Metrological Traceability in the Social Sciences: A Model from Reading Measurement

A Jackson Stenner
CEO, MetaMetrics, Inc.
Durham, North Carolina, USA
E-mail: jstenner@lexile.com

William P Fisher Jr
Research Associate, BEAR Center, Graduate School of Education, University of California, Berkeley, CA (USA) & Principal, LivingCapitalMetrics Consulting
E-mail: william@livingcapitalmetrics.com

Abstract. The central importance of reading ability in learning makes it the natural place to start in formative and summative assessments in education. The Lexile Framework for Reading constitutes a commercial metrological traceability network linking books, test results, instructional materials, and students in elementary and secondary English and Spanish language reading education in the U.S., Canada, Mexico, and Australia.

1. Introductory history

The ability to read is fundamental to education, and it is accordingly tested and measured more often than any other subject area. The index to the eighteenth edition of the Buros Mental Measurements Yearbook [1] includes over 140 tests with the word “reading” in their titles. This count does not include tests focused on vocabulary or word meaning, which are also numerous.

Though the issues are complex, literacy remains essential to productivity in the global economy [2]. The need for effective and efficient reading education will only intensify as communication, teamwork, information, and information management are increasingly demanded as basic skills [3].

2. Theory for reading measurement

Despite the longstanding fundamental importance of reading as the tool most essential to learning, reading research remained atheoretical until 1953 [4], and interest in a unified theory of reading is a relatively new phenomenon [5]. Further, in the years since 1953, available reading theories have not been used to inform the design or interpretation of assessments of reading ability[4].

Though it may seem counterintuitive, this failure to apply theory in the course of empirical measurement research is not unusual, nor is it restricted to reading research. On the contrary, measurement technologies in the natural sciences have historically been developed through socially-contextualized trial-and-error solutions to practical engineering problems, such as consistent, stable results, and not directly from theoretical principles [6,7]. Theory generally comes later, after researchers have had the opportunity to employ standardized technologies in the routine and repeated reproduction of a controlled phenomenon. Only then do applicable general principles emerge as useful insights that can be fed back into technical refinements.

Published under licence by IOP Publishing Ltd

1.1 Syntactic and semantic elements

Science inevitably involves reducing complex and rich phenomena to simpler models. The truth of the models is less an issue than their usefulness [8,9]. Simplification is usually achieved only in a context that respects necessary constraints and accepts the realization of a limited goal. The efficiency and power obtained when limited goals can be realized in the form of a useful tool, however, can sometimes confer great value on a simplified process.

In the 1950s, Rasch's formulation of a parameter separability theorem, the concept of specific objectivity, and models useful in practical measurement applications combined into an important step forward in educational measurement [10]. These developments were followed by Wright's introduction of improved estimation algorithms, model fit tests, and software in the 1960s, along with his vigorous championing of Rasch's ideas [11]. By the 1970s, enough data from reading tests had been successfully fit to Rasch models in the U.S. to support the viability of the Anchor Test Study, an equating of seven major reading tests involving over 350,000 students in all 50 U.S. states [12].

Lacking a theory capable of uniting different tests into a common framework, however, the uniform scale resulting from the equating study was made obsolete as soon as the tests were changed, which was, of course, immediately. Successful research predicting item difficulties on the Peabody Vocabulary Test and the Knox Cube Test (a measure of short term memory and attention span) [13,14], however, led to a new effort focused on reading.

Reading theories build on the fact that all symbol systems share two features: a semantic component and a syntactic component. In language, the semantic units are words. Words are organized according to rules of syntax into sentences [15]. Semantic units vary in familiarity and the syntactic structures vary in complexity. The readability of a text passage is dominated by the familiarity of the semantic units and by the complexity of the syntactic structures used in constructing the message. Many readability equations therefore use a two-variable equation to forecast text difficulty. The word-frequency and sentence-length measures combine to produce a regression equation, known as a construct specification equation [13,14]. This equation provides a theoretical model that is evaluated in terms of the proportion of the variance of reading comprehension task difficulties (or, more recently, the means of specification-equivalent ensembles of item difficulties [16]) that can be explained.

