Does the Reader Comprehend the Text Because the Reader Is Able or Because the Text Is Easy?

A. Jackson Stenner

MetaMetrics, Inc.

Mark H. Stone

Adler School of Professional Psychology

Text measure: 1500L

Presented at

International Reading Association

Reno–Tahoe, Nevada

May 4, 2004

For correspondence:
A. Jackson Stenner, MetaMetrics, Inc.
1000 Park Forty Plaza Drive, Suite 120
Durham, NC 27713 USA
919-547-3402, jstenner@lexile.com

Abstract

Does the reader comprehend the text because the reader is able or because the text is easy? Localizing the cause of comprehension in either the reader or the text is fraught with contradictions. A proposed solution models comprehension as the difference between reader ability and text readability. Computing such a difference requires that reader ability and text readability be measured on a common scale. Thus, the puzzle is solved by positing a single continuum along which texts and readers can be conjointly ordered. A reader's comprehension of a text is a function of the difference between reader ability and text readability. This solution forces recognition that generalizations about reader performance can be text independent (reader ability) or text dependent (comprehension). The article explores how reader ability and text readability can be measured on a single continuum, and the implications that this formulation holds for reading theory, the teaching of reading, and the testing of reading.

Introduction

A *koan* is a Zen riddle designed to provoke meditation and open the mind to new ways of thinking and understanding. We begin by offering a koan: Does the reader comprehend the text because the reader is able or because the text is easy? To seek an answer to this riddle, we must question how we think about the relative contributions of reader and text to comprehension.

One possibility, which was popular for most of the last century up to about 25 years ago, is the "behaviorist" argument that the text is preeminent in dictating the meaning that a reader takes from a reading. To a behaviorist, readers are exchangeable, and what matters most is the schedule of challenges and rewards that surrounds the exercise of reading. A second possibility, which has been the dominant approach in reading research in the last few decades, is the "constructivist" argument that the reader's prior knowledge and expectations shape the meaning that is constructed from a reading, so that comprehension lies in the mind of the reader. A third possibility is that the riddle may be a trap, in the sense that reader and text measurement are incommensurable and so their contributions to reading comprehension cannot be usefully compared.

It is characteristic of riddles that they only rarely can be solved by adhering to the rules. The reading riddle asks for an either-or choice; that is, does comprehension result from reader ability *or* text readability? Our answer to the riddle is to break out of the either-or choice, and to answer that *both* reader and text share in accounting for comprehension. We begin this paper by arguing that much of contemporary research on reading, in which reading comprehension is placed in the mind of the reader, is based on

a fundamental conceptual flaw that makes the results open to multiple interpretations. We will then argue that reader ability and text readability can be measured on a single numerical scale, called Lexiles. We introduce and discuss the Lexile Framework and both practical and conceptual aspects of how it measures text readability, reader ability, and comprehension. Using this framework, reader comprehension is determined by the gap between reader ability and text readability. Using both reader ability and text readability to explain reading comprehension leads to new perspectives for research on reading, the teaching of reading, and reading assessment.

The Problem of Placing Reading Comprehension in the Mind of the Reader

For the past quarter century, much reading research has ignored text when accounting for what happens when a reader engages a text. Instead, much contemporary reading research and theory attempts to solve the "reading riddle" by placing reading comprehension in the mind of the reader. From this "constructivist" perspective, meaning is constructed in the reader's head, and because readers vary greatly in their background knowledge, a single text evokes such varied meanings that any universal representation of text (which, say, a measurement of the readability of the text necessarily would be) is a chimera. Constructivists argue that text is needed to trigger a reader reaction, but what the reader brings to an encounter with text is the dominant explanatory mechanism in accounting for what the reader comprehends.

In this view, it becomes unnecessary even to contemplate measuring the readability of text. This perspective explains why approximately 3,000 pages of reading

theory and research spread across three volumes and two decades in the three *Handbooks of Reading Research* (Barr, Kamil, & Mosenthal, 1984; Barr, Kamil, Mosenthal, & Pearson, 1996; Kamil, Mosenthal, Pearson, & Barr, 2000) can manage with less than a handful of references to text readability. Similarly, the Rand Corporation published a study about framing future research on reading comprehension that includes one disparaging reference to readability (Snow, 2002): although "text" is frequently mentioned, its measurement is not. In the modern vernacular, "text is not where it's at" (Hiebert, 2004).

A weaker version of this ontology holds that although text measurement can be helpful in some circumstances, this admission should not mask the reality that the reader is the dominant actor in reading comprehension. In this version, the role of text is not openly dismissed, but rather quietly neglected. This strategy of quiet neglect has been rhetorically quite successful. After all, an open assertion that a well-researched construct with a century-long tradition, like the readability of text, should simply be dismissed from modern reading theory and practice would have required assuming a heavy intellectual burden of explanation and justification. Simple neglect, on the other hand, leaves the unwary with the incorrect notion that, in human science research, constructs come and go, and the time for looking at readability of text has passed.

However, choosing reader over text, as many contemporary reading theorists have done, leads to certain confusions (Stenner & Wright, 2002). To understand the implications of this problem for reading research, suppose an investigator is interested in the relationship between reading and summarizing. Data on 100 fourth graders' reading ability is collected by administering a grade-appropriate reading achievement test. After

examinees read and answer written questions, they are asked to summarize orally the content of the passage. These oral summaries are evaluated for quality by five raters. A correlation is computed between the reading scale scores and the summary measure, and the researchers discover a relatively high correlation of 0.7 (where 0 represents no correlation and 1.0 represents a perfect correlation). The investigators conclude, “Good readers possess summarization skills that are much better developed than those of poor readers.” Thousands of studies using this general design have allowed researchers to estimate statistical relationships between reading performance and dozens of so-called reading process correlates (Pressley, 2000).

