One Fish, Two Fish

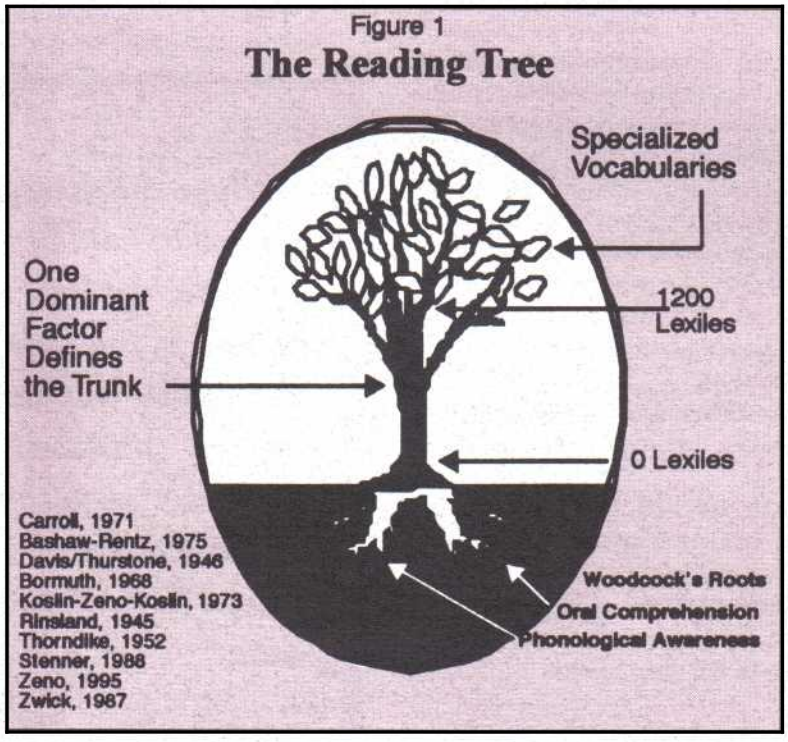
Rasch Measures Reading Best

Benjamin D. Wright
and
A. Jackson Stenner

Think of reading as the tree in **Figure 1**. It has roots like oral comprehension and phonological awareness. As reading ability grows, a trunk extends through grade school, high school, and college branching at the top into specialized vocabularies. That single trunk is longer than many realize. It grows quite straight and singular from first grade through college.

Reading has always been the most researched topic in education. There have been many studies of reading ability, large and small, local and national. When the results of these studies are reviewed, one clear picture emerges. Despite the 97 ways to test reading ability, many decades of empirical data document definitively that no researcher has been able to measure more than one kind of reading ability (Mitchell, 1985). This has proven true in spite of intense interest in discovering diversity. Consider three examples: the 1940s Davis Study, the 1970s Anchor Study, and six 1980s and 1990s ETS studies.

Figure 1
The Reading Tree



Davis - 1940s

Fred Davis went to a great deal of trouble to define and operationalize nine kinds of reading ability (1944). He made up nine different reading tests to prove the separate identities of his nine kinds. He gave his nine tests to hundreds of students, analyzed their responses to prove his thesis, and reported that he had established nine kinds of reading. But when Louis Thurstone reanalyzed Davis' data (1946), Thurstone showed conclusively that Davis had no evidence of more than one dimension of reading.

Anchor Study - 1970s

In the 1970s, worry about national literacy moved the U.S. government to finance a national Anchor Study (Jaeger, 1973). Fourteen different reading tests were administered to a great many children in order to uncover the relationships among the 14 different test scores. Millions of dollars were spent. Thousands of responses were analyzed. The final report required 15,000 pages in 30 volumes — just the kind of document one reads overnight, takes to school the next day and applies to teaching (Loret et al., 1974). In reaction to this futility, and against a great deal of proprietary resistance, Bashaw and Rentz were able to obtain a small grant to reanalyze the Anchor Study data (1975, 1977). By applying new methods for constructing

objective measurement (Wright and Stone, 1979), Bashaw and Rentz were able to show that all 14 tests used in the Anchor Study — with all their different kinds of items, item authors, and publishers — could all be calibrated onto one linear "National Reference Scale" of reading ability.

