The Objective Measurement of Reading Comprehension In Response to Technical Questions Raised by the California Department of Education Technical Study Group

A. Jackson Stenner
MetaMetrics, Inc.
2327 Englert Drive, Suite 300
Durham, NC 27713

Donald S. Burdick
Institute of Statistics and Decision Sciences
Duke University

January 3, 1997

Contents

1 MEASUREMENT DEFINED

Measurement is the process of converting observations into quantities through theory. Measurement as a "process" implies an "act of ascertainment of finding out" (Leonard, 1962, p. 4). The term "observation" refers to the qualitative observation or count, be it the color of a blood glucose strip or the position of a column of mercury. The "quantity", or measure, is the number assigned to the attribute of the object of measurement. The term "theory" in this definition makes clear that "every instance of measurement presupposes an extensive background of explicitly confirmed, scientific theory" (Leonard, 1962, p. 4).

A construct theory in its more colloquial form, is just a story about what it means to move up and down a scale. It is used to calibrate indicants. Examples of calibration include the placement of lines on the tube of

a liquid-in-a-glass thermometer or the assignment of difficulty calibrations to a series of vocabulary test items. The theory creates a context in which the observation can be understood as the data for an estimator for the measure. In the case of the attribute"reading comprehension", the "process" is the act of ascertaining the level of reading comprehension attained by a person.

The process of measurement results in a quantity. The "observation" is often a raw score or count correct on some set of items. The "quantity" is the amount of reading comprehension ability that a person possesses expressed in some metric. The conversion of observations into measures through theory is accomplished using the Rasch (1980) model, which states a requirement for the way that theory (expressed as item calibrations) and observations (count of correct items) interact in a probability model to make useful measures.

The above definition of measurement can be implemented by an equation that expresses the relationships among an observation (O_p), theory ($C_1, \ldots, C_L$), and a measure (M_p):

$$O_p = \sum_{i=1}^{L} \frac{e^{(M_p - C_i)}}{1 + e^{(M_p - C_i)}} \tag{1}$$

The Rasch model stipulates that the probability of a response to an indicant is governed by the difference between the indicant calibration (C_i) and the person's measure (M_p). The number correct for a person, i.e. the observation O_p, is set equal to the sum of these modeled probabilities yielding Equation (1), which is then solved for the measure M_p. When a person's measure greatly exceeds the items' calibrations, then the probabilities will be high and the sum of these probabilities will correspond to a high number correct. Conversely, when the item calibrations generally exceed the person measure, the modeled probabilities of a correct response will be low which corresponds to a low number correct. When we know the observation (O_p) and the indicant calibrations (C_i), we can use an iterative procedure to find the measure (M_p) that will make the sum of the modeled probabilities equal to the observation (O_p).

Formula (1) possesses several distinguishing characteristics:

- The key terms from the above definition of measurement are placed in a precise relationship to one another.

- The individual responses of, say, a person to each item on an instrument

are absent from the equation. The only piece of data that survives the act of observation is the number correct , thus confirming that this number is sufficient for estimating the measure.

- For any set of items we know the possible raw counts. When it is possible to know the indicant calibrations from theory, the only parameter that must be estimated in (1) above is the measure that corresponds to each observable raw count (i.e., number correct). Thus, when the calibrations (C_i) are given by theory, a correspondence table linking observation and measure can be constructed without reference to data on other individuals. Lest this seem surprising, we emphasize that theory-based calibration of instrumentation is the norm throughout science, engineering and commerce. Due largely to a dearth of good theory, the behavioral sciences have been forced to rely on data alone for instrument calibration.

The measure (M_p) in (1) depends only upon the particular indicant calibrations provided by theory and on the observation. The observation (i.e., count correct) is completely sufficient as an estimator of the measure. That is, there is no further information in the data or in the context of measurement that can improve the estimate of the measure. This "complete sufficiency" is possible only in that specific arrangement of observation, theory and measure expressed in (1) above. Only the Rasch model, informed by theory-based calibrations, offers complete sufficiency.

Even more important than sufficiency is the property of objectivity. The Rasch model (Wright and Stone, 1979), in combination with a construct theory (Stenner, Smith, and Burdick, 1983), allows the complete separation of the measure from particulars associated with the act of measuring. A consequence of this complete separation is that measures achieve general objectivity, a concept we explore in depth in the next section.

1.1 Specific and General Objectivity

In this section, we review the calls for objectivity in behavioral science measurement and differentiate between two kinds of objectivity: specific and general.

Measurement is objective if it is independent of the conditions of measurement, e.g. which instrument is used to do the measuring. If the measured

difference between objects is independent of conditions, the objectivity is specific or local. If the scale value of a single object is independent of conditions, the objectivity is general. Because of the ambiguity in the location of the origin, the Rasch model by itself yields local but not general objectivity. In combination with a construct theory like the Lexile Framework, a criterion referenced interpretation is obtained which resolves the ambiguity and yields general objectivity.

Objectivity is the foundation of valid measurement. Indeed, it is central to the idea of measurement. When a number is reported as a measure, the underlying assumption is that the measure is objective, that is, it has been sufficiently well-separated from the conditions of measurement that we can ignore these conditions when reporting the measure. For example, if we say that it is 80 degrees Fahrenheit, the inherent assumption is that the validity of this measure does not depend on any conditions of measurement, such as which thermometer was used. This objectivity, which is the foundation of physical science measurement, has not been achieved to the same degree in the behavioral sciences, although it has long been sought after, witness Thurstone:

> "It should be possible to omit several test questions at different levels of the scale without affecting the individual score.
>
> It should not be required to submit every subject to the whole range of the scale. The starting point and the terminal point being selected by the examiner should not directly affect the individual score" (Thurstone, 1926, p. 446).

It is clear that Thurstone believed that person measures should be independent of the particular items used in the measurement instrument. What is not clear is whether or not Thurstone intended that relative scores (i.e., differences) or absolute scores (i.e., point locations) should be free of effects due to indicants. Two years later, he stated:

> " **The scale must transcend the group measured.** One crucial experimental test must be applied to our method of measuring attitudes before it can be accepted as valid. A measurement instrument must not be seriously affected in its measuring function by the object of measurement. To the extent that its measuring function is so affected, the validity of the instrument

> is impaired or limited. If a yardstick measured differently because of the fact that it was a rug, a picture or a piece of paper that was being measured, then to that extent the trustworthiness of that yardstick as a measuring device would be impaired. Within the range of objects for which the measuring instrument is intended, its function must be independent of the object of measurement" (Thurstone, 1928, p. 547).

¿From 1926 to 1931, Thurstone published examples of "objectivity", including weight and height, that suggest an interest in sample-free, absolute measurement, although his models result in only an approximation of specific objectivity. Thus Thurstone's philosophizing on the attributes of "good" measures focused on general objectivity (absolute scale locations are independent of the instrument), whereas his mathematical models and research applications realized only local objectivity (differences among persons are independent of the measuring instrument) (Stenner, 1994).

Although Thurstone did not use the word objectivity, he clearly had this concept at the forefront of his thinking. Georg Rasch made objectivity the centerpiece of a new psychometric model. Rasch (1960) states:

> "Individual-centered statistical techniques require models in which each individual is characterized separately and from which, given adequate data, the individual parameters can be estimated. It is further essential that comparisons between individuals become independent of which particular instruments—tests or items or other stimuli—within the class considered have been used. Symmetrically, it ought to be possible to compare stimuli belonging to the same class 'measuring the same thing' independent of which particular individuals within a class considered were instrumental for the comparisons" (Rasch, 1980).
>
> "Where this law can be applied, it provides a principle of measurement on a ratio scale of both stimulus parameters and object parameters, the conceptual status of which is comparable to that of measuring mass and force. Thus, by way of an example, the reading accuracy of a child—as ascertained by means of any of the oral reading tests catalogued in the appendix—can be measured with the same kind of objectivity as we may tell its weight,

though not with the same degree of precision, to be sure, but that is a different matter" (Rasch, 1980, p. 115).

"Thus, if a set of empirical data cannot be described by the [Rasch] model, then complete specifically objective statements cannot be derived from them. Firstly, the failing of specific objectivity means that the conclusions about, say, any set of person parameters will depend on which other persons also are compared. As a parody we might think of the comparison of the volumes of a glass and a bottle as being influenced by the heights of some books on a shelf.

"Secondly, the conclusions about the persons would depend on just which terms were chosen for the comparison, a situation to which a parallel would be that the relative height of two persons would depend on whether the measuring stick was calibrated in inches or in centimeters" (Rasch, 1968, p. 7).

"Thus, in principle, the [M_p's] stand for properties of the objects per se, regardless of which [C_i's] might be used for locating them. Therefore, they really ought to be appraised without any reference to the [C_i's] actually employed for this purpose, just like reading the temperature of an object should give essentially the same result whichever adequate thermometer was used" (Rasch, no date, p. 5).

Rasch (1960) coined the term "specific objectivity" and realized that his model achieved a separation of instrument and measure long sought after, but never achieved. Whether or not Rasch distinguished between specific and general objectivity is not clear.

