. The organization develops scientific measures of academic achievement and complementary technologies that link assessment results with instruction. For more than twenty years, MetaMetrics' work has been increasingly recognized worldwide for its distinct value in differentiating instruction and personalizing learning. Its products and services for reading, mathematics and writing provide valuable insights about academic ability and the potential for growth, enabling students to achieve their goals at every stage of development.

MetaMetrics®
LINKING ASSESSMENT WITH INSTRUCTION