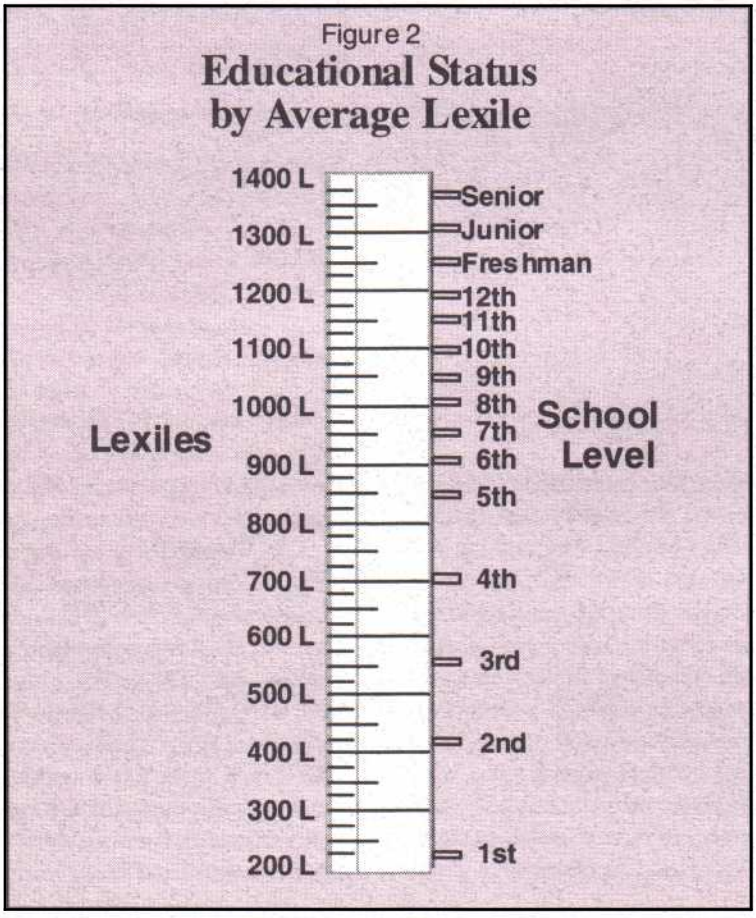
The essence of the Bashaw and Rentz results can be summarized on one easy-to-read page (1977) — a bit more useful than 15,000 pages. Their one-page summary shows how every raw score from the 14 Anchor Study reading tests can be equated to one linear National Reference Scale. Their page also shows that the scores of all 14 tests can be understood as measuring the same kind of reading on one common scale. The Bashaw and Rentz National Reference Scale is additional evidence that, so far, no more than one kind of reading ability has ever been measured. Unfortunately, their work had little effect on the course of U.S. education. The experts went right on claiming there must be more than one kind of reading — and sending teachers confusing messages as to what they were supposed to teach and how to do it.

ETS Studies - 1980s and 1990s

In the 1980s and 1990s, the Educational Testing Service (ETS) did a series of studies for the U.S. government. ETS (1990) insisted on three kinds of reading: prose reading, document reading, and quantitative reading. They built a separate test to measure each of these three kinds of reading —greatly increasing costs. Versions of these tests were administered to samples of school children, prisoners, young adults, mature adults, and senior citizens. ETS reported three reading measures for each person and claimed to have measured three kinds of reading (Kirsch & Jungeblut, 1986). But reviewers noted that, no matter which kind of reading was chosen, there were no differences in the results (Kirsch & Jungeblut, 1993, 1994; Reder, 1996; Zwick, 1987). When the relationships among reading and age and ethnicity were analyzed, whether for prose, document, or quantitative reading, all conclusions came out the same.