Rasch was always careful to point out that it was comparisons (i.e., relative measures) that were independent of the instrument, suggesting that he clearly understood the distinction. But his favorite physical science examples (mass and temperature) clearly possess a more general and complete objectivity not shared by the reading comprehension tests he developed.

Finally, Wright (1968) offered an accessible and complete statement on specific objectivity. He wrote:

"Let us call measurement that possesses this property 'objective'. Two conditions are necessary to achieve it. First, the calibration

> of measuring instruments must be independent of those objects that happen to be used for calibration. Second, the measurement of objects must be independent of the instrument that happens to be used for measuring. In practice, these conditions can only be approximated, but their approximation is what makes measurement objective.
>
> Object-free instrument calibration and instrument-free object measurement are the conditions that make it possible to generalize measurement beyond the particular instrument used, to compare objects measured on similar but not identical instruments, and to conceive or partition instruments to suit new measurement requirements" (Wright, 1968).

After a quarter-century of exploration of "objectivity" as a fundamental requirement of measurement:

> "Objectivity is the expectation and, hence, requirement that the amount and meaning of a measure has been well enough separated from the measuring instrument and the occasion of measurement that the measure can be used as a quantity without qualification as to which was the particular instrument or what was the specific occasion.
>
> Although a measuring occasion is necessary for a measure to result, the utility of the measure depends on the specifics of the occasion disappearing from consideration. It must be possible to take the occasion for granted and, for the time being, to forget about it. Were such a separation of meaning from the circumstances of its occasion not possible, not only science, but also commerce, and even communication, would become impossible" (Wright, 1991, p. 1).

Objectivity is the cornerstone of measurement. Measures must be completely independent of the particular instruments used and the particular conditions of measurement surrounding their use. A critical distinction, however, exists between specific or local objectivity, as achieved by Rasch, and the general objectivity that is inherent in measures from physical science.

"Local objectivity" is a consequence of a set of data fitting the Rasch model. When the data fit, differences between object measures and indicant

calibrations are shown to be sample-independent. This means that apart from random error two indicants must differ by the same amount no matter which sample of objects actually responds to the indicants. Similarly, two objects must differ by the same amount no matter which samples of indicants (from the relevant universe) are used to implement the measurement procedure. Consequently, when data fit the Rasch model, then the relative locations of objects and indicants on the underlying continuum for a construct are sample and test independent.

An ideal, approximated by measures in physics and chemistry (e.g., temperature measurements), is that absolute location of an object on, for example, the Celsius scale, is independent of the instruments and conditions of measurement. Temperature theory is well enough developed that thermometers can be constructed without reference to any data. In fact, routine manufacture of thermometers occurs without checking the calibrations against data with known values prior to shipping the instruments to customers. Such is our collective confidence in temperature theory. We know enough about liquid expansion coefficients, gas laws, glass conductivity and fluid viscosity to construct a usefully precise measurement device with recourse to theory only.

The consequence of operating with a construct theory and associated calibration equations is that general objectivity is achieved. Measurement of the temperature of two objects results in not just sample independence for the difference between their temperatures but sample independence for the point estimate of each object's temperature reading.

In summary, specific or local objectivity as achieved with the Rasch model ensures that relative measures, that is, the differences between objects or between indicants, are independent of the conditions of measurement. In contrast, general objectivity ensures that absolute measures, the amounts themselves, are similarly independent.

The fundamental requirement of a fully objective measurement procedure, therefore, is that it be capable of converting an observation (raw count) into a measure without recourse to individual or group data on indicants or objects. We call this feature of a measurement procedure "general objectivity":

> "The difference between local and general objectivity is seen not to be a consequence of the fundamental natures of the social and physical sciences, nor to be a necessary outcome of the method

> of making observations, but to be entirely a matter of the level of sophistication of the theory underlying the construction of the particular measurement instruments" (Stenner, 1990, p. 111).

We turn now to an application of these methods to the measurement of reading comprehension.

2 THE LEXILE FRAMEWORK

Reading comprehension is the most tested construct in education. It is probable that reading comprehension ability is measured more frequently than temperature, height, or weight among students ages 6 to 18. Reading comprehension ability is widely recognized as the best predictor of success in higher education and on-the-job performance. Economists and educators have joined in identifying low literacy as a primary causal factor in the United States' dwindling economic productivity. In an information age, reading comprehension is a survival skill.

The Ninth Mental Measurements Yearbook (Mitchell, 1985) reviews 97 reading comprehension tests. Associated with each of these tests is a conceptual rationale (however primitive) and a scale. Thus there are 97 different reading measures. The current status of reading comprehension measurement is reminiscent of late seventeenth-century temperature measurement, in which the absence of a unifying temperature theory resulted in about 30 different scales competing for favor throughout Europe. The consequence for science and commerce was chaos. In a similar fashion, the presence of dozens of competing reading comprehension scales produces confusion among educators, researchers, policy makers, parents and students.

2.1 The Lexile Theory

We communicate using various symbol systems including mathematics, music, and language. All symbol systems share two features: a semantic component and a syntactic component. In mathematics, the semantic units are numbers, variables, and operators, which are combined according to rules of syntax into mathematical expressions. In music, the semantic units are notes, arranged according to rules of syntax to form chords and phrases. In language, the semantic units are words. Words are organized according to

rules of syntax into thought units and sentences (Carver, 1974). In all cases, the semantic units vary in familiarity and the syntactic structures vary in complexity. The comprehensibility or difficulty of a message is dominated by the familiarity of the semantic units and by the complexity of the syntactic structures used in constructing the message.

2.2 The Semantic Component

As far as the semantic component is concerned, it is clear that most operationalizations are proxies for the probability that an individual will encounter a word in a familiar context and thus be able to infer its meaning (Bormuth, 1966). This is the basis of exposure theory, which explains the way receptive or hearing vocabulary develops (Miller and Gildea, 1987; Stenner, Smith, and Burdick, 1983). Klare (1963) builds the case for the semantic component varying along a familiarity-to-rarity continuum, a concept that is further developed by Carroll, Davies, and Richman (1971), whose word-frequency study examined the reoccurrence of words in a five-million-word corpus of running text. 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 thus become a part of an individual's receptive vocabulary.

Variables such as the average number of letters or syllables per word have been found to be proxies for word frequency. There is a high negative correlation between the length of words and the frequency of word usage. Polysyllabic words are used less frequently than monosyllabic words, making word length a good proxy for the likelihood of an individual being exposed to them.

Stenner, Smith, and Burdick (1983) analyzed more than 50 semantic variables in order to identify those elements that contributed to the difficulty of the vocabulary items on Forms L and M of the Peabody Picture Vocabulary Test—Revised (Dunn and Dunn, 1981). Variables included parts of speech, number of letters, number of syllables, the modal grade at which the word appeared in school materials, content classification of the word, the frequency of the word from two different word counts, and various algebraic transformations of these measures. Correlations were calculated between the logit difficulties of the test items and each predictor variable. The best operationalization of the semantic component of reading was found to be word

frequency.

The word frequency measure used was the raw count of how often a given word appeared in a corpus of 5,088,721 words sampled from a broad range of school materials (Carroll, Davies, and Richman, 1971). Exploratory data analysis was performed to test the explanatory power of this variable. This analysis involved calculating the mean word frequency for each of 66 reading comprehension test passages from the Peabody Individual Achievement Test (Dunn and Markwardt, 1970). Correlations were obtained between algebraic transformations of these means and the rank order of the test items. Since the items were ordered according to increasing difficulty, the rank order was used as a proxy for item difficulty. The mean log word frequency provided the highest correlation with item rank order.

2.3 The Syntactic Component

Sentence length is a powerful proxy for the syntactic complexity of a passage. But an important caveat is that sentence length is not the underlying causal influence (Chall, 1988). Researchers sometimes incorrectly assume that manipulation of sentence length will have a predictable effect on passage difficulty. Davidson and Kantor (1982), for example, illustrate rather clearly that sentence length can be reduced and difficulty increased and vice versa.

Klare (1963) provides a possible interpretation for how sentence length works in predicting passage difficulty. He speculates that the syntactic component varies with the load placed on short-term memory. This explanation is supported by Crain and Shankweiler (1988), Shankweiler and Crain (1986), and Liberman, Mann, Shankweiler, and Westelman (1982), whose work has provided evidence that sentence length is a good proxy for the demands that structural complexity places upon verbal short-term memory.

Algebraic transformations of the mean sentence length for the 66 Peabody Individual Achievement Test (PIAT) reading comprehension items were again correlated with item rank order. It was found that the log of the mean sentence length was the best predictor of passage difficulty. Thus, the two constructs in the Lexile Theory of reading (semantics and syntax) are operationalized in terms of word frequency and sentence length into a calibration equation which can be used to scale the difficulty of English language text and reading comprehension test items.

2.4 The Calibration Equation

The word-frequency and sentence-length measures were combined in a regression equation designed to "explain" most of the variance found in any set of reading comprehension task difficulties. A provisional equation was developed from a regression analysis of the PIAT reading comprehension items. The log of the mean sentence length and the mean of the log word frequencies combined to explain 85 percent of the variance (r=.92) in PIAT item rank order.