2.1 The specification equation

One approach to such a specification equation first employs the mean of the logarithm of the frequencies with which words in a text appear in a 550-million word corpus of K-16 texts. More specifically, the log frequency of the word family, which is more highly correlated with word difficulty, comprises one term in the equation. Word families include the stimulus word, all plurals, adverbial forms, comparatives, superlatives, verb forms, past participles, and adjectival forms. The frequencies of all words in the family are summed and the log of that sum is used in the specification equation.

The second term of the specification equation is the logarithm of the text's mean sentence length. This parameter is operationalized simply by counting and averaging the number of words in each sentence. The relationship of word frequency and sentence length to text readability was investigated in research that extended a previous study on semantic units [14]. This analysis involved calculation of the mean word frequency and the log of the mean sentence length for each of the 66 reading comprehension passages on the Peabody Individual Achievement Test. The observed difficulty of each passage was the mean difficulty of the items associated with the passage (provided by the publisher) converted to the logit scale.

A regression analysis based on the word-frequency and sentence-length measures produced a regression equation that explained much of the variance found in the set of reading comprehension tasks. The resulting correlation between the observed logit difficulties and the theoretical calibrations was 0.97 after correction for range restriction and measurement error [14].

The regression equation was further refined based on its use in predicting the observed difficulty of the reading comprehension passages on eight other standardized tests (see Table 1). Repeated and ongoing comparisons of theoretically expected calibrations with data-based estimates produced from test data analysis provide continually updated validity evidence.

The regression equation links the syntactic and semantic features of text to the empirically determined difficulty of text. That link, in turn, is reproduced across thousands of test items and millions of examinees in applications. The consistent display of the link over time provides a basis for using the equation to perform theory-based calibrations of test items and texts, thus rendering empirical calibrations necessary only as checks on the system.

Table 1 Correlations of theory-based calibrations produced by the specification equation and data-based item difficulties.

Test	# of Questions	# of Passages	$r_{(OT)}$ [a]	$R_{(OT)}$ [b]	$R^*_{(OT)}$ [c]
SRA	235	46	.95	.97	1.00
CAT-E	418	74	.91	.95	.97
Lexile	262	262	.93	.95	.97
PIAT	66	66	.93	.94	.97
CAT-C	253	43	.83	.93	.96
CTBS-U	246	50	.74	.92	.95
NAEP	189	70	.65	.92	.94
Battery	26	26	.88	.84	.87
Mastery	85	85	.74	.75	.77
Totals	1780	722			
Means			.84	.91	.93

[a] $r_{(OT)}$ = raw correlation between observed difficulties (O) and theory-based calibrations (T)
[b] $R_{(OT)}$ = correlation between observed difficulties (O) and theory-based calibrations (T) corrected for range restriction
[c] $R^*_{(OT)}$ = correlation between observed difficulties (O) and theory-based calibrations (T) corrected for range restriction and measurement error.

The theory-based calibration process begins from an analysis of a text in which the numbers of words per sentence are counted, and the commonality of the individual words is estimated. The theoretical logit is then a function of sentence length and word frequencies in the language stated in the specification equation:

$$(9.82247*\mathrm{LMSL})-(2.14634*\mathrm{MLWF})-\mathrm{constant} \qquad (1)$$

where LMSL is the log of the mean sentence length and MLWF is the mean of the log of the word frequencies. LMSL and MLWF are used as proxies for syntactic complexity and semantic demand [17]. These logits are then scaled as follows:

$$(\mathrm{logit} + 3.3)*180 + 200 \qquad (2)$$

The uncertainty (standard error) of the individual measures [18] is

$$SE = X * [L/r(L-r)]\text{^}1/2 \qquad (3)$$

which is the square root of the test length L divided by the count correct r times the L-r count incorrect, times an expansion factor X that depends on test width. This logit is then converted to the standard unit. The original study found success on items at about -3.3 logits as indicating the earliest reading ability, and set that level at 200. A practical top to the scale was at 2.3 logits, and this was set 1000 units higher, to 1200. A standard unit uncertainty for a well targeted 36-item test measuring with an uncertainty of about .40 logits is the original logit range of 2.3 - (-3.3) = 5.6 divided into the 1000 L range, times .40, which comes to about 71.