Figure 2

Educational Status by Average Lexile



Later, when the various sets of ETS data were reanalyzed by independent researchers, no evidence for three kinds of reading measures could be found (Bernstein, & Teng, 1989; Reder, Rock and Yamamoto, 1994; 1996; Salganik and Tal, 1989; Zwick, 1987). The correlations among ETS prose, document, and quantitative reading measures ranged from 0.89 to 0.96. Thus, once again and in spite of strong proprietary and theoretical interests in proving otherwise, nobody had succeeded in measuring more than one kind of reading ability.

Lexiles

Figure 2 is a reading ruler. Its Lexile units work just like the inches. The Lexile ruler is built out of readability theory, school practice, and educational science. The Lexile scale is an interval scale. It comes from a theoretical specification of a readability unit that corresponds to the empirical calibrations of reading test items. It is a readability ruler. And it is a reading ability ruler.

Readability formulas are built out of abstract characteristics of language. No attempt is made to identify what a word or sentence means. The idea is not new. The Athenian Bar Association used readability calculations to teach lawyers to write briefs in 400 B.C. (Chall, 1988; Zakaluk and Samuels, 1988). According to the Athenians, the ability to read a passage was not the ability to interpret what the passage was about. The ability to read was just the ability to read. Talmudic teachers who wanted to regularize their students' studies, used readability measures to divide the Torah readings into equal portions of reading difficulty in 700 A.D. (Lorge). Like the Athenians, their concern in doing this was not with what a particular Torah passage was about, but rather the extent to which passage readability burdened readers.

In the twentieth century, every imaginable structural characteristic of a passage has been tested as a potential source for a readability measure: the number of letters and syllables in a word; the number of sentences in a passage; sentence length; balances between pronouns and nouns, verbs and

prepositions (Stenner, 1997). The Lexile readability measure uses word familiarity and sentence length.

Lexile Accuracies

Table 1 lists the correlations between readability measures from the ten most studied readability equations and student responses to different types of reading test items. The columns of Table 1 report on five item types:

Lexile Slices;
SRA Passages;
Battery Test Sentences;
Mastery Test Cloze Gaps;
Peabody Test Pictures.

The item types span the range of reading comprehension items. The numbers in the table show the correlations between theoretical readability measures of item text and empirical item calibrations calculated from students' test responses. Consider the top row. The Lexile readability equation predicted

Table 1
Correlations between
Empirical & Theoretical Item Difficulties

Ten Readability Equations	Five Test Item Types				
	Lexile Slice	SRA Passage	Battery Sentence	Mastery Cloze	Peabody Picture
Lexile	.90	.92	.85	.74	.94
Flesch	.85	.94	.85	.70	.85
ARI	.85	.93	.85	.71	.85
FOG	.85	.92	.73	.75	.85
Powers	.82	.93	.83	.65	.74
Holquist	.81	.91	.81	.84	.86
Flesch-1	.79	.92	.81	.61	.69
Flesch-2	.75	.87	.70	.52	.71
Coleman	.74	.87	.75	.75	.83
Dale-Chall	.76	.92	.82	.73	.67

Adapted from Stenner, 1997

how difficult Lexile slices would be for persons taking a Lexile reading test at a correlation of 0.90, the SRA passage at 0.92, the Battery Sentence at 0.85, the Mastery Cloze at 0.74, and the Peabody Picture at 0.94 (Stenner, 1996). With the exception of the cloze items, these predictions are nearly perfect. Also note that the simple Lexile equation, based only on word familiarity and sentence length, predicts empirical item responses as well as any other readability equation— no matter how complex. **Table 1** documents, yet again that one, and only one, kind of reading is measured by these reading tests. Were that not so, the array of nearly perfect correlations could not occur. **Table 1** also shows that we can have a useful measurement of text readability and reader reading ability on a single reading ruler!