But this research design suffers from a fatal flaw. There are two conceptually distinct reader behaviors: reading ability, which does not depend on the text being read, and reading comprehension, which does depend on the text being read. In any study where all examinees respond to the same text, those examinees with high measured reading ability also experience high comprehension rates, and those examinees with low reading ability experience correspondingly low comprehension rates. Thus, those who are identified as “good” readers in this experimental design are both readers with high reading ability and readers who enjoy (with this text) high comprehension rates. So-called “poor” readers in this experimental design have both low reading ability, relative to the other fourth graders in the study, and experience low comprehension rates.

Even after finding a strong positive correlation between the written reading test and the oral summary, this research design cannot conclude whether the oral summarization performance is correlated with reading ability or comprehension rate. Perhaps persons with high reading ability summarize well regardless of what they read,

or perhaps when readers engage text that they comprehend well – regardless of their reading ability -- they summarize well. Conversely, perhaps those with low reading ability have difficulty summarizing, or perhaps when anyone engages text that they comprehend poorly, they will summarize poorly, whatever their reading ability. In this latter interpretation, summarization performance results from reading comprehension and only appears to be correlated with reading ability (in the hypothetical study) because the research design confounds text-dependent performance and text-independent performance.

Thousands of studies have sought to contrast “good” or “mature” readers on the one hand and “poor” or “immature” readers on the other. Usually, the research report makes a claim about what good readers can do that poor readers cannot, but it is often not clear to what kind of reading behavior the epithets *good* and *poor* refer. Two extended quotations below summarize the reader characteristics that differentiate *good, skilled,* and mature readers on the one hand and *poor*, *less skilled,* and *immature* readers on the other. The first quotation is from Daneman (1996); the second is from (Pressley, 2000). Each quotation includes at least a dozen generalizations about good vs. poor readers. Consider which of these generalizations are text-independent claims about reading ability and which generalizations are text-dependent claims about reading comprehension – or if it is even possible to tell which argument is being made.

> Relatively few studies have examined exactly which of the high-level processes shared by reading and listening are responsible for the individual differences. However, from the little evidence we do have, a consistent picture seems to

emerge. Poor readers are at a particular disadvantage when they have to execute a process that requires them to integrate newly encountered information with information encountered earlier in the text or retrieved from semantic memory. So, for example, poor readers have problems interrelating successive topics (Lorch, Lorch, & Morgan, 1987) and integrating information to derive the overall gist or main theme of a passage (Daneman & Carpenter, 1980; Oakhill, 1982; Palincsar & Brown, 1984; Smiley, Oakley, Worthen, Campione, & Brown, 1977). They have more difficulty making inferences (Masson & Miller, 1983; Oakhill & Yuill, 1986) and tend to make fewer of them during text comprehension (Oakhill, 1982). Poor readers also have more difficulty computing the referent for a pronoun (Daneman & Carpenter, 1980; Oakhill & Yuill, 1986). Other researchers have found that poor readers do not demand informational coherence and consistency in a text, and often fail to detect, let alone repair, semantic inconsistencies (Garner, 1980).

Good readers are extremely active as they read, as is apparent whenever excellent adult readers are asked to think aloud as they go through text (Pressley & Afflerback, 1995). Good readers are aware of why they are reading a text, gain an overview of the text before reading, make predictions about the upcoming text, read selectively based on their overview, associate ideas in text to what they already know, note whether their predictions and expectations about text content are being met, revise their prior knowledge when compelling new ideas conflicting with prior knowledge are encountered, figure out the meanings of

> unfamiliar vocabulary based on context clues, underline and reread and make notes and paraphrase to remember important points, interpret the text, evaluate its quality, review important points as they conclude reading, and think about how ideas encountered in the text might be used in the future. Young and less skilled readers, in contrast, exhibit a lack of such activity (e.g. Cordón & Day, 1996).

Neither quotation contains any reference to text measurement, so the presumption appears to be that claims about good and bad readers are independent of the text. But even a reader with a fairly high reading ability might do a poor job of comprehending, say, the relatively difficult text of a Supreme Court decision. Conversely, even a reader of fairly low ability may provide a rich summary of a children's story like *Frog and Toad Are Friends* (Lobel, 1970).

To the unwary, a research design in which all participants have the same text may seem like a perfect way to ensure that text plays no role in drawing inferences about good and bad readers, but this approach is deeply misguided. In the attempt to focus on only the reader, rather than embracing the duality of reading ability and text readability, comprehension rate and reading ability are hopelessly confounded. We are left with the current predicament: thousands of studies that calculate correlations between a reading test performance and some process that may be connected to reading, and no way at all to determine which of these correlations support text-dependent generalizations about reading comprehension and which support text-independent generalizations about reading ability.

The Lexile Framework, discussed in the next section, will separate reading ability, text readability, and reading comprehension. Thus, it requires that claims about "good" and "poor" readers must be qualified as to whether the discussion is about reading ability or reading comprehension, and further requires that claims about "easy" or "hard" texts must be qualified as to whether they refer to the text in isolation from readers or whether they refer to a particular sample of readers having difficulty comprehending the text (Stenner, Horabin, Smith & Smith, 1988; Stenner & Stone, 2003; Stone, Wright & Stenner, 1999; Wright & Stone, 2004).

A Lexile Framework Primer

When a reader confronts a text, is required to carry out some task in response to that text, and the reader's response is subsequently rated by some mechanism, the situation may appear so amorphous and complex that it is easy to despair that any simple model can account for and measure what happens. Yet all science progresses by inventing workable simplifications of complex reality. The Lexile Framework for Reading purports to measure in a common unit, called Lexiles, the traits of reader ability and text readability. Based on these measures, reading comprehension can be calculated based on the gap between reader ability and text readability. When reader ability far exceeds text readability, then comprehension should approach unity. Conversely, when text measure far exceeds reader measure, then the probability of little or no comprehension should approach unity.

The way of formulating the connection from reader ability and text readability to reader comprehension raises a number of concerns, which will be discussed in this section: 1) the readability of text must be measured; 2) reading ability must be measured, which given the readability of the text will lead to a prediction about reading comprehension; 3) the conceptual framework and assumptions that allow a researcher to subtract text readability from reader ability and calculate a comprehension rate must be spelled out; 4) the approach should be flexible enough to include other aspects of the reading environment like subjective assessments of reading and different tasks for measuring reading. This section will discuss how the Lexile framework addresses these concerns.