3. Practical implementation issues

Projected comprehension rates should not be the only factor influencing text selection. To make the quantified measure the sole determinant of a curricular decision would be analogous to reducing a table to its physical dimensions when its sentimental or historical value might also be relevant.

Initial efforts at deploying the unit of measurement quickly encountered a chicken and egg question from book publishers: why should they adopt the unit as a means of indicating the text complexity of their books and articles if there were no schools or students prepared to take advantage of that information? Conversely, state departments of education and school districts asked, why should they be interested in a universally uniform measure of reading ability if there were no books or articles to match with students' ability measures?

The solution arose when one publisher incorporated the unit in their own system, which involved both a reading curriculum and a reading assessment system. This bootstrapping of one instance of coordinated reader-text matching made the link to the unit more attractive to testing agencies, who could now point to an additional use for their results; to book publishers, who now were assured of a population of students with measures to match with their books; and to state departments of education and school districts, who could now effectively put the matching system to work.

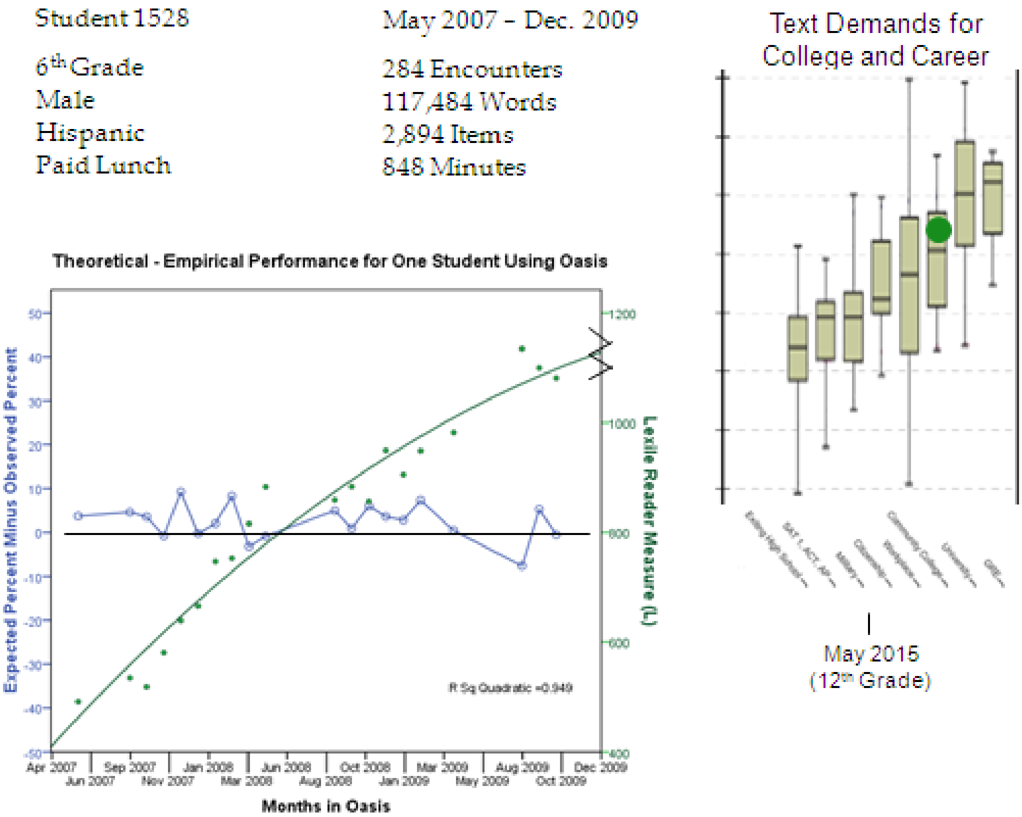

Figure 1. Individual student online reading measurement tracking system report

4. Conclusion

Over 28 million U.S. students, over half of all children in school, currently receive at least one reading ability measure per school year in the standard unit. The text complexity of tens of millions of magazine articles and hundreds of thousands of books are expressed in the standard unit, and all major reading assessments report measures in it.

The English-based system is in use in the U.S., Canada and Australia (with applications emerging in New Zealand, South Africa, and England), and in ESL applications in Korea, Japan, Malaysia, Hong Kong, and elsewhere in Asia. A Spanish system for matching readers and texts in the same unit is in use in Mexico and the Philippines. Researchers in various parts of the world are exploring possibilities for expanding the reader-text matching system to Mandarin, French, and other languages.

Educational textbook and curriculum publishers have developed online software applications for tracking individual student growth in reading ability. A report from one such system is shown in Figure 1. The value of repeated measures of a student over time and across texts is evident in the growth trajectory and the expected convergence of the student's ability with the reading demands of adult life.