The regression equation produced by this analysis was used to assign theoretical difficulties to 400 pilot test items (see Figure 1). The pilot items were ordered by difficulty and administered to approximately 3,000 students from grades 2 to 12. Misfitting items were removed, leaving 262 test items for which observed logit difficulties were obtained using the Rasch analysis program MSCALE. (Wright, Rossner, and Congdon, 1985).

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.

The final specification equation was based upon the observed logit difficulties for the remaining 262 pilot test items. A regression analysis was performed using sentence length and word frequency as predictors of these logit difficulties. The adjusted correlation between the observed logit difficulties and the theoretical calibrations was .97. The weights produced by the regression analysis yielded the following equation:

$$\text{Theoretical Logit} = (9.82247 \times \text{LMSL}) - (2.14634 \times \text{MLWF}) - \text{constant} \quad (2)$$

where LMSL = log of the mean sentence length and MLWF= mean of the log word frequencies.

2.5 The Lexile Scale

The logit scale obtained from a Rasch analysis using MSCALE has its zero located at the mean difficulty of the items used. An item would therefore experience a shift in its logit difficulty obtained using MSCALE if it were transferred to a test with different mean difficulty, which violates general objectivity. General objectivity requires that scores obtained from different test administrations be tied to a common zero. To achieve general objectivity the theoretical logit difficulties obtained from Equation (2) must be transformed to a scale in which the ambiguity regarding the location of zero is resolved.

The method for setting a scale with a fixed zero is easily described. First, identify two anchor points for the scale. They should be intuitive, easily reproduced, and widely recognized. For most thermometers, the anchor points are the freezing and boiling points of water. For the Lexile Scale, the anchor points are text from seven basal primers for the low end and text from the Electronic Encyclopedia (Grolier, 1986) for the high end. These points correspond to middle of first grade text and the midpoint of workplace text.

Second, using Equation (2), obtain the logit difficulty of the two anchors. The mean logit difficulty of the primer material was -3.3 and the mean logit difficulty for the encyclopedia samples was $+2.26$.

Third, determine the unit size. For the Celsius thermometer, the unit size (a degree) is 1/100 of the difference between freezing (0 degrees) and boiling (100 degrees) water. For the Lexile scale the unit size was defined as 1/1000. Therefore, a Lexile by definition equals 1/1000th of the difference between the comprehensibility of the primers and the encyclopedia.

Fourth, assign a value to the lower anchor. The low-end anchor on the Lexile scale was assigned a value of 200. Zero was not used as the low-end value in order to minimize the occurrence of negative Lexile values.

Finally, a linear equation of form

$$(\text{logit score} + \text{constant}) \times \text{CF} + 200 = \text{Lexile text measure} \qquad (3)$$

was developed to convert logit difficulties to Lexile calibrations. The values of the conversion factor CF and the additive constant are determined from

the anchors. Equation (2) yields difficulties of -3.3 logits for the primers and 2.26 logits for the encyclopedia. Plugging these values for the logit score into (3) and setting the respective Lexile scores equal to 200 and 1200 yields the following equations:

$$\begin{aligned} (-3.3 + \text{constant}) \times \text{CF} + 200 &= 200\text{Lexiles} \\ (2.26 + \text{constant}) \times \text{CF} + 200 &= 1200\text{Lexiles} \end{aligned} \tag{4}$$

Solving these equations yields 3.3 for the constant and 180 for the conversion factor. Thus, the final equation that converts theoretical logit difficulties produced by the Equation (2) into Lexile units is:

$$[(\text{Logit} + 3.3) * 180] + 200 = \text{Lexile text measure} \tag{5}$$

Measurements for all persons and all texts are now reportable in a common unit, a Lexile, which is similar to the degree calibrations on a thermometer. The higher the Lexile measure for a text, the more difficult the material is to read and the more reading ability a student must possess to comprehend the text. Text measures are located on the Lexile map (see attachment) at the point corresponding to a person with the ability to achieve 75 percent comprehension. People are located on the scale by analyzing their performance on calibrated reading tasks. They are located on the map at the point where they are forecast to realize 75 percent comprehension. A person with a Lexile measure of 1000L is expected to answer correctly 75 percent of native Lexile items sampled from a text with a 1000L measure. This provides the means for matching a person's Lexile literacy measure with reading materials at his comprehension level. The attached Lexile map is a poster-sized graphic that can be used to bring meaning to the relation between text and person measures.

2.6 Testing the Lexile Equation

A computer program incorporating the Lexile equation has been developed that analyzes continuous prose and reports text measures in Lexiles (MetaMetrics, 1995). In order to test the utility of the theory, 1,780 reading comprehension test items appearing on 9 nationally normed tests were analyzed (Stenner, Smith, Horabin, and Smith, 1987). The study correlated empirical item difficulties provided by the publisher with the Lexile calibrations

specified by the computer analysis of the text of each item. In those cases where multiple questions were asked about a single passage, empirical item difficulties were averaged to yield a single observed difficulty for the passage. See Table 1 for detail on this analysis.

The empirical difficulties were obtained in one of three ways. Three of the tests included observed logit difficulties from either a Rasch or three-parameter analysis (e.g., NAEP). For four others, logit difficulties were estimated from item p-values and raw score means and standard deviations (e.g., CAT). TestCalc (Horabin, 1989), a computer program for analyzing test data, was used to estimate these logit difficulties. Two of the tests provided no item parameters, but in each case items were ordered on the test in terms of difficulty (e.g., PIAT). For those tests, the empirical difficulties were approximated by the difficulty rank order of the item.

Once theory-specified calibrations and empirical item difficulties were computed, the two arrays were correlated and plotted separately for each test. The plots were checked for unusual residuals and curvature, and it was discovered that the equation did not fit poetry items or non-continuous prose items (e.g., recipes, menus, or shopping lists). This indicated that the universe to which the Lexile equation could be generalized was limited to continuous prose. The poetry and non-continuous prose items were removed and correlations were again obtained and used to describe the fit of observation to theory.

Model misspecification is not the only influence which operates to deflate the correlation between theory and observation. Another is range restriction in the item difficulties. Some tests do not cover the full developmental continuum for reading comprehension. The NAEP (1983), for example, is administered to grades four, eight, and eleven. The resulting restriction in the range of item difficulties attenuates the correlation between theory and data. Thorndike (1949) gives a procedure for correcting a correlation for restriction in range where the range of the variable in the unrestricted group is known.

Table 1 presents the results of correlating the theoretical calibrations and observed difficulties for nine tests. The last two columns of the table show the raw correlations between observed item difficulties and theoretical item calibrations, and these same correlations corrected for restriction in range. The Fisher Z mean of the raw correlations is $R_{(OT)} = .84$. When corrections are made for range restriction, the Fisher Z mean disattenuated correlation

TABLE 1
Correlations Between Theory-Based Calibrations Produced by the Lexile Equation and Data-Based Item Difficulties

Test	Number of Questions	Number of Passages	Mean	SD	Range	Min	Max	$r_{(OT)}$	$R_{(OT)}$	$R'_{(OT)}$
SRA	235	46	644	353	1303	33	1336	.95	.97	1.00
CAT-E	418	74	789	258	1339	212	1551	.91	.95	.98
Lexile	262	262	771	463	1910	-304	1606	.93	.95	.97
PIAT	66	66	939	451	1515	242	1757	.93	.94	.97
CAT-C	253	43	744	238	810	314	1124	.83	.93	.96
CTBS	246	50	703	271	1133	173	1306	.74	.92	.95
NAEP	189	70	833	263	1162	169	1331	.65	.92	.94
Battery	26	26	491	560	2186	-702	1484	.88	.84	.87
Mastery	85	85	593	488	2135	-586	1549	.74	.75	.77
Totals Grand Means	**1780**	**722**	**767**	**343**	**1441**	**50**	**1491**	**.84**	**.91**	**.93**

$r_{(OT)}$ = raw correlation between observed difficulties (O) and theory-based calibrations (T).

$R_{(OT)}$ = correlation between observed difficulties (O) and theory-based calibrations (T) corrected for range restriction.

$R'_{(OT)}$ = correlation between observed difficulties (O) and theory-based calibrations (T) corrected for range restriction and measurement error.

*Means are computed on Fisher $\bar{Z}$ transformed correlations.

between theory-based calibration and empirical difficulty in an unrestricted group of reading comprehension items is $R'_{(OT)} = .91$. These results show that most attempts to measure reading comprehension, no matter what the item form, type of skill objectives intended or response requirement used, measure the common comprehension factor specified by the Lexile theory.

A second study was performed in which Lexile calibrations were obtained for units in 11 basal series. It was presumed that each basal series was sequenced by difficulty. So, for example, the latter portion of a third-grade reader is presumably more difficult than the first portion of the same book. Likewise, a fourth-grade reader is presumed to be more difficult than a third-grade reader. Observed difficulties for each unit in a basal series were estimated by the rank order of the unit in the series. Thus, the first unit in the first book of the first-grade was assigned a rank order of one and the last unit of the eighth-grade reader was assigned the highest rank order number. Correlations were computed between the ranked order and the Lexile calibration of each unit. After correction for range restriction the Fisher Z average correlation between the Lexile calibration of text comprehensibility and the rank order of the basal units was .97 (see Table 2).