Measuring the Readability of Text

All systems of communication through symbols share two features: a semantic component and a syntactic component. In language, the semantic units are words, which are organized according to rules of syntax into thought units and sentences (Carver, 1974). Thus, the readability of text is governed largely by the familiarity of the words and by the complexity of the syntactic structures used in constructing the message.

Regarding the semantic component, most methods of measuring difficulty are proxies for the probability that a person will encounter a word in context and infer its meaning (Bormuth, 1966). “Exposure theory,” which is the main explanation for the development of receptive or hearing vocabulary, is based on how often people are exposed to words (Miller & Gildea, 1987; Stenner, Smith, & Burdick, 1983). Knowing the frequency of words as they are used in written and oral communication provides the

best means of inferring the likelihood that a word will be encountered and become a part of an individual's receptive vocabulary. Klare (1963) built the case for the semantic component varying along a continuum of familiarity to rarity, a concept that was further developed by Carroll, Davies, and Richman (1971), whose word-frequency study examined the reoccurrences of words in a five-million-word corpus of running text.

Sentence length is a powerful proxy for the syntactic complexity of a passage. However, sentence length is not the underlying causal influence (Chall, 1988). Davidson and Kantor (1982), for example, illustrated that reducing sentence length can increase difficulty and vice versa. The underlying factor in the difficulty of syntax probably involves the load placed on short-term memory, and a body of evidence suggests that sentence length is a good proxy for the demands of structural complexity on verbal short-term memory (Klare, 1963; Crain and Shankweiler, 1988; Shankweiler and Crain, 1986; Liberman, Mann, Shankweiler, and Werfelman, 1982).

The Lexile equation combines into an algebraic equation the measurements of word frequency and sentence length drawn from a certain text, and produces a measure in Lexiles. A simple children's book would have a Lexile score of about 200; a complex and specialized work might have a Lexile score of 1700 or more. The body of text used for measuring the likelihood of encountering particular words makes use of a 550-million-word corpus, comprising the full text for more than 30,000 books. The Lexile Analyzer—a software program for measuring text readability—is freely available for noncommercial use at <http://www.Lexile.com>. Also, the Lexile website contains a search program that allows you to insert book titles and see their Lexile scores, or to insert a range of Lexile scores and see a list of book titles in that range.

Measuring Reader Ability and Comprehension

Reader ability is defined in reference to the text readability – although not in reference to any particular tests. For example, *Harry Potter* has a text readability of 880L. Imagine that the book was turned into a test with 1000 questions, each question appearing in a standard Lexile form called the "native form," illustrated in Figure 1. We then say that a reader with the ability to answer correctly 750 of the 1000 items correctly has a Lexile reading ability of 880L.

An important conceptual insight of the Lexile framework is that text itself could be viewed as a virtual test made up of standard-sized "native form" 125-word passages. (If the 125th word is not the end of a sentence, the passage is extended until it reaches the sentence-ending punctuation.) The degree of difficulty of each passage can be calibrated using the Lexile equation that takes words and syntax into account. Thus, this approach produces a measure of readability expressed in the same metric used to express reader ability.

Choosing one type of reading task, like the "native item" format, does not cause a loss of generality. Any reading behavior that systematically varies as a function of the reader/text difference can in principle be used to measure reading ability; for example, the accuracy with which words are read aloud from text can be used as a measure of reading. As we will discuss in a later section, the Lexile approach can allow making comparisons across reading tasks, so that the results from different tasks can still be measured in Lexiles.

***Figure 1*. An example Lexile test item.**

Wilbur likes Charlotte better and better each day. Her campaign against insects seemed sensible and useful. Hardly anybody around the farm had a good word to say for a fly. Flies spent their time pestering others. The cows hated them. The horses hated them. The sheep loathed them. Mr. And Mrs. Zuckerman were always complaining about them, and putting up screens. **Everyone _____ about them.**

A. agreed
B. gathered
C. laughed
D. learned

From *Charlotte's Web* by E. B. White, 1952, New York: Harper & Row.

A passage of text can be understood as entailing a large number of propositions, each of which can form the basis for making a "native item" reading test question. The set of all such allowable items for a passage is the "ensemble." Each individual test item in the ensemble can be imagined to have an observed difficulty, and the average over this distribution of difficulties is the ensemble mean. The Lexile Theory claims that these ensemble means are predictable from knowledge of the semantic and syntactic features of text passages (Stenner, Burdick, Sanford and Burdick, 2004). Thus, the Lexile Theory replaces statements about individual reading item difficulties with statements about ensembles (KAC, 1959). Because text readability is a property of the entire text passage and not a characteristic of any particular test item, the ensemble interpretation produces the right level of aggregation for the empirical variation that the Lexile Theory attempts to explain.

The choice of a 75% comprehension rate on the virtual test items is arbitrary, but highly useful. At a conceptual level, this thought experiment of breaking a text into virtual test questions enables combining a substantive theory of text difficulty, based on

word familiarity and syntax, with a psychometric model for reader ability. At a practical level, 75% comprehension represents a balance between reading skill and difficulty of text that allows the reading experience to be both successful and challenging.

When given a measure of text readability, it is often useful to imagine what reader characteristics match this text readability; for example, a text at 880L like Harry Potter would match the 880L reading ability of a fiftieth percentile sixth grader. Similarly, when given a reader measure (1200L), imagine the texts that this reader can read with 75% comprehension; someone with reading ability of 1200L, for example, could read *Cold Mountain (*1210L) or *The Trumpeter of Krakow (*1200L) with 75% comprehension.

The gap between reader ability and text difficulty will determine the extent of comprehension, as measured by the proportion of correct answers on the reading test. For example, a 1200L reader can read text such as in *USA Today* (1200L) with 75% comprehension and would have 92% comprehension of *Harry Potter* (880L) but would have only 60% comprehension of the typical College Board SAT text (1330L). Figure 2 shows the expected level of comprehension for a reader with ability of 750L with texts at different levels of readability.

Figure 2.

