Background on Empirical Validation of Text Complexity

Hal Burdick
New Technologies, MetaMetrics inc.

Without some form of brain imaging, the best way to determine the complexity of a text is by imposing a response requirement on a reader to check for understanding. It is important that item writing protocols focus on the complexity of the text itself avoiding pitfalls in item writing that add easiness or hardness to the understanding of the text.

Early in the development of the *Lexile Framework for Reading*, researchers evaluated the predictability of reading items and determined which item type was best suitable for predicting a text's complexity. The embedded sentence completion item, often called Lexile native items, won the day. Figure 1 shows an example of a native Lexile item.

Figure 1 Native Lexile Item written from R. L. Stine's *The Barking Ghost*.

She disappeared through the trees. "Fine with me," I thought angrily. It would be fine with me if I never saw her again. I am glad she is _____.

A. gone
B. first
C. best
D. sitting

The item consists of authentic text written by professional authors and an embedded sentence produced by an item writer. The protocol emphasizes the complexity of the text more than the stem or foils. The reader's task is to replace a word in the blank which best completes the sentence given the text. All options make sense in the embedded sentence, but only one choice is unambiguously correct given the passage. The text selected is source targeted so that highly difficult items are not written from low Lexile sources and vice versa.

Table 1 shows how this task type performs in terms of root mean square error (RMSE). RMSE is calculated as the standard deviation of the differences between theoretical and observed measures for the items.

Table 1. Root Mean Square Error of Native Lexile Items from two large sample studies of over 10,000 respondents per item

	Grade	*n* of items	RMSE
Study 1	4	40	154L
	5	40	151L
	6	40	160L
	7	60	150L
	8	60	151L
	9	60	137L
Study 2	1	40	163L
	2	40	173L
	3	45	170L
	4	54	150L
	5	54	150L
	6	54	198L
	7-8	70	172L
	9-11	70	191L

.
Notice the fairly consistent RMSE for each grade for both studies which has a floor around 150L (one exception is grade 9 of 137L for Study 1). Despite many attempts at improving the theory this 150L floor remained the best that theory could do for many years.

During that time characteristics of both text and item-type were evaluated. Table 2 shows a list of the types of variables researchers investigated.

Table 2. List of text and item features with little or no impact on reducing RMSE of Lexile specification equation.

Source of Predictor	Predictor
Text	Dialogue (number of characters between quotations)
	Fiction vs. Non-Fiction
	Passage Length
	Picture Support*
	Lexile measure of the source
	Subjective Review
Item	Correct Foil position**
	Item writer bias
	Lexile level of correct option
	LSA of correct option with passage
	Maximum Lexile level of any foil
	Stem Length

* added easiness of 120L, but no reduction in RMSE.
** first position mildly significant. Other positions the same.

Given that the difference from grade level to grade level is roughly 100L, 150L is a fairly large error for targeting a reader to text. A field study in 1998 evaluated the effects on reader measures when using items calibrated by theory. Table 3 shows the median Lexile measure for a sample population of students was highly reproducible even with large errors at the individual item level. Medians were used instead of means to avoid problems with scoring perfect scores.

Table 3. Reproduction of Lexile Medians given forms using theoretical measures for item difficulties.

	Grade 2 Test (n=40)	Grade 3 Test (n=40)	Grade 4 Test (n=40)	Grade 5 Test (n=40)	Grade 6 Test (n=40)
3rd Grade Readers (n=107)	345L	340L			
4th Grade Readers (n=123)		685L	679L		
5th Grade Readers (n=123)			834L	818L	852L

This startling reproduction of the medians caused us to reevaluate our approach to theory. We knew there was variability created by item type. Figure 2 shows the empirical Lexile measures for three items written to the same passage shown in Figure 1.

Figure 2. Three Empirical Lexile Measures for items written to the same passage.

Passage (430L) *R. L. Stine The Barking Ghost*	**Observed Difficulty**	**Embedded Completion Items**				
She disappeared through the trees. "Fine with me," I thought angrily. It would be fine with me if I never saw her again.	**269L**	I am glad she is _____.	gone	first	best	sitting
	632L	I was _____.	upset	happy	polite	hungry
	740L	I _____ her.	disliked	forgot	told	chased

.

The range of 471L in the item difficulties is severe, but the effect on reader measures is negligible. This shows that the observed variability in item difficulties is likely random and unbiased which suggests taking an ensemble interpretation when evaluating theory.

The ensemble interpretation for Lexiles is explained in detail in the cited paper published in the Journal of Applied Measurement.

REFERENCES

Stenner, A. J., Burdick, H., Sanford, E. E., & Burdick, D. S. (2006). How Accurate are Lexile Text Measures? Journal of Applied Measurement. 7(3), 307-322.