The fact that Lexile theory accounted for the unit rank ordering of 11 basal series is all the more noteworthy when we recognize that the series differ in prose selections, the developmental range addressed, the types of prose introduced (i.e., narrative versus expository), and the purported skills and objectives they emphasize. The theory works throughout the full developmental range from pre-primer (-200 Lexiles to 200 Lexiles) through advanced graduate school material (1400 Lexiles to 1800 Lexiles).

2.7 Interpreting Lexile Measures

One of the biggest shortcomings of many current testing procedures is the limited usefulness of the normative interpretation of a score. A normative interpretation only expresses how a student did on the test compared to other students of the same age or grade. A student's performance is typically reported as a percentile. A percentile of 65 for a third-grade girl indicates that she scored better than 65 percent of the third-grade students involved in the norming study. Percentile scores on standardized reading tests do not provide any information about what a student can or cannot read. What can a teacher or parent do with a percentile score? What kind of instruction can a

TABLE 2

Correlations Between the Lexile Calibration and The Rank Order of Unit

BASAL SERIES	Number of Units	Mean	Standard Diviation	Minimum	Maximum	$r_{(OT)}$	$R_{(OT)}$	$R'_{(OT)}$
Ginn Rainbow Series (1985)	53	522	272	-109	1101	.93	.98	1.00
HBJ Eagle Series (1983)	70	549	251	38	1030	.93	.98	1.00
Scott Foresman Focus Series (1985)	92	552	159	145	871	.84	.99	1.00
Riverside Reading Series (1986)	67	609	231	96	1034	.87	.97	1.00
Houghton-Mifflin Reading Series (1983)	33	449	284	-209	964	.88	.96	.99
Economy Reading Series (1986)	67	639	265	-1	1198	.86	.96	.99
Scott Foresman American Tradition (1987)	88	696	239	106	1211	.85	.97	.99
HBJ Odyssey Series (1986)	38	662	209	276	1123	.79	.97	.99
Holt Basic Reading Series (1986)	54	615	299	10	1214	.87	.96	.98
Houghton-Mifflin Reading Series (1986)	46	707	256	242	1236	.81	.95	.98
Open Court Headway Program (1985)	52	694	197	-133	1018	.54	.94	.97
Grand Means	**660**	**609**	**242**	**42**	**1091**	**.86**	**.97**	**.99**

$r_{(OT)}$ = raw correlation between observed difficulties (O) and theory-based calibrations (T).

$R_{(OT)}$ = correlation between observed difficulties (O) and theory-based calibrations (T) corrected for range restriction.

$R'_{(OT)}$ = correlation between observed difficulties (O) and theory-based calibrations (T) corrected for range restriction and measurement error.

*Means are computed on Fisher Z transformed correlations.

teacher give a student when the only information provided is that a particular child is reading at the 65th percentile of some sample of third-graders?

An important feature of the Lexile Framework is that it provides criterion-referenced interpretations of every measure. When a person's measure is equal to the task's calibration, then the Lexile scale forecasts that the individual has a 75 percent comprehension rate on that task. When 20 such tasks are given to this person, one expects three-fourths of the responses to be correct.

There is some empirical evidence supporting the choice of a 75 percent target comprehension rate, as opposed to, say, a 50 percent or 90 percent rate. Squires, Huitt, and Segars (1983) found that reading achievement for second-graders peaked when the success rate reached 75 percent. A 75 percent success rate also is supported by the findings of Crawford, King, Brophy, and Evertson (1975).

Since the Lexile theory provides complementary procedures for measuring people and text, the scale can be used to match a person's level of comprehension with books that the person is forecast to read with a high comprehension rate. Up to this time, trying to identify possible supplemental reading for students has, for the most part, relied on a teacher's familiarity with the titles. For example, an eighth-grade girl who is interested in sports but is not reading at grade level might be able to handle a biography on Chris Evert. The teacher may not know, however, whether that biography is too difficult or too easy for the student. The Lexile Framework provides a reader measure and text measure on the same scale. Armed with this information, a teacher, librarian, student, or parent can plan for success.

Students develop reading comprehension skills by reading. Skill development is enhanced when their reading is accompanied by frequent response requirements. Response requirements may be structured in a variety of ways. An instructor may ask oral questions as the reader progresses through the prose or written questions may be embedded in the text, much as is done with Lexile test items. Response requirements are important; unless there is some evaluation, there can be no assurance that the reader is properly targeted and comprehending the material. Students need to be given text on which they can practice being a competent reader (Smith, 1973). The above approach does not complete a fully articulated instructional theory, but its prescription is straightforward. Students need to read more and teachers need to monitor this reading with some efficient response requirement. One

implication of these notions is that some of the time spent on skill sheets might be better spent reading targeted prose with concomitant response requirements (Anderson, Hiebert, Scott, and Wilkinson, 1984).

As the reader improves, new titles with higher text measures can be chosen to match the growing person measure, thus keeping the comprehension rate at the most productive level. We need to locate a reader's "edge" and then expose the reader to text that plays on that edge. When this approach is followed in any domain of human development, the edge moves and the capacities of the individual are enhanced.

What happens when the edge is over-estimated and repeatedly exceeded? In physical exertion, if you push beyond the edge you feel pain; if you demand even more from the muscle, you will experience severe muscle strain or ligament damage. In reading, playing on the edge is a satisfying and confidence-building activity, but exceeding that edge by over-challenging readers with out of reach materials reduces self-confidence, stunts growth and results in the individual "tuning out". The tremendous emphasis on reading in daily activities, makes every encounter with written text a reconfirmation of a poor reader's inadequacy. Is it any wonder that 15 to 20 percent of US high school students decide to find some other way to spend their days (Hahn, 1987)?

For individuals to become competent readers, they need to be exposed to text that results in a comprehension rate of 75 percent or better. If the match between reader and text results in a 50 percent comprehension rate, there will be too much unfamiliar vocabulary and too much of a load placed on the reader's tolerance for syntactical complexity for that reader to attend to meaning. The rhythm and flow of familiar sentence structures will be interrupted by frequent unfamiliar vocabulary, resulting in inefficient chunking and short-term memory overload. When readers are correctly targeted, they read fluidly with comprehension; when incorrectly targeted, they struggle both with the material and with maintaining their self-esteem.

2.8 Forecasting Comprehension Rates

A student with a measure of 600L who is given a text measured at 600L is expected have a 75 percent comprehension rate. This 75 percent comprehension rate is the basis for selecting text that is targeted to a student's reading ability, but what exactly does it mean? And what would the comprehension rate be if this same student is given a text measured at 350L or one at 800L?

The 75 percent comprehension rate for a student-text pair can be given an operational meaning by imagining the text to be carved into item-sized slices of approximately 120 words each with a question embedded in each slice. A student who answers 3/4 of the questions correctly has a 75 percent comprehension rate.

Suppose instead that the text and student measures are not the same. It is the difference in Lexiles between person and text that governs comprehension. If the text measure is less than the student measure, the comprehension rate will exceed 75 percent. If not, it will be less. The question is: by how much? What is the expected comprehension rate when a 600L student reads a 350L text?

If all the item-sized slices in the 350L text had the same calibration, which would, of course, be 350L, the 250L difference between the 600L reader and the 350L text could be converted to logits with the conversion factor of 180 and, after adding the 1.1 logit offset, plugged into the Rasch model to obtain the expected comprehension rate. Unfortunately, comprehension rates calculated by this straightforward procedure will be biased because the slice calibrations in ordinary prose are not all the same. The average difficulty level of the slices and their variability both affect the comprehension rate.

Although the exact relationship between comprehension rate and the pattern of slice calibrations is complicated, experience has shown that a useful approximation results if we use the straightforward procedure just described but change the conversion factor from 180 to 225. This yields the following equation for comprehension rate of a text:

$$\text{Rate} = \frac{e^{\text{Eld}+1.1}}{1 + e^{\text{Eld}+1.1}} \tag{6}$$

where Eld is the "effective logit difference" given by

$$\text{Eld} = \frac{\text{Person Lexile measure} - \text{Text Lexile measure}}{225}.$$

Tables 3 and 4 show the comprehension rates calculated from Equation (6) for various combinations of person and text measures.