Expected comprehension rates for a 750L reader reading books with varying text measures.

Reader Ability	Text Readability	Title	Expected Comprehension Rate
750L	250L	*Play Ball Amelia Bedelia*	95%
750L	500L	*Harold and The Purple Crayon*	90%
750L	750L	*The Adventures of Pinocchio*	75%
750L	1000L	*Island of The Blue Dolphins*	50%
750L	1250L	*The Midwife's Apprentice*	25%

Finally, this framework implies that differences in text readability can be traded off for differences in reader ability to hold comprehension constant, which as we shall see is a protean concept with important implications for reading theory and practice.

The Conceptual Framework

Perhaps the preeminent idea underlying the Lexile Framework is the measurement of reading ability and text readability on a common scale, followed closely by the realization that text and task used to evaluate reading must be conceptually separated if the measurement of reading is to be unified (Stenner, Smith, & Burdick, 1983; Stenner, Horabin, Smith, & Smith, 1988; Wright & Stone, 1979). This movement from measures

of reading ability and text readability to a measure of reading comprehension rests on a solid and well-developed conceptual framework.

This conceptual framework is based on four key assumptions. The first assumption is *sufficiency*, which means that the number of correct answers on a reading test contains all of the information in the response record that is informative about a reader's ability (Fisher, 1922; Andersen, 1977). In particular, sufficiency means that which specific items the reader got correct or incorrect contributes no information about the reader's ability beyond what is given by the total count correct. In addition, the sufficiency assumption means that the measure of text readability summarizes everything about the text that is important in accounting for comprehension.

The second assumption is *separability*, which means reader ability and text readability have separate effects that can be isolated from one another (Rasch, 1960 [1980]). The third assumption, *specific objectivity*, means that reader measures can be estimated independently of the particular text used, the particular form of the response (for example, whether the response is an essay or some other response), and the method used to rate the essay or constructed response (Rasch, 1960 [1980]).

A final assumption of *latent additivity* means that the measures of reader ability and text readability, when related to reading comprehension, are connected to one another by addition or subtraction (Luce, Tuckey, 1964). This represents a test of the quantitative hypothesis that the contruct being measured adds up in the same way that the numbers representing it do. (Michell, 1999). An important implication of latent additivity is that reader ability and text readability are measured on a common scale and differences between them can be traded off to hold success rate constant. Lest the assumption of an

additive representation seem overly restrictive, we note that an additive representation can be transformed into a multiplicative representation and vice versa quite easily (by using logarithms). Thus, the assumption of an additive form actually includes the possibility of a multiplicative relationship, suitably transformed. Because it is often easier to conceptualize addition (or subtraction) than multiplication (or division), from this point on we will stipulate that reader ability and text readability are related through subtraction and have equal potential in affecting the outcome of an encounter between reader and text. Thus:

reader ability - text readability = probability of comprehending the text.

Only one model meets the four assumptions set forth above, and thus allows us to subtract text readability from reader ability to determine the probability of comprehension. It is one of a family of models named after the Danish statistician, George Rasch (1960 [1980]). The Rasch model is a mathematical function that relates the probability of a correct response on an item/text to the difference between one reader parameter (in this case, reader ability) and one text parameter (in this case, text readability). This probability may be interpreted as a comprehension rate for the item/text. Reader comprehension for a multi-item passage is a function of the sum of these modeled probabilities of a correct answer. A high reader ability relative to the text readability produces a high probability of a correct response to a test item – that is, a high comprehension rate. Conversely, a low reader ability relative to a high text readability results in a low-modeled probability of a correct answer. In this latter case when the

probabilities associated with a number of these reader/paragraph encounters are summed, the result is a low forecasted comprehension rate. This probabilistic perspective thus admits reasonable uncertainty regarding what may happen with a particular reader, reading a particular text on a particular occasion, while focusing attention on average performances of readers of a particular ability, on typical occasions, reading texts of a particular difficulty.

The Rasch model states a set of requirements for the way observations and theory combine in a probability model to make meaningful measures. The Rasch Model combines the three components in any definition of measurement -- observation, theory, and measure -- into a simple, elegant, and, in some important respects, unique representation (Wright & Stone, 2004).

This algebraic framework will also allow for testing the assumptions behind the idea that reader comprehension can be modeled as the gap between reader ability and text readability. For example, it will test whether differences in text readability can be traded off against differences in reader ability to keep comprehension constant. These kinds of trade-offs or invariance's are only possible if additivity obtains.

Text readabilities computed using up-to-date technology like the Lexile method are highly reliable (Stenner, Burdick, Sanford, & Burdick, 2004). Indeed, uncertainty over text readability can be effectively ignored in many applications of the Lexile Framework. Uncertainty in reader abilities, as opposed to text readabilities, is by far the major source of error in forecasted comprehension.

The combination of the Lexile Framework and the Rasch model is thus not a matter of collecting a body of data and then trying "to model the data" with strategies like

fitting various functional forms, inserting different variables, and trying out different interaction terms between variables. Instead, the Lexile Framework in cooperation with the Rasch model set forth a set of requirements that data must meet if the data are to be useful in constructing meaningful reader measures. If the data do not meet these requirements, then the appropriate response is to question the way that the observations were made and ask what contaminants to the observation process might have introduced unintended dependencies in the data.

The overall goal of this conceptual framework is to create measures of reading ability and text readability that transcend the initial conditions of measurement. This standard holds true for most measurement procedures in use in the so-called "hard" sciences – for example, thermometers for measuring temperature. The way is now clear to achieve this standard in the measurement of reading. Measures of reading should provide the same result, even if they are carried out using different methods, just like temperature measurements using Fahrenheit or Celsius scales, or using different types of physical thermometers, provide a common result. By using the Lexile framework, we can leave behind the particulars of the method and moment of measurement as we move forward to use the measure in description and prediction of reader behavior.