Figure 2 shows the relationship between expected and observed text complexity measures in the online system. This plot illustrates the power of theory. Traceability to the standard unit is achieved not only by estimating student reading ability measures from data, but by gauging text complexity from its syntactic and semantic makeup. Given theory-based estimates of item difficulty, items can be adaptively selected for custom-tailored individualized administration, and those students' measures may then be estimated from their comprehension rates relative to the scale values of those items.

The specification equation operationalizes Rasch's notion of a frame of reference in a way that extends the frame beyond the specific objectivity obtained in the context of a particular test or set of equated tests to an indefinitely large collection of actual or virtual instruments. Theory-based instrument calibration eliminates the need to use data to both calibrate instruments and measure persons. The pay-off from using theory instead of data to calibrate instruments is large and immediate. When data fit a Rasch model, *differences* among person measures are free of dependencies on other facets of the measurement context (i.e., the differences are specifically objective). When data fit a causal or theory-enhanced Rasch model, *absolute* person measures are free of the conditions of measurement (items, occasions, etc.) making them objective beyond the limits of a specific frame of reference. In the theory -referenced context, person measures are individually-centred statistics in that no reference to another person(s) figures in their estimation.

One of the most important uses of reading test scores is to predict how a reader will perform on non-test tasks. For example, educators may be interested in the likelihood that a particular reader completes her first year of studies at a four year college. A regression equation could be fit to first year completion outcome data using reading test scores as predictors to estimate an individual first year completion likelihood on the basis of the relationship between test scores and completion found to hold in the population at large.

An alternative strategy is to imagine that first year college textbooks are virtual reading tests with item calibrations provided by the specification equation. Arbitrarily, but usefully, fixing a success rate (say 75%) on the virtual items for each textbook enables solving for the reader measure needed to correctly answer 75% of those virtual items. The individual reader's measure is then interpreted relative to the text complexity measure for each text in the freshman book bag. If the likely success rate in correctly answering the virtual items is high, so is the expectation of completing the first year.

High school graduates' reading measures can thus be compared to college text demands and a theoretically satisfying prediction can be made about the likelihood of first year completion. The efficiencies this system realizes from its use of scientifically validated predictive theory shows special promise as a tool for tracking reading readiness for post-secondary experiences in college, the work place, and the responsibilities of citizenship [19,20].