2.9 Ergonomics of the Lexile Framework

The last section reviewed the evidence supporting the validity of the Lexile Framework. The next section looks at how measurement error propagates

TABLE 3

Comprehension Rates for the Same Individual With Materials of Varying Comprehensibility

Person Measure	Text Calibration	Sample Titles	Forecast Comprehension
1000	500	Are You There God? It's Me Margaret - Blume	96%
1000	750	The Martian Chronicles - Bradbury	90%
1000	1000	Reader's Digest	75%
1000	1250	The Call of the Wild - London	50%
1000	1500	On the Equality Among Mankind - Rousseau	25%

TABLE 4

Comprehension Rates of Different Ability Person With the Same Material

Person Measure	Calibration for Sports Illustrated	Forecast Comprehension Rate
500	1000	25%
750	1000	50%
1000	1000	75%
1250	1000	90%
1500	1000	96%

through the Framework and influences the certainty with which texts and readers are located on the Lexile Scale. Intertwined with and influenced by the reliability and validity of the framework is the issue of how well fitted the product is to the form and function required by the user community, i.e. the product's ergonomics. An assessment framework may produce reliable and valid measures and still fail to achieve its promise due to inattention to "soft" product features like beauty, accessibility, believability, and extensibility. At present the evidence supporting the Framework's ergonomic fitness is largely anecdotal but these anecdotes serve to add context to the validity coefficients and standard errors of measurement reported in the adjoining sections.

In addition to ordering texts in a manner consistent with how basal publishers order selections and how test publishers assign calibrations to test items, the Lexile Framework orders literature titles in a way consistent with teacher judgments about which titles are read at different grade levels. Scholastic Publishing, Inc. reported Lexile text measures for many of the books offered in their 1997 catalog and employed the Framework to link products developed in different divisions of the corporation (Scholastic Catalog, 1997). The North Carolina State Administration for School Libraries has begun publishing Lexile text measures for all titles approved for purchase by media coordinators (Infotech). These independent applications of the Lexile Framework show that the Framework orders text in a way found useful by diverse user groups.

Reader measures produced by the Framework have been found to correlate in expected ways with age, grade, sex, SES, grade attained, books in the home, income, occupations, and many other demographic variables. Lexile reader measures have been found to correlate with other reading tests as highly as parallel forms of those tests correlate among themselves. Attempts to isolate second and third dimensions of reading have been singularly unsuccessful (Zwick, 1984). Finally, teachers report that Lexile measures order students in ways consistent with reading group formation and their judgments about reading proficiency.

There are dozens of readability equations that can be used, almost interchangeably, to order the comprehensibility of text. Likewise, there are hundreds of reading tests that order readers. What distinguishes the Lexile Framework is its ability to conjointly order texts and readers on the same scale. The ability to characterize a reader as 1000L and a text as 1000L enables a forecast of the comprehension rate that that reader will have with

that particular text. The difference between an absolute measure of the text and of the reader is used to forecast the relative construct called "comprehension". Comprehension, itself, is not an absolute; rather it is the consequence of an encounter between a reader and a text.

The subjective experience of 50%, 75%, and 90% comprehension as reported by young readers varies greatly. A 1000L reader of 1000L text (75% comprehension) reports confidence and competence. Teachers listening to such a reader report that the reader can sustain the meaning thread, and reads with motivation, appropriate emotion and emphasis. In short, such readers sound like they comprehend what they are reading. A 1000L reader of 1250L text (50% comprehension) encounters sufficient unfamiliar vocabulary and syntactic structures that the meaning thread is frequently lost. Such readers report frustration and will seldom choose to read independently at this level of comprehension. Finally, a 1000L reader of 750L text reports total control of the text, reads with speed and appears automatic. Dick Woodcock is linking the Woodcock Johnson-Revised (1995) to the Lexile Framework and as part of this study is looking at the kinds of errors that readers make when comprehending at 50%, 75%, and 90% rates.

A primary utility of the Lexile Framework is in large measure its ability to forecast what happens when readers confront text. With every application by teacher, student, librarian or parent, there is a test of the Frameworks accuracy. The Framework makes a point prediction every time a text is chosen for a reader. The anecdotal evidence suggests that the Framework works as intended. That is not to say that there is an absence of error in forecasted comprehension. There is error in text measures, reader measures, and their difference modeled as forecasted comprehension. However, the error is sufficiently small that our judgments about readers, texts, and comprehension rates are useful.

3 MEASUREMENT ERROR

Measurement is the process of converting observations into quantities via theory. Repeated observation of what is intended to be the 'same thing results in a series of non-identical numbers. When these observations (say, counts correct for a person on various reading comprehension tests) are converted into quantities (e.g., Lexile measures) via calibrations provided by a theory

(e.g., Lexile Framework), the resulting quantities are distributed about a mean that is taken to be the measure of the person's reading comprehension. The standard deviation of this distribution divided by the square root of the number of measurements is the standard error of measurement, (SEM). Each measure resulting from repetition of a measurement procedure is assumed to be exchangeable with any other of the possible measures that might have been made. In this approach there is no a priori reason for favoring a measure obtained on Monday versus Tuesday, or one based on multiple choice versus constructed response. Any measure from this class of measures is specified to be equivalent. Asserting the exchangeability of a defined class of measures does not imply that any measure at all will do, only any measure from the defined class. The arithmetic mean of a sample of measures approaches a limit as the sample size increases. The "closeness together" of these measures, expressed as a standard deviation, is an index of uncertainty regarding the magnitude of the quantity.

In practice we rarely have available large numbers of measurements on each object of measurement. Typically, there is only one measure. But there is still the necessity to attach some estimate of uncertainty to that measure. Seventy years of psychometric research has yielded dozens of proposed solutions in the form of reliability coefficients and associated SEMs. Reliability coefficients, despite their well known sample dependencies, are defended on the basis that they represent a "unitless" measure of precision that can be compared across scales (Note 1). The SEM is defended for its stability over samples but is, of course, scale dependent. We find the legion of reliability coefficients and SEMs deficient in one or more of the following ways, they: (1) do not correctly model error variation due to methods and moments (e.g., retest coefficients and KR-21), (2) ignore instrument/item main effects that are "error" in Rasch- and theory-referenced measurement applications where *absolute* rather than relative scale location is the focus (e.g., coefficient alpha), (3) yield an average group statistic that is uncritically attached to each individual's measure (e.g., alternate forms coefficients), and (4) ignore moment to moment variation. Generalizability theory (Brennan, 1980) remedies some of these deficiencies; but as applied still suffers from ailment (3). What is needed is an approach to measurement error that (a) admits all species of intra-individual method and moment variation, (b) can be calculated from a single response pattern without recourse to data on other instruments or persons, and (c) yields a SEM that is easy to understand and calculate.

Resampling theory meets the above objectives. Measures that result from a theory referenced measurement model are "generally objective", i.e., absolute measures are independent of the instrument used. Thus, a person's 50 item response pattern can be treated as a personal item bank. One thousand "replicates" of 50 items each can be sampled *with replacement* from this response record. The standard deviation of this 1,000 measure distribution is a standard error of measurement which meets the three criteria above.

Another way to look at a resampled SEM is as an answer to the laymen's question, "What would happen, if we did it again?" The result of doing it again and again is a distribution of measures the standard deviation of which is the standard error of measurement. Time and cost always limit our ability to measure again. But we still want to describe our uncertainty about a measure's long run stability. The best evidence of what would happen if we measure again is gained by "measuring" again and again with the data in hand. Resampling enables us to simulate "measuring again" and to measure the dispersion in the replicated measures resulting from the resampling procedure. The more dispersion or spread in our replicates, the more uncertainty we have about the measures long run value.

Here is an example that clarifies how the resampling procedure works and shows its sensitivity to theory misfit and mistargeting. Table 5 presents several response patterns to 10 Lexile reading comprehension items uniformly distributed over a 5.5 logit test width. The first 3 people get 5 items correct and a measure of 813L, but the misfit increases as we move across the table. Person 3 has missed easier questions and answered harder questions correctly compared to persons 1 and 2. Note how the SEM increases as theory misfit increases. As theoretical expectations and what is actually observed diverge, the caution index (SEM) increases.

The last two people score 8 correct and are each assigned a measure of 1238L. Since these individuals are less well targeted than the first three, we expect to see increases in the SEMs and we do. Again, theory misfit contributes to the size of the SEM because person 5 has greater misfit and a higher SEM than person 4. Error variance accumulates with each decrease in targeting efficiency and with each increase in theory misfit. This example is too small to illustrate how other sources of method and moment variance increase the intra-individual variance, but the general idea should be clear. The lesson is: one cannot estimate what one does not replicate. So design into the observation model variation in method and moment facets

Table 5: Measures and SEM's for Five Hypothetical Patterns

Item Calibrations	Person1	Person2	Person3	Person4	Person5
250L	1	1	1	1	1
375L	1	1	0	1	1
500L	1	0	1	1	1
625L	1	1	0	1	0
750L	1	1	1	1	1
875L	0	0	0	1	1
1000L	0	0	1	1	1
1125L	0	0	0	1	1
1250L	0	1	1	0	0
1375L	0	0	0	0	1
Person Measure	813L	813L	813L	1238L	1238L
SEM	113L	216L	256L	413L	496L

that are considered important, then resample over these facets consistent with your definition of "do it again". The standard deviation of the resulting distribution describes uncertainty about the measure more completely than conventional approaches to measurement error.