Different Tasks and Subjective Raters

Along with measuring text readability and reader ability, a full version of the Lexile Framework will also address two other important aspects. A full model of reading would include not only reading ability and text difficulty, but also whether the reading task involved something other than the "native form," and whether a computer based or

human rating system was employed to judge a readers performance. For example, observations on readers might include the read-aloud accuracy rate, words read correctly per minute, number of correct answers to multiple choice questions, quality rating of summary, and retelling. These tasks are all different from the 125-word “native form” at the basis of the Lexile framework. However, these different task types can be shown to measure the same reading ability as is measured by traditional reading tests and therefore, can be rescaled in Lexiles (Stenner and Wright, 2002).

In its simplest manifestation, which has been the focus up to this point, the Lexile approach uses a single multiple-choice task type, thus eliminating the need for a rater/observer parameter. Moreover, the task type can be restricted to a basic common format, the so-called “native item” format illustrated earlier. Thus, in its simple form, the Lexile Framework reduces to a measure of reader ability and a measure of text readability. But the Lexile framework can be used to adjust for the severity of subjective raters or observers and to adjust for the difficulty of the tasks demanded across different tests.

An expanded version of the Rasch model enables the measurement of four facets of reading -- readers, texts, tasks, and raters -- on a common scale (Linacre, 1987). If data fit the multi-faceted Rasch model, so that the four key assumptions discussed in the previous section are met, then 1) each facet can be estimated independently of the other facets; 2) the facets can be added to each other; 3) differences in measurements on one facet can be traded off for equal differences on other facet(s) to hold constant the probability of reading comprehension. In this case, the many-facet Rasch model enables

the estimation of measures of reader ability, text readability, the reading task, and the subjectivity of the rater all expressed in equal interval Lexile measures.

For an intuitive sense of how such conversions might work, note that it is straightforward to describe how performance on other tasks will correspond to a particular comprehension rate on a "native form" Lexile test. For example, perhaps a 75% comprehension rate on a "native form" Lexile test represents performance equal to 98% word call accuracy in a read-aloud study. There are probably dozens of task types that can be ordered on a "task continuum" for reading, all measuring the same reading ability construct but doing it with added easiness or hardness relative to the native item format. Once the demands of a certain task type have been located on the task continuum, then test results using this type of task can be converted into Lexiles (Wright & Stone, 1979; Wright & Masters, 1982).

There are numerous philosophical, mathematical, and practical implications of the way that substantive reading theory and the many-facet Rasch Model are combined in the Lexile Framework for Reading. The reader interested in more extensive treatments might begin with Linacre (1989) and Boomsma, van Duijn, and Snigders (2001).

Implications for Reading Research

The last 50 years of reading research has amassed a staggering array of studies, which seek to correlate some measure of reading performance with some "process variable," which is a putative cause or effect of reading performance. These relationships are sometimes presented as correlations between variables and sometimes presented as

differences in the averages between groups of “good” and “poor” readers. However, results presented in either form can be re-expressed in the other form, and so in this section, we will for convenience refer only to correlations.

About the process variables that are more or less correlated with reading performance, Stanovich (1986) asked an embarrassingly simple question: “Is there evidence that the correlate is in fact a cause (worthy of instructional focus), or is it in fact an effect or consequence of reading performance?” After all, if the process variable causes better reading performance, then it might be worth spending precious instructional time and money on trying to use that variable to improve reading performance, but if the variable is a result of better reading performance, then while it may serve as a marker of pedagogical success, it cannot be used to improve reading performance.

The previous discussion has argued that there are two kinds of reading performance: “reading ability,” which is a reader trait that is independent of text, and “comprehension,” which is modeled as a function of the difference between reader ability and text readability. These two kinds of reading performance can be crossed with four kinds of relationships: 1) The process variable causes one of the two reading performances; 2) The process variable is caused by (is a consequence of or is an effect of) one of the two reading performances; 3) There is a reciprocal causal relationship between the process variable and one of the two reading performances; or 4) An observed correlation between the process variable and reading performance is spurious (perhaps due to a third variable that is causing both of them).

Figure 3 juxtaposes the two kinds of reading performance -- reading ability and comprehension -- with three kinds of relationship -- cause, effect, and reciprocal

causation. (Spurious relationships are not considered further in this discussion.) The candidate variables listed in this table should be taken as provisional. We have not conducted a proper review of the literature to solidify their placement in the classification. Thus, we offer these candidate variables to provoke thought and, perhaps, to inflict some gentle bruises upon prevailing intuitions.

Type 1 relationships are candidate causes of reading ability. The first mechanism suggests that readers cannot improve their reading if they cannot access text (Krashen, 2002). A second candidate cause of reading ability is that exposure to text that is well matched to the reader's ability promotes faster development than repeated exposure to text that is too hard or too easy (Carver, 2000).

Type 2 relationships include effects of reading ability. For example, continuing in college is a difficult prospect for an 800L reader, because college texts have readabilities in the 1200L–1400L range, which results in an expected comprehension rate approaching 25% (Williamson, 2004). In this scenario, a decision to forego college is an effect of low reading ability. Many correlates of reading ability that have been proposed as causes are probably effects of comprehension (Stanovich, 1986).

Type 3 relationships share a reciprocal causation with reading ability. Vocabulary grows as a result of listening up to about 600L–700L, and then reading becomes a major avenue for vocabulary growth. However, because vocabulary (semantic facility) is one of the two key variables in the Lexile equations, it is viewed as causal for reading ability. A virtuous circle characterizes high ability readers: they read more so their vocabulary grows and because their vocabulary grows, they can comprehend increasingly difficult texts.

Type 4 relationships include those variables that, when experimentally manipulated, change the reader's comprehension. The Lexile theory asserts that the "match" between reader ability and text readability is the major cause of comprehension.