An important tool in reading education is the basal reader. The teaching sequence of basal readers records generations of practical experience with text readability and its bearing on student reading ability. **Table 2** lists the correlations between Lexile Readability and Basal Reader Order for the eleven basal readers most used in the United States. Each series is built to mark out successive units of increasing reading difficulty. Ginn has 53 units — from book 1 at the easiest to book 53 at the hardest. HBJ Eagle has 70 units. Teachers work their students through these series from start to finish. **Table 2**

Table 2
Correlations between
Basal Reader Order & Lexile Readability

Basal Reader Series	Basal Units	r	R	R'
Ginn	53	.93	.98	1.00
HBJ Eagle	70	.93	.98	1.00
SF Focus	92	.84	.99	1.00
Riverside	67	.87	.97	1.00
HM (1983)	33	.88	.96	.99
Economy	67	.86	.96	.99
SF Amer Trad	88	.85	.97	.99
HBJ Odyssey	38	.79	.97	.99
Holt	54	.87	.96	.98
HM (1986)	46	.81	.95	.97
Open Court	52	.54	.94	.97

Adapted from Stenner, 1997

r = raw R = corrected for attenuation R' = corrected for attenuation and range restriction

shows that the correlations between Lexile measures of the texts of these basal readers and their sequential positions from easy to hard are extraordinarily high. In fact, when corrected for attenuation and range restriction, these correlations approach perfection (Stenner, 1997)

Each designer of a basal reader series used their own ideas, consultants, and theory to decide what was easy and what was hard. Nevertheless, when the texts of these basal units are Lexiled, these Lexiles predict exactly where each book stands on its own reading ladder — more evidence that, despite differences among publishers and authors, all units end up benchmarking the same single dimension of reading ability.

Finally there are the ubiquitous reading ability tests administered annually to assess every student's reading ability. **Table 3** shows how well theoretical item text Lexiles predict actual readers' test performances on eight of the most popular reading tests. The second column shows how many passages from each test were Lexiled. The third column lists the item type. Once again there is a very high correlation between the difficulty of these items as calculated by the entirely abstract Lexile specification equation and the live data produced by students answering these items on reading tests. When we correct for attenuation and range restriction, the correlations are just about perfect. Only the Mastery Cloze test, well-known to be idiosyncratic, fails to conform fully.

What does this mean? Not only is only one reading ability being measured by all of these reading comprehension

READING RULER

Table 3
Correlations between
Passage Lexiles & Item Readabilities

Tests	Passages Analyzed	Item Type	r	R	R'
SRA	46	Passage	.95	.97	1.00
CAT-E	74	Passage	.91	.95	.98
CAT-C	43	Passage	.83	.93	.96
CTBS	50	Passage	.74	.92	.95
NAEP	70	Passage	.65	.92	.94
Lexile	262	Slice	.93	.95	.97
PIAT	66	Picture	.93	.94	.97
Mastery	85	Cloze	.74	.75	.77

Adapted from Stenner, 1997

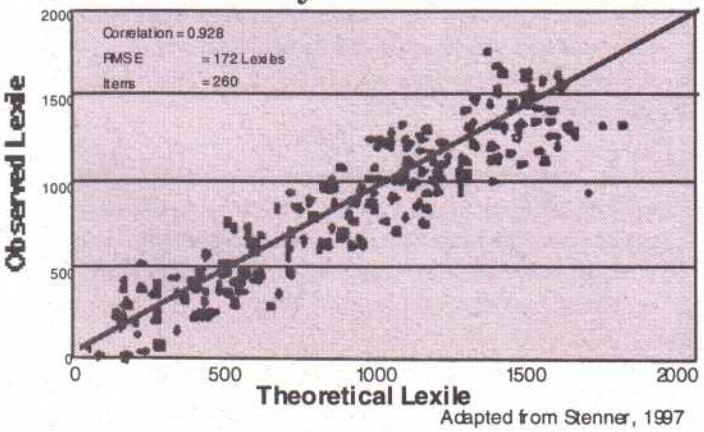
r = raw R = corrected for attenuation R' = corrected for attenuation and range restriction

tests, but we can replace all the expensive data used to calibrate these tests empirically with one formula — the abstract Lexile specification equation. We can calculate the reading difficulty of test items by Lexiling their text without administering them to a single student!