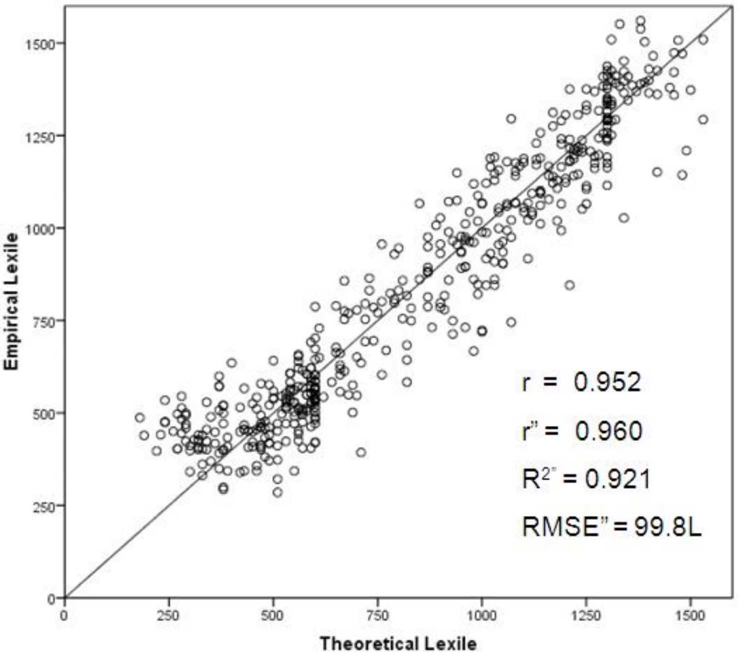

Figure 2. Empirical vs. theoretical Lexile text complexity estimates

5. References

[1] Spies R A, Carlson J F, Geisinger K F (Eds) 2010 *The Eighteenth Mental Measurements Yearbook* (Lincoln, Nebraska: University of Nebraska Press)

[2] Hamilton M, Pitt K 2011 *International Journal of Educational Development* **31** 596-605

[3] Neuman S B, Roskos K 2012 *The Reading Teacher* **66** 207-210

[4] Engelhard G 2001 *J App Meas* **2** 1-26

[5] Sadoski M, Paivio A 2007 *Scientific Studies of Reading* **11** 337-356

[6] Wise M N (Ed) 1995 *The values of precision* (Princeton, New Jersey: Princeton University Press)

[7] Bijker W E, Hughes T P, Pinch T (Eds) 2012 *The social construction of technological systems: New directions in the sociology and history of technology* (Cambridge, Massachusetts: MIT Press)

[8] Rasch G 1973/2011 *Rasch Measurement Transactions* **24** 1309

[9] Box G E P 1979 In R L Launer & G N Wilkinson (Eds.) *Robustness in statistics* pp. 201-235 (New York: Academic Press)

[10] Rasch G 1960 *Probabilistic models for some intelligence and attainment tests* (Reprint, with Foreword and Afterword by B. D. Wright, Chicago: University of Chicago Press, 1980 (Copenhagen, Denmark: Danmarks Paedogogiske Institut)

[11] Rasch G 1972/1988 *Rasch Measurement Transactions* **2** 19

[12] Rentz R R, Bashaw W L 1977 *J Ed Meas* **14** 161-179

[13] Stenner A J, Smith M 1982 *Perceptual and Motor Skills* **55** 415-426

[14] Stenner A J, Smith M, Burdick D S 1983. *J Ed Meas* **20** 305-316

[15] Carver R P 1974. *J Reading Behav* **6** 249-74

[16] Burdick D S, Stone M H, Stenner A J 2006 *Rasch Measurement Transactions* **20** 1059-60

[17] Stenner A J, Burdick D S 1997 The objective measurement of reading comprehension. www.lexile.com (visited 10 March 2013) (Durham, North Carolina: MetaMetrics, Inc.)

[18] Wright B D, Stone M H 1979 *Best test design* (Chicago, Illinois: MESA Press)

[19] Williamson G L 2008 *J Advanced Academics* **19** 602-632

[20] Williamson G L, Fitzgerald J, Stenner A J 2013 *Educational Researcher* **42** 59-69