Type 5 relationships are probably vast in number. Many constructs have been proposed as causes of reading performance that are, in fact, effects of reader comprehension. As one example, Stanovich (1986, p. 365) reviewed the evolving view in the research on eye movements as a cause of reading performance:

> The relationship of certain eye movement patterns to reading fluency has repeatedly, and erroneously, been interpreted as indicating that reading ability was determined by the efficiency of the eye movements themselves. For example, researchers have repeatedly found that less skilled readers make more regressive eye movements, make more fixations per line of text, and have longer fixation durations than skilled readers (Rayner, 1985a, 1985b). The assumption that these particular eye movement characteristics were a cause of reading disability led to the now thoroughly discredited "eye movement training" programs that repeatedly have been advanced as "cures" for reading disabilities. Of course, we now recognize that eye movement patterns represent a perfect example of a causal connection running in the opposite direction. Poor readers do show the inefficient characteristics listed above; but they also comprehend text more poorly. In fact, we now know that eye movements rather closely reflect the efficiency of ongoing reading —with the number of regressions and fixations per line increasing as the material becomes more difficult, and decreasing as reading efficiency increases

(Aman & Singh, 1983; Just & Carpenter, 1980; Olson, Kliegal, & Davidson, 1983; Rayner, 1978, 1985a, 1985b; Stanley, Smith, & Howell, 1983; Tinker, 1958)—and this is true for all readers, regardless of their skill level.

In short, “eye movement” migrated from what we are calling a Type 1 relationship to a Type 5 relationship -- that is, from a cause of reading ability to an effect of reading comprehension. We wonder how many of the currently popular “reading process variables” or “comprehension processes” are in fact consequences of reader comprehension rates.

We should avoid, however, being too quick to dismiss all process variables that are correlated with reading comprehension as unimportant if they are not causal. In some cases, effects of reading comprehension may be important outcomes in their own right. For example, perhaps manipulations of reading fluency are not a useful way to cause greater comprehension, but gains in comprehension can help to encourage greater reading speed – and reading speed (or fluency) may be important in its own right. Given two prospective employees both reading at 1300L, the employer may choose the one who reads at 250 wpm over one who reads at 175 wpm.

Finally, Type 6 process variables exhibit reciprocal causation with reading comprehension. A reader who is motivated to read about basketball will comprehend more and, because of the comprehension, will be further motivated to read. Again, a virtuous circle on a shorter time scale develops.

The classification scheme in Figure 3 has a number of useful applications. It emphasizes that reading research should always specify whether reading ability or

comprehension is under study, and should attempt to specify the causal status of each process variable. This matrix of causes and consequences of reading ability and comprehension can be used to consider instructional implications of research findings. Also, the classification scheme may prove useful in carrying out meta-analyses that attempt to summarize the results of many studies in the reading literature, by helping such analyses avoid the "apples-and-oranges criticism" that they are jumbling together causes and effects for different reading performances.

Figure 3.

A Framework for Thinking About Reading Performance, Its Correlates, and Causations

	Causes (Type 1)	Effects (Type 2)	Reciprocal Causation (Type 3)
Trait: Reading Ability (text independent)	Accessibility of text Targeting text on reader ability (long-term)	Last grade completed Many process variables Life Earnings	Vocabulary Amount of text read Syntactic facility
State:	(Type 4)	(Type 5)	(Type 6)
Reading Comprehension (text dependent)	Reader ability minus text readability Rereading	Eye movement Retelling Inferencing Many process variables: Accuracy, fluency, etc.	Engagement Motivation Distractibility

Implications for Reading Instruction

Reading research has more than academic implications. Its results play out in the development of reading programs and technologies and influence how teachers view the reading process, both of which influence how reading is taught. Indeed, today's top-selling reading technologies each include one or more process variables that were originally developed in academic research before finding their way from the researcher's lab to the publisher's boardroom.

But there are also more subtle examples of how research on reading ability, text readability, and comprehension might directly affect pedagogy. For example, the dominant instructional model used in U.S. middle school through college courses, with the exception of some classes that use laboratories or case studies, relies on core textbooks. All students use the same textbook, regardless of their reading ability. However, in a typical classroom, student reading abilities vary over an 800L to 900L range, which implies the stunning conclusion that comprehension rates vary within a class from less than 25% to more than 95%. Thus, many students cannot access the majority of content in the textbook because their comprehension rates are too low.

If students were better targeted to textbook material according to their reading ability, would their content knowledge increase? Krashen (2000, p. 30) argues the skeptical view that "there is no evidence at all for the 'targeting hypothesis' as a cause of non-reading," and states, "In fact, what little data there is on this issue suggests that matching for reading level is not the problem." Carver (2000), on the other hand, devoted a chapter in his book, *The Causes of High and Low Reading Achievement*, to reviewing

the research that the “match” – that is, that reader ability matches text readability -- is causally implicated in a wide range of important behaviors, not the least of which is the improvement of reading ability itself. If the “match” is, in fact, causally implicated in promoting growth in reading ability and also a controlling influence in how much science, social studies, health, and so on a student learns in different classes, then the one-textbook-fits-all instructional philosophy may need to be revisited.

More broadly, the prospect of linking reading test scores to particular books (or more generically, connecting readers to text) has been the single most compelling application that has sustained the unification of reading. Teachers and parents want reading test scores to be more actionable and refrigerator-friendly. It is empowering for teachers, students, and parents to be able to forecast the success that a reader will have with a text.

Implications for Reading Assessment

Reading assessments may be the most common tests in education, so a change in perspective could have substantial consequences for test theory and practice.

The evidence seems overwhelming that we can usefully treat reading ability, text readability, and comprehension as if they are one-dimensional constructs. The strongest support for such a treatment comes from the fact that when using the Lexile measures, differences between two reader ability measures can be traded off for an equivalent difference in two text readability measures, while holding comprehension constant. Consequently, there is no justification for reporting separate components of reading

ability, other than the semantic and syntactic facility measures that are implicit within the Lexile measures.

If reading ability is “one thing,” then it makes sense to unify the measurement and reporting of this quantity around a single metric. At present, hundreds of reading tests report in proprietary and non-exchangeable metrics. The life of the professional educator can be simplified tremendously by unifying the reading construct. In the 1700s, countries unified the measurement of temperature (Celsius and Fahrenheit). In the 1800s, nations unified the measurement of time (Greenwich mean time). All major norm-referenced reading tests have now been linked to the Lexile scale so as to enable test publishers to report in both their own proprietary metrics and a common supplemental metric. In particular, the linking technology already exists to translate reading scores into Lexiles from the SAT-10, Iowa Tests, Terra Nova, NWEA-MAPS, Gates MacGinitie, Stanford Diagnostic Reading Test, and Scholastic Reading Inventory, among others. Moreover, dozens of text publishers and text aggregators have adopted the Lexile scale as a standard for representing the readability of text and 19 million K–12 students now get a Lexile measure at least once a year.