Figure 3 puts the relationship between theoretical Lexiles and observed item difficulties into perspective. The uncorrected correlation of 0.93, when disattentuated for error and corrected for range restrictions, approaches 1.00. The Lexile equation produces an almost perfect correlation between theory and practice.

Figure 3 shows the extent to which idiosyncratic variations in student responses and item response options enter the process. Where does this variation come from? Item response options have to compete with each other or they do not work. But there has to be one correct answer. Irregularity in the composition of multiple-choice options, even when they are reduced to choosing one word to fill a blank, is unavoidable. What the item writer chooses to ask about a passage and the options

Figure 3
Theory into Practice



Adapted from Stenner, 1997

they offer the test taker to choose among are not only about reading ability. They are also about personal differences among test writers.

There are also variations among test takers in alertness and motivation that disturb their performances. In view of these unavoidable contingencies, it is surprising that the correlation between Lexile theory and actual practice is so high. How does this affect the measurement of reading ability? The root mean square measurement error for a one-item test would be about 172 Lexiles. What are the implications of that much error? The distance from First Grade school books to Second Grade school books is 200 Lexiles. So we would undoubtedly be uneasy with measurement errors as large as 172 Lexiles. However, when we combine the responses to a test of 25 Lexile items, the measurement error drops to 35 Lexiles. And when we use a test of 50 Lexile items, the measurement error drops to 25 Lexiles — one-eighth of the 200 Lexile difference between First and Second Grade books. Thus, when we combine a few Lexile items into a test, we get a measure of where a reader is on the Lexile reading ability ruler, precise enough for all practical purposes. We do not plumb their depths of understanding. But we do measure their reading ability.

Sources

Bashaw, W.L. & Rentz, R.R. (1975). Equating Reading Tests with the Rasch Model, v1: Final Report & vol1 & vol2: Technical Reference Tables. Final Report of U.S. Department of Health, Education, and Welfare Contract OEC-O72-5237. Athens, GA: The University of Georgia. (ERIC Document Reproduction Nos. ED 127 330 & ED 127 331.)

Bashaw, W.L. & Rentz, R.R. (1977). The National Reference Scale for Reading: An Application of the Rasch Model. Journal of Educational Measurement, 14:161-179.

Bernstein, I.H., & Teng, G. (1989). Factoring Items and Factoring Scales are Different: Spurious Evidence for multidimensionality due to Item Categorization. Psychological Bulletin, 105(3):467-477.

Bormuth, J.R. (1966). Readability: New Approach. Reading Research Quarterly, 7:79-132.

Carroll, J.B., Davies, P. & Richmond, B. (1971). The Word Frequency Book, Boston: Houghton Mifflin,.

Campbell, A., Kirsch, I.S., & Kolstad, (1992). A. Assessing Literacy: The Framework for the National Adult Literacy Survey. Washington, DC: National Center for Education Statistics, U.S. Department of Education.

Chall, J.S. (1988). "The Beginning Years." In B.L. Zakaluk and S.J. Samuels (Eds.), Readability: Its Past, Present and Future, Newark, DE: International Reading Association.

Davis, Fred. (1944). Fundamental Factors of Comprehension in Reading, Psychometrika, 9:185-197.

Educational Testing Service. (1990). ETS Tests of Applied Literacy Skills. NY: Simon & Schuster Workplace Resources.

Jaeger, R.M. (1973). The National Test Equating Study in Reading (The Anchor Test Study). Measurement in Education, 4:1-8.

Kirsch, I.S., Jungeblut, A. & Campbell, A. The ETS Tests of Applied Literacy. Princeton, MJ: Educational Testing Service, 1991.

Kirsch, I.S., Jungeblut, A., Jenkins, L., & Kolstad, A. Adult Literacy in America: A First look at the Results of the National Adult Literacy Survey. Washington, DC: National Center for Education Statistics, U.S. Department of Education, 1993.