This perspective on error, which treats uncertainty as the dispersion observed in resampled replicate measures, forces us to make explicit what we mean by "measure again". For example, what facets of the process of measurement are expected to vary with each prospective replication (in ANOVA terms these facets are "random") and what facets are expected to stay the same with each replication (in ANOVA terms such facets are "fixed")? There is no one "right" answer to the question of how we decide to define "do it again" (i.e., measure again). In some research contexts it may be useful to treat test items as fixed (e.g., when linking one instrument to another) whereas in most applications items would be "random". In psychological research on state anxiety, day-to-day fluctuations in measures are represented as construct variance, whereas, in studies of trait anxiety, day-to-day fluctuations are treated as error. In general, there is no "cookbook" for defining the resampling design that is "best" for describing uncertainty. Your definition of "do it again" must be made explicit and then the resampling process must be executed in each measurement application in conformance with that

definition.

With the above perspective on measurement error as a foundation, we turn to a discussion of four kinds of error that arise in the use of the Lexile Framework: text measure error, reader measure error, error in forecasted comprehension rate error, and error in linking tests to the Lexile Framework.

3.1 Text Measure Error

When determining the Lexile measure for a book or manual, the standard procedure is to sample 20 pages randomly from the work. These pages are concatenated into a text file that is passed to a software package called the Lexile Analyzer. The analyzer "slices" the text file into as many 125 word passages as possible, and passes the set of slices through an analysis process that calculates a Lexile calibration for each slice. That set of calibrations is then passed to an equation that solves for the Lexile measure corresponding to a 75% comprehension rate. The analyzer uses the slice calibrations as test item calibrations and then solves for the measure corresponding to a relative raw score of 75% (e.g., 30 out of 40 correct, as if the slices were test items). Obviously, the measure corresponding to a relative raw score of 75% on *Goodnight Moon* (300L) slices would be lower than the measure corresponding to a comparable raw score on *USA Today* (1080L) slices. The Lexile Analyzer automates this process and thousands of books have been measured in this way. But what "certainty" can we attach to these text measures?

Our perspective on assessing uncertainty (i.e., error) requires an answer to the question "what would happen if we measured again?" We could measure again by sampling another 20 pages and repeating the above analysis. The result would be a text measure which differs from the first. We could repeat the sampling process over and over until we exhausted available time and resources and then take the standard deviation of the resulting distribution of text measures as our standard error of measurement. But what if resource constraints dictate that we can only sample 20 pages resulting in, say, 49 calibrated slices. Is there a way to simulate the act of continuing to repeat the measurement process a large number of times. The answer is "yes". We can use resampling methods to simulate repeated measurements. This works as follows: sample with replacement 49 calibrations from the set of 49 slice calibrations and solve for the measure. Because the resampling is done

Table 6: Standard Errors for Selected Text Measures

TITLE	Number of Slices	Text Measure	Resampled Mean	S.E.
Equality Among Mankind	80	1501	1501	32
Ivanhoe	92	1427	1426	26
David Copperfield	130	1196	1197	29
Swiss Family Robinson	44	1167	1168	28
Treasure Island	25	1081	1081	75
The Hobbit	85	1068	1068	24
Dr. Zhivago	48	1031	1026	43
20,000 Leagues Under the Sea	37	990	988	37
The Old Man and the Sea	55	905	908	50
Little House on the Prairie	67	754	753	29
Encyclopedia Brown	25	634	632	34
It's Me, Margaret	29	511	509	37
$N_r = 1000$				

with replacement, the resampled 49 slices will differ from the original set of 49 because the "with replacement" feature insures that some slices will be sampled more than once in some replicates and not at all in others.

Each replication results in a "replicate" text measure. The standard deviation over, say, 1000 replicate measures is the standard error of measurement and describes the uncertainty with which we locate a title in the Lexile Framework. Table 6 presents standard errors (SE) for a group of well known titles. The standard errors vary from a low of 26L to a high of 75L. Most text measures in the Lexile Library have standard errors from 30L to 40L (Notes 2 and 3).

3.2 Reader Measure Error

What do we mean by "do it again" when measuring reader performance? Measuring again implies a different set of items (method) on a different occasion (moment) meaning that method and moment are random facets and are expected to vary with each replication of the measurement process. With

this definition of a replication there is nothing special about one particular set of items or test, nor is there anything special about one particular Tuesday morning. Any calibrated set of items given on any day within a two-week period is considered exchangeable with any other method and moment. By "exchangeable" we mean that we have no a priori basis for believing that one particular method-moment combination will yield a higher or lower measure than any other. That is not to say that we expect resulting measures to be the same. On the contrary we expect them to be different. We just don't know which method-moment combination will produce lower measures and which higher. The anticipated variance among replications due to methods/moments and their interactions is error. A better understanding of how these sources of error come about can be gained by describing some of the behaviors and measurement contexts that may vary from replication to replication.

Suppose that most of the items used to measure Sally are sampled from books in the "Baby Sitter" series and that this is Sally's favorite series. When Sally is measured again, items are sampled from less familiar texts. The differences in Lexile measures coming from highly familiar and unfamiliar texts would be error. Now suppose that the particular response format used for all items administered to Sally results in slightly higher calibrations than other item formats and that this slight advantage is constant for all items of this type. This constant main effect for items also contributes to error in measuring Sally's reading performance.

Characteristics of the moment and context of measurement can contribute to variation in replicate measures. Suppose, unknown to the test developers, that measures go up with each replication because of practice effects. This "occasion main effect" also would be treated as error. Suppose Sally is fed breakfast and rides the bus on Tuesdays and Thursdays, but on Monday, Wednesdays, and Fridays her parent has early business meetings and she gets no breakfast and must walk one mile to school. Some of the test replications are given on what for Sally are "good days" and some are given on "bad days". Variation in her reading performance due to these context factors contribute to the error. Yet another source of error arises if a particular kind of item, say, gets easier, relative to other items, as readers get more practice with it. This item by occasion interaction contributes to error.

Familiar reliability coefficients and SEMs do not reflect the uncertainty in reader measures that arise from all of the sources described above. IRT model

errors, equivalence coefficients, stability coefficients, and alternate forms coefficients all underestimate the reader measure error under the Lexile Framework. The Lexile Framework produces absolute measures and as such, treats as error sources of variance that a relative measurement model either ignores or treats as construct (i.e., wanted) variance.

The best approach to attaching uncertainty to a readers measure is to resample the item response record, i.e., simulating what would happen if we actually measured again. Suppose 10 year old Jose takes two 50 item reading tests one week apart. Occasions and the 50 items nested within occasion can be independently resampled (two stage resampling) and the resulting two measures averaged for each replicate. One thousand replications would result in a distribution of replicate measures. The standard deviation of this distribution is the resampled SEM and it describes uncertainty in Jose's reading measure under a definition of "do it again" that treats methods (items), moments (occasion and context), and their interactions as error. Furthermore, in computing Jose's reading measure and the error in that measure he is treated as an individual without reference to other people's performance.

3.3 Comprehension Forecast Error

The difference between a text measure and a reader measure can be used to forecast the readers comprehension with that text. If an 1100L reader reads *USA Today* (1100L) the Framework forecasts 75% comprehension. This forecast means that if an 1100L reader takes 100 Lexile test items taken from *USA Today* the count correct is estimated to be 75 or 75% of the items taken. The same 1100L reader is forecast to have 50% comprehension of freshman college texts (1350L) and 90% comprehension of *Sounder* (830L). How much error is there in such a forecast? That is, if we made the forecast again what kind of variability in the comprehension rate would we expect to observe.

How do we define "do it again" when we are talking about comprehension rate? Comprehension rate is determined from the reader measure and the text measure. Consequently, error variation in comprehension rate derives from error variation in those two quantities. When we "do it again" we expect to sample another 20 pages from the title in question, and we expect to test the reader again. The result is a new text measure and a new reader measure which combine to forecast a new comprehension rate. Thus, errors

in reader measure and text measure combine to generate variability in the replicated comprehension rate. This kind of replication can be simulated by resampling a text measure replicant and a reader measure replicant that combine to forecast a comprehension rate replicant. Repeating this resampling procedure, say, 1,000 times will yield 1,000 comprehension rates that can be used to build a confidence interval around the mean comprehension rate. Unlike text and reader error, the comprehension rate error will not be symmetrical about the forecasted comprehension rate.

3.4 Linking Standard Errors

A linking study results in a table with three columns. Column 1 includes all possible scale scores on the target test; Column 2 reports the Lexile equivalent for each scale score, and Column 3 gives the linking standard error (LSE) in Lexiles for the scale score to Lexile conversion. Table 7 presents the results from linking the North Carolina End of Grade (NCEOG) test to the Lexile Framework. The linking standard error describes the expected variation in Lexiles associated with repeating the linking study a large number of times. If a linking study produces a correspondence between target scale score and Lexile of 158:980L, then each person scoring 158 on the target test would be assigned a Lexile measure of 980L. The correspondence is symmetrical in that a 980L on the Lexile scale corresponds to a 158 on the target test. Thus, target test scores are converted to Lexile measures (and back again) in the same way that Fahrenheit temperatures are converted into Celsius temperatures (and back again).

The equation for converting target scores to Lexile measures is based on a linear linking design. Each of the 956 students in the study took the NCEOG and a Lexile test of comparable length. Some of the students took the NCEOG first and others took the Lexile test first. Less than two weeks separated the two test administrations. The NCEOG score (transformed 3p IRT measure) and the Lexile measure (transformed 1p Rasch measure) were plotted and an *sd* line (geometric mean of the two regressions) was fitted to the data (Figure 2). The equation for the *sd* line was used to build the correspondence Table 7. The procedure for computing the standard error for each correspondence of NCEOG scale score and Lexile measure is described below.

A linking standard error is an answer to the question "what would happen

TABLE 7

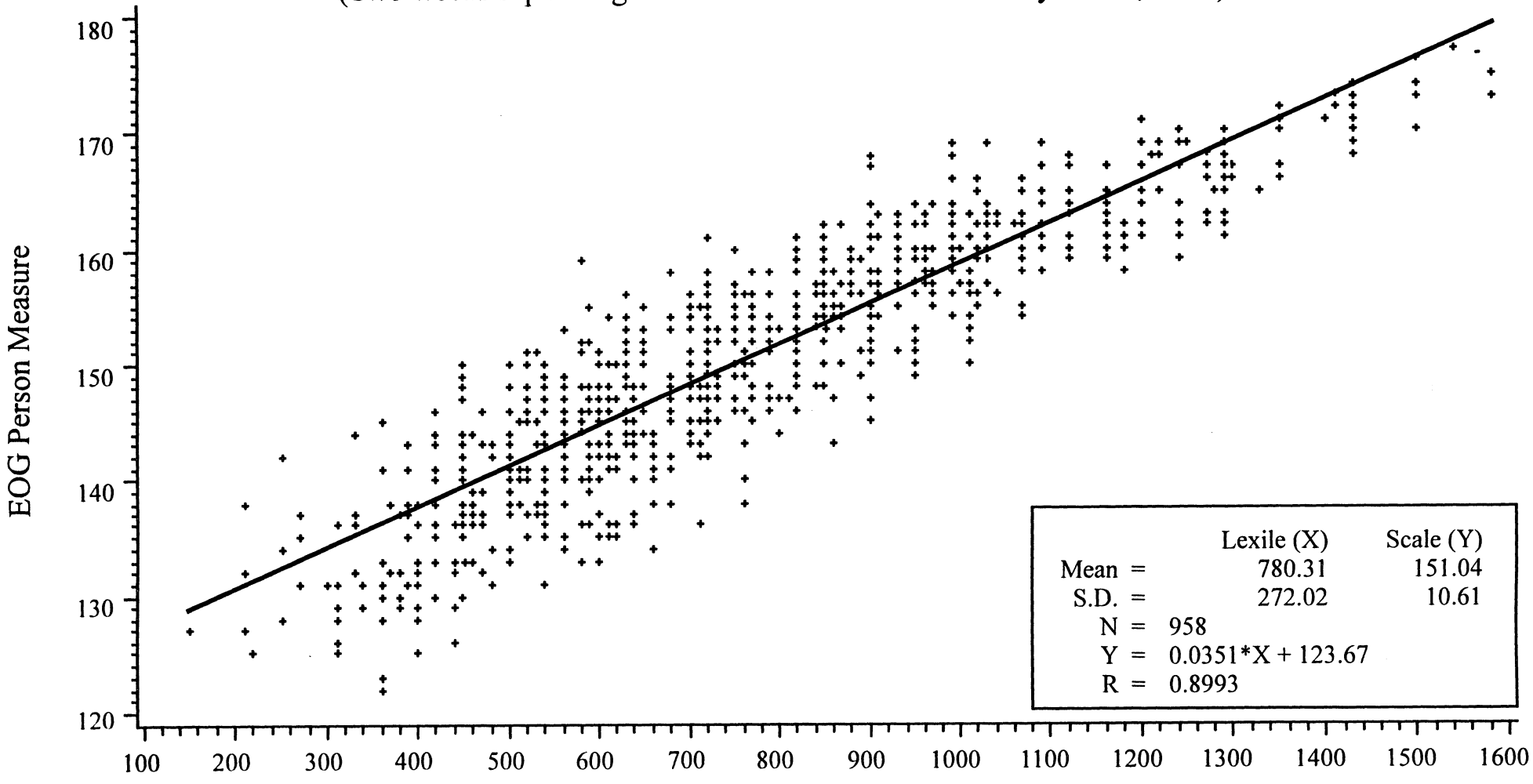
N. C. End of Grade
Scale Score to Lexile Conversion Table
Grades 3 through 8

Scale Score	Lexile	Std. Error	Scale Score	Lexile	Std. Error
130	240	18.3	153	830	6.6
131	265	17.6	154	855	6.8
132	290	16.8	155	880	7.1
133	315	16.1	156	910	7.5
134	345	15.4	157	935	7.9
135	370	14.6	158	960	8.4
136	395	13.9	159	985	8.9
137	420	13.2	160	1010	9.4
138	445	12.5	161	1035	10.0
139	470	11.9	162	1060	10.7
140	500	11.2	163	1090	11.3
141	520	10.6	164	1110	12.0
142	550	10.0	165	1140	12.6
143	575	9.4	166	1165	13.3
144	600	8.8	167	1190	14.0
145	625	8.3	168	1215	14.7
146	650	7.8	169	1240	15.5
147	675	7.4	170	1265	16.2
148	700	7.1	171	1290	16.9
149	730	6.8	172	1317	17.7
150	755	6.6	173	1345	18.4
151	780	6.5	174	1370	19.2
152	805	6.5	175	1395	19.9

Figure 2

Analysis of NC End of Grade vs. Lexile Measures For All Grades

(Two weeks separating the two test administrations: May - June, 1995)



Legend: +++ Measures ——— Regression Line

if we linked again?" The best evidence of what would happen if we linked again is to link again with the data in hand. The definition of "do it again" used in Table 7 assumes that different persons and different Lexile test items would be used in each replication of the linking design. Thus. persons and Lexile items are random (resampled), but NCEOG items are fixed (not resampled). The NCEOG scale scores are treated differently from the Lexile scores because of an inherent difference in the nature of the two scales. The Lexile Framework is a free-standing criterion-referenced scale that is not tied to any particular test instrument. In contrast the NCEOG scores are tied to the instrument used in the linking study.

The resampling procedure for computing a LSE in conformance with the above definition of "do it again" would proceed as follows:

1. Sample *with replacement* 958 persons from the 958 person data set. For each person, resample his/her Lexile response record and compute a replicate Lexile measure. Use the person's NCEOG measure without resampling. If a person appears six times in a replicated data set, he/she will have a different Lexile measure each time but the same NCEOG.

2. Plot the NCEOG scale score and the resampled Lexile measure for the 958 resampled persons.

3. Compute the *sd* line and build the table of correspondence between NCEOG and Lexile measures.

4. Repeat steps 1–3 ninety-nine more times.

5. Compute the standard deviation of the 100 Lexile measures corresponding to each NCEOG scale score and report this standard deviation as the linking standard error (LSE).

A large LSE undermines our confidence in the correspondence between target scale scores and Lexile measures. When, however, the correspondence is generalizable over items, persons and occasions, we move one step closer to a common framework for reading measurement and instruction.

3.5 How Errors Combine

When native Lexile items are used to measure readers, the reader error, previously discussed, correctly describes uncertainty about a reader's measure. When, however, a test using a non-native item format (e.g., NCEOG or ITBS) is linked to the Lexile Framework, reader error must be inflated by the linking error. The previously discussed reader error is combined with the linking error according to the square root law:

$$\text{Total Reader Error} = \text{SEM}_r = \sqrt{\text{SEM}^2 + \text{LSE}^2} \tag{7}$$

Thus, the nominal reader error SEM is inflated by the corresponding linking error whenever a linked test is used to generate a reader measure. When linking is involved, reader error will be specified to include linking error as in Equation 7 above.

As an example suppose a fifth grader scores 158 on the NCEOG. Assume the SEM for this score to be 2.73. To convert this SEM to Lexiles we must multiply by (272.02/10.61) [See Figure 2] yielding 70L. Using Table 7, we look up the corresponding Lexile measure of 980L and find the LSE to be 12L. Applying Equation 7, we obtain

$$\text{SEM}_r = \sqrt{70^2 + 12^2} = 71L.$$

Note that reader error combines intra-individual variance due to method and moment of measurement with linking variance.

The principle of combining sources of error can also be applied to assess the error in a forecasted comprehension rate. Since comprehension rate for an encounter between reader and text is determined by the reader measure and the text measure, errors in either of those measures will cause error in the comprehension rate. Since comprehension rate is a non-linear function of the difference between reader measure and text measure, we cannot use a square-root law to obtain a standard error for the comprehension rate. We can, however, use a square-root law to calculate a standard error for the difference(SED), which can then be used to calculate a confidence interval for the comprehension rate.

The standard error for the difference is computed as follows:

$$\text{SED} = \sqrt{\text{SEM}_r^2 + \text{SEM}_t^2}, \tag{8}$$

where SEM_r = Reader error (Intra-individual + Linking) and SEM_t = Text measure error.

Suppose a 1000L reader measured with 71L of error reads a 1000L text measured with 30L of error. What is the 90% confidence interval about the 75% forecasted comprehension rate? The answer is that the error of the difference (712 L + 302 L) is 77L. Table 8 can be consulted for an estimate of the 90% confidence interval about a 75% comprehension rate, given that the error of the difference is approximately 80L (rounding up from 77L to the nearest tabled value). The 90% confidence interval for the comprehension rate is 63% to 84%.

In summary, there are three kinds of error of which users of the Lexile Framework should be aware:

- Reader error describes uncertainty in the location of a reader on the Lexile Map. Reader error reflects the fact that measuring again with different items on different occasions with tests that are linked to the Lexile Framework would result in a series of non-identical measures with a standard deviation equal to the SEM_r.

- Text error describes uncertainty in the location of a title on the Lexile Map and reflects the fact that repeatedly drawing 20 page samples from that title would result in a series of text measures with a standard deviation equal to SEM_t.

- Error in forecasted comprehension rate, which is a consequence of error in the difference between reader measure and text measure. Because of the nonlinear asymmetric relationship between comprehension rate and the reader-text difference, this error is expressed in terms of confidence intervals.

TABLE 8

90% Confidence Intervals for Various Combinations of Comprehension Rate and Error of Difference Between Reader and Text Measures

Reader ^{L} - Text ^{L}	Forcasted Comprehension Rate	Error of Difference				
		40	60	80	100	120
-250	50%	43-57	39-61	36-64	33-67	30-70
-225	53%	46-60	42-63	38-67	35-70	32-73
-200	55%	48-62	45-66	41-69	38-72	34-75
-175	58%	51-65	47-68	44-71	40-74	37-77
-150	61%	54-67	50-71	47-73	43-76	39-79
-125	63%	56-70	53-73	49-76	46-78	42-81
-100	66%	59-72	56-75	52-78	48-80	45-82
-75	68%	62-74	58-77	55-79	51-82	48-84
-50	71%	64-76	61-79	57-81	54-83	50-85
-25	73%	67-78	64-81	60-83	57-85	53-87
-0	75%	69-80	66-82	63-84	59-86	56-88
+25	77%	72-82	68-84	65-~~66~~86	62-87	58-89
+50	79%	74-83	71-85	68-87	64-89	61-90
+75	81%	76-85	73-87	70-88	67-90	64-91
+100	82%	78-86	75-88	72-89	69-91	66-92
+125	84%	80-87	77-89	74-90	72-91	69-93
+150	85%	81-89	79-90	77-91	74-92	71-93
+175	87%	83-90	81-91	78-92	76-93	73-94
+200	88%	84-91	82-92	80-93	78-94	76-95
+225	89%	86-92	84-93	82-94	80-94	77-95
+250	90%	87-92	85-93	83-94	81-95	79-96

Notes:

1. Why do we need scale free indices of uncertainty? Because we are awash in multiple scales for the same construct (e.g., there are several hundred scales for reading comprehension). When we realize that theory-referenced measurement can enable us to adopt a single scale (but continue to use the two hundred different tests if we desire) we will be less attracted to reliability as a "unitless" index of uncertainty and instead will embrace, as do all other sciences, the standard error of measurement.

2. If the research focus is on disattenuating correlations then reliabilities based on relative error models are the best choice. However, see Schmidt and Hunter (1996) for a discussion of the ways that the wrong relative error model is often used to disattenuate correlations.

3. A more conventional approach to computing a SEM for these twelve titles might proceed as follows: Compute measures separately for the odd and even numbered slice calibrations; correlate the odd measures and even measures over the twelve titles to estimates the reliability (r_{yy}); multiply the standard deviation of the title measures by $\sqrt{1 - r_{yy}}$. The resulting standard error of measurement (SEM) is used to describe the "typical" uncertainty in the text measures.

   There are two reasons why the conventional approach does not work for estimating uncertainty in Lexile text measures. First, the Lexile Framework is an absolute measurement model meaning that SEMs computed using relative error models (e.g., Cronbach's alpha, KR-20, split half, test-retest, and alternate forms coefficients) underestimate the error and therefore overestimate the reliability coefficient and consequently overestimate the certainty with which titles are measured. Secondly, conventional approaches to reliability average variances and co-variances over objects of measurement (in this case book titles) and thus are best viewed as group statistics. When the focus is on the individual case, whether it be an individual book title or reader, these group averages are very crude guides to the uncertainty expected in a particular text measure or reader measure. A cursory glance at Table 6 should make clear that a summary SEM is a poor substitute for individual SEMs (Note 2).

REFERENCES

Anderson, R. C. and Davison, A. (1988). Conceptual and empirical bases of readability formulas. In A. Davison and G. M. Green (Eds.), *Linguistic complexity and text comprehension: Readability issues reconsidered.* Hillsdale, NJ: Erlbaum.

Anderson, R. C., Hiebert, E. H., Scott, J. A., and Wilkinson, I. (1985). *Becoming a nation of readers: The report of the commission on reading.* Washington, DC: U. S. Department of Education.

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

California Achievement Test: Form C (1977). New York: McGraw-Hill.

California Achievement Test: Form E (1985). New York: McGraw-Hill.

Carroll, J. B. (1980). Measurement of abilities constructs. In U. S. Office of Personnel Management, *Construct Validity in Psychological Measurement.* Princeton, NJ: Educational Testing Service.

Carroll, J. B., Davies, P., and Richman, B. (1971). *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.

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.

Comprehensive Test of Basic Skills: Form U (1981). New York: McGraw-Hill.

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

Crawford, W. J., King, C. E., Brophy, J. E. (1975). *Error rates and question difficulty related to elementary children's learning*. Paper presented at the annual meeting of the American Educational Research Association, Washington, D. C.

Crick, J. E. and Brennan, R. L. (1982). *GENOVA: A generalized analysis of variance system* [computer program]. Dorchester, MA: University of Massachusetts at Boston.

Davidson, A. and 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.

Dunn, L. M. and Dunn, L. M. (1981). *Peabody Picture Vocabulary Test-Revised: Forms L and M.* Circle Pines, MN: American Guidance Service.

Dunn, L. M. and Markwardt, F. C. (1970). *Peabody Individual Achievement Test.* Circle Pines, MN: American Guidance Service.

Electronic Encyclopedia (1986). Danbury, CT: Grolier.

Hahn, A. (1987). Reaching out to America's dropouts: What to do? *Phi Delta Kappan*, 67, 256-263.

Hitch, G. J. and Baddeley, A. D. (1974). Verbal reasoning and working memory. *Journal of Experimental Psychiatry*, 28, 603-621.

Horabin, I. (1989). *TestCalc* [computer program]. Durham, NC: Ivan Horabin.

Horabin, I. (1989). *TestCalc* [computer program]. Durham, NC: MetaMetrics.

Horabin, I. (1987). PC-LEX: A computer program for rating the difficulty of continuous prose in Lexiles [computer program]. Durham, NC: MetaMetrics.

Klare, G. R. (1974). Assessing readability. *Reading Research Quarterly*, 1, 63-102.

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

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

Lord, F. M. (1980). Applications of item response theory to practical testing problems. Hillsdale, NJ: Erlbaum.

Luce, R. D. and Tukey, J. W. (1964). Simultaneous conjoint measurement: A new type O fundamental measurement. *Journal of Mathematical Psychology*, 1,1-27.

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

Mitchell, J. V. (1985). *The Ninth Mental Measurements Yearbook.* Linclon, Nebraska: University of Nebraska Press.

National Assessment of Educational Progress (1984). Princeton, NJ: Educational Testing Service.

Rasch, G. On Objectivity and Specificity of the Probabalistic Basis for Testing, mimeographed, no date, 1 -19.

Rasch, G. A. (1968). Mathematical theory of objectivity and its consequences for model construction. In report from *European Meeting on Statistics, Economics, and Management Sciences*, Amsterdam.

Rasch, G. (1980). *Probabalistic Models for Some Intelligence and Attachment Tests*. Chicago: The University of Chicago Press (first published in 1960).

Readability Calculations [computer program] (1984). Dallas, TX: Micro Power and Light Company.

Thorndike, R. L. (1949). *Personnel Selection*. New York: Wiley.

Thurstone, L. L. (1925). A method of scaling psychological and educational tests. *Journal of Psychology*, October 1925.

Thurstone, L. L. (1928). Attitudes can be measured. *American Journal of Sociology*, 33, 529-544.

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

Stanley, J. C. (1971). Reliability in R. C. Thorndike (Ed.) *Educational Measurement*: 2nd edition. Washington, D. C.: American Council on Education.

Stenner, A. J. General objectivity, *Transactions of the Rasch Measurement SIG*, 3, 1.

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

Stenner, A. J., Smith, D. R., Horabin, I., and Smith, M. (1987). Fit of the Lexile Theory to Item Difficulties on Fourteen Standardized Reading Comprehension Tests. Durham, NC: MetaMetrics.

Squires, D. A., Huitt, W. G., and Segars, J. K. (1983). *Effective Schools and Classrooms*. Alexandria, VA: Association for Supervisor and Curricular Development.

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

White, E. B. (1952). *Charlotte's Web*. New York: Harper and Row.

Wright, B. D. (