Once unification of reading assessments has been realized, it will prove useful to link progress assessments to the common metric, so that reading assessment isn’t done only once a year, but monthly and even weekly. Periodic classroom-based progress assessments that report in the same metric as the high-stakes instruments will shift the reporting focus away from status (“how well is the student reading on this May morning?”) to growth (“on what growth trajectory is this reader, and what do we forecast reading ability to be at high school graduation?”). Indeed, a common metric spanning the

full developmental continuum makes it possible to build individual growth trajectories for each student.

When the causes of test item difficulty are known, it even becomes possible to engineer reading test items on demand, either by having human item writers follow a protocol or by teaching a computer to develop reading items with associated theoretical calibrations. The implications of this possibility are far-reaching and a little unsettling. Suppose, for example, that a future edition of a high stakes reading test did not involve a test nor bank of questions, but instead comprised a set of rules for generating appropriate test items, along with an adaptive algorithm for choosing the next best text/question for an examinee depending on what reading ability is revealed, and a richly annotated scale that links reported reader ability to appropriate texts (Stenner & Stone, 2003). Reproducible measures of reader ability can thus be made from counts correct on items that have never been administered before and may never be used again. In this scenario, test items are disposable commodities. Test security is guaranteed because no one sees the items until the computer builds them. In this case, releasing “the test” would mean either releasing sample items from the test or releasing the item-generating algorithm itself.

Perhaps the main difficulty with this approach is that among the randomly generated individual test questions, some will inevitably work better than others to measure ability. Because each item is used only once with one examinee, there would be no means of comparing whether some questions create a wider or narrower spread of answers than other questions, or whether certain groups of students perform better on a certain question. With each test item used only once, the problem of comparing test

questions that are better or worse would be less accessible to study. But the hope would be that in a test of reasonable length, these random differences would largely balance out.

This ultimate unification in the measurement of reading ability need not diminish the role of normative metrics for reading ability such as percentiles, stanines (in which students are ranked in nine groups, with the bottom three being below-average, the middle three being average, and the upper three being above average), and "normal curve equivalent" scores (in which students are measured along a scale from 1 to 99 based on how their score would rank them in a normal-curve distribution). However, measuring reading performance in terms of grade-equivalents should be abandoned in favor of text-based descriptions of reader performance, such as "This reader can read typical third grade text (500L to 600L) with 75% comprehension." With this formulation, it becomes possible to give a coherent answer to a common refrain of parents and school board members, "What does it mean to be on grade level?" For example, minimally acceptable end of third grade performance can be described as the level needed to read fourth grade textbooks with 60% or better comprehension, which emphasizes that the reading goal is not being set relative to other students, but is in place to ensure that students can handle the text readability of the grade to which they are being promoted.

Conclusion

We claim that the measurement of reading is philosophically, mathematically, and practically analogous to the measurement of temperature. Temperature theory is adequately developed so that thermometers can be constructed without reference to any

data. We know enough about liquid expansion coefficients, the gas laws, glass conductivity, and fluid viscosity to construct a remarkably precise temperature measurement device with recourse only to theory. Routine manufacture of thermometers occurs without even checking the calibrations against data with known values prior to shipping the instruments to customers. Furthermore, a second instrument developer can follow the same specification and produce another thermometer that measures temperature on the same scale and with a precision comparable to the first instrument. If two people are measured with two different thermometers, our confidence in temperature measurement is such that the readings can be usefully compared.

In recent decades, reading measurement has proceeded in a radically different manner: instruments (collections of items) are field-tested on thousands of readers, and empirical difficulties are estimated for each item. The potential is now upon us to move reading assessment from a data-driven to a theory-driven enterprise. The consequences of this shift for research and practice are hard to over-estimate (Stone, 2002). The Lexile approach is built on a theory of what makes text readable (word choice and syntax), how a measure of reader ability can be connected to this measure of text readability, and how to combine the readability of text and the ability of readers into a useful measure of comprehension. The ultimate result of building measures of reading on a strong theoretical base may be that if two people are tested for reading ability with two different tests, their scores will be comparable. Moreover, if anyone wants to generate a new reading test, then as long as they follow the same underlying specification, anyone using that test will have comparable results as well.

Text matters! Paradoxically, the last 25 years of reading research has celebrated the role of text but, for the most part, avoided measuring it. In much of the current literature, “reading performance” conflates reader ability, which does not depend on text, and reader comprehension, which does. However, reader ability and reading comprehension are conceptually and operationally separable. We have described how reader ability and text readability, measured in the common scale of Lexiles, can be used to model reading comprehension.

References

Andersen, E. B. (1977). Sufficient statistics and latent trait models. *Psychometrika*, 42(1), 69-81.

Barr, R., Kamil, M. L., & Mosenthal, P. B. (Eds.). (1984). *Handbook of reading research* (Vol. I). New York: Longman.

Barr, R., Kamil, M. L., Mosenthal, P. B., & Pearson, P. D. (1996). *Handbook of reading research* (Vol. II). Mahwah, NJ: Lawrence Erlbaum.

Boomsma, A., van Duijn, M. A. J., & Snigders, T.A.B., (Eds.). (2001). *Essays on item response theory*. New York: Springer.

Bormuth, J. R. (1966). Readability: New approach. *Reading Research Quarterly*, *7*, 79–132.

Carroll, J. B., Davies, P., & Richman, B. (1971). *The word frequency book*. Boston: Houghton Mifflin.

Carver, R. P. (1974). Measuring the primary effect of reading: Reading storage technique, understanding judgments and cloze. *Journal of Reading Behavior*, *6*, 249–274.

Carver, R. P. (2000). *The causes of high and low reading achievement*. Mahwah, NJ, Lawrence Erlbaum.

Chall, J. S. (1988). The beginning years. In B. L. Zakaluk & S. J. Samuels (Eds.), *Readability: Its past, present, and future*. Newark, DE: International Reading Association.

Crain, S., & Shankweiler, D. (1988). Syntactic complexity and reading acquisition. In A. Davidson & G. M. Green (Eds.), *Linguistic complexity and text comprehension: Readability issues reconsidered*. Hillsdale, NJ: Erlbaum Associates.

Daneman, M. (1996). Individual differences in reading skills, IV. M.L. Kamil, P.B. Mosenthal, P.D. Pearson & R. Barr (Eds.), Handbook of reading research (Vol. III, pp 512-538). Mahwah, NJ: Erlbaum Associates.

Davidson, A., & Kantor, R. N. (1982). On the failure of readability formulas to define readable text: A case study from adaptations. *Reading Research Quarterly, 17,* 187–209.

Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London, A,* 222, 309-368.

Hiebert, E. H. (2004) Standards, assessment and text difficulty. In A. E. Fastrup and S. J. Samuels (Eds.), *What research has to say about reading instruction* (3rd ed.) Newark, DE: Instructional Reading Association.

KAC. M. (1959). Statistical independence in probability, analysis, and number theory. New York: Wiley.

Kamil, M. L., Mosenthal, P. B., Pearson, P. D., & Barr, R. (2000). *Handbook of reading research* (Vol. III). Mahwah, NJ: Lawrence Erlbaum.

Klare, G. R. (1963). *The measurement of readability*. Ames, IA: Iowa State University Press.

Krashen, S. (2002). The Lexile framework: The controversy continues. *California School Library Journal*, 25(*2*), 29–31.

Liberman, I. Y., Mann, V. A., Shankweiler, D., & Werfelman, M. (1982). Children's memory for recurring linguistic and non-linguistic material in relation to reading ability. *Cortex, 18*, 367–375.

Linacre, J. M. (1987). *An extension of the Rasch model to multi-facet situations*. Chicago: University of Chicago Department of Education.

Linacre, J. M. (1989). *Many-faceted Rasch measurement*. Chicago: MESA Press.

Lobel, A. (1979). *Frog and toad are friends*. New York: HarperCollins.

Luce, R. D., & Tukey, J. W. (1964). Simultaneous conjoint measurement: a new type of fundamental measurement. *Journal of Mathematical Psychology, 1*, 1–27.

Michell, J. (1999). *Measurement in psychology: A critical history of a methodological concept.* Cambridge, MA: University Press.

Miller, G. A., & Gildea, P. M. (1987). How children learn words. *Scientific American, 257*, 94–99.

Palincsar, A. S., & Brown, A. L. (1984). Reciprocal teaching of comprehension-fostering and comprehension-monitoring activities. *Cognition and Instruction, 1*, 117–175.

Pressley, M. (2000). What should comprehension instruction be the instruction of? In M. L. Kamil, P. B. Mosenthal, P. D. Pearson, & R. Barr (Eds.), *Handbook of reading research* (Vol. III, pp. 545-561). Mahwah, NJ: Erlbaum.

Rasch, G. (1960 [1980]). *Probabilistic models for some intelligence and attainment tests.* Copenhagen and Chicago: University of Chicago Press.

Shankwiler, D., & Crain, S. (1986). Language mechanisms and reading disorder: A modular approach. *Cognition, 14*, 139–168.

Snow, C. E. (2002). *Reading for understanding: Toward a R&D program in reading comprehension*. Washington, DC: Rand Corporation.

Stanovich, K. E. (1986). Matthew effects in reading: some consequences of individual differences in the acquisition of literacy. *Reading Research Quarterly*, 21(*4*), 360–406.

Stenner, A. J., Burdick, H., Sanford, E., & Burdick, D. S. (2004). *How accurate are lexile text measures?* Durham, NC: MetaMetrics.

Stenner, A. J., Horabin, I., Smith, D. R., & Smith, M. (1988, June). Most comprehension tests do measure reading comprehension: A response to McLean and Goldstein. *Phi Delta Kappan*, 765–769.

Stenner, A. J., Smith, M., & Burdick, D. S. (1983). Toward a theory of construct definition. *Journal of Educational Measurement*, *20*(4), 305–315.

Stenner, A. J., & Stone, M. H. (2003, in press). Item specifications vs. item banking. *Rasch Measurement Transactions, 17*(3), 929–930.

Stenner, A. J., & Wright, B. D. (2002). *Readability, reading ability, and comprehension*. In B. D. Wright & M. H. Stone. (2004). *Making Measures*. Chicago: Phaneron Press.

Stone, M. H. (2002). *The Knox cube test: A manual for clinical and experimental uses*. Wood Dale, IL: Stoelting Company.

Stone, M. H., Wright, B. D., & Stenner, A. J. (1999). Mapping variables. *Journal of outcome measurement, 3*(4), 308–322.

Williamson, G. L. (2004). Student Readiness for Postsecondary Options. [Online].

Available:http://www.lexile.com/lexilearticles/Student%20Readiness%20for%20Postsecondary%20Options%20(v4_1).pdf

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis: Rasch measurement*. Chicago, IL: MESA Press.

Wright, B. D., & Stone, M. H. (1979). *Best test design*. Chicago: MESA Press.

Wright, B.D., & Stone, M.H. (2004). *Making Measures*. Chicago: Phareon Press.

Author Note

It is impossible for us to express fully our appreciation to Benjamin D. Wright for decades of dialogue, inspiration, scholarship, and friendship. Thanks to David P. Pearson for suggesting this revision and Tim Taylor for many helpful suggestions. Thanks also to Donald S. Burdick, William P. Fisher, Jr., Eleanor E. Sanford, Hal Burdick, Carl Swartz, Robert C. Calfee, Jack Carroll (deceased), and Richard Venezky.

Correspondence concerning this article should be addressed to A. Jackson Stenner, MetaMetrics, Inc., 1000 Park Forty Plaza Dr., Suite 120, Durham, NC 27713 USA, (919) 547-3402, jstenner@lexile.com.