Kirsch, I.S., Jungeblut, A., & Mosenthal, P.B. (1994). "Moving Toward the Measurement of Adult Literacy," paper presented at the March NCES Meeting, Washington, DC, 1994.

Loret, P.G.; Seder, A.; Bianchini, JC. & Vale, C.A. (1974). Anchor Test Study Final Report: Project Report and vols. 1-30. Berkeley, CA: Educational Testing Service. (ERIC Document Nos. Ed 092 601 - ED 092 631.

Lorge, I. (1939). Predicting Reading Difficulty of Selections for

School Children. Elementary English Review, 16:229-233.

Mitchel, J.V. (1985). The Ninth Mental Measurements Yearbook. Lincoln, NE: University of Nebraska Press.

Reder, Stephen, (1996). "Dimensionality and Construct Validity of the NALS Assessment," in M.C. Smith (Ed.) Literacy for the 21st Century: Research, Policy and Practice, Greenwood Publishing, in Press.

Rentz, R.R. & Bashaw, (1975) W.L. Equating Reading Tests with the Rasch Model, v1: Final Report & vol1 & vol2: Technical Reference Tables. Final Report of U.S. Department of Health, Education, and Welfare Contract OEC-O72-5237. Athens, GA: The University of Georgia. (ERIC Document Reproduction Nos. ED 127 330 & ED 127 331.

Rentz, R.R. & Bashaw, W.L. (1977). The National Reference Scale for Reading: An Application of the Rasch Model. Journal of Educational Measurement, 14:161-179.

Rinsland, H.D. A Basic Vocabulary of Elementary School Children, 1945.

Rock, D.A., & Yamamoto, K. (1994). Construct Validity of the Adult Literacy Subscales. Princeton, NJ: Educational Testing Service.

Salganik, L.H., & Tal, J. (1989). A Review and Reanalysis of the ETS/NAEP Young Adult Literacy Survey. Washington, DC: Pelavin Associates.

Stenner, A.J., & Burdick, D.S. (1997) "The Objective Measurement of Reading Comprehension: In Response to Technical Questions Raised by the California Department of Education Technical Study Group." Durham, NC: Metametrics.

Thorndike, E. L. & Lorge, I. (1952). The Teacher's Word Book of 30,000 Word.

Thurstone, L.L. (1946). "Note on a Reanalysis of Davis' Reading Tests," Psychometrika, v11, n2, 185ff.

Woodcock, R.W., Woodcock (1974). Reading Mastery Tests. Circle Pines, MN: American Guidance Service.

Wright, B.D., & Stone, M. H. (1979). Best Test Design. Chicago: MESA Press.

Zakaluk and S.J. Samuels (Eds.) (1988). Readability: Its Past, Present and Future, Newark, DE: International Reading Association.

Zeno, S.M., Ivens, S.H., Millard, R.T. & Davvuri, Raj. (1995). The Educators Word Frequency Guide, Touchstone.

Zwick, R. (1987). Assessing the Dimensionality of the NAEP Reading Data, Journal of Educational Measurement, 24:293-308.

**The authors are grateful to Ed Bouchard for helping with this report.*

A. Jackson Stenner, Ph.D.

Jack Stenner is co-founder and Chairman of MetaMetrics, Inc. MetaMetrics is a privately held corporation that specializes in research and development in the field of education. He has been Principal Investigator on five grants from the National Institute of Health, (1984-1996) dealing with the measurement of literacy.

Jack Stenner is also former Chairman and co-founder of National Technology Group, a 700-person firm specializing in computer networking and systems integration which was sold to VanStar Corporation in December 1996. He holds a Ph.D degree from Duke University and Bachelor degrees in Psychology and Education from the University of Missouri.

Jack is President of the Institute for Objective Measurement in Chicago, Illinois. He serves as a board member for The National Institute for Statistical Sciences (NISS) and is Immediate Past President of the Professional Billiard Tour Association (PBTA).

Jack resides in Chapel Hill, North Carolina with his wife, Jennifer, and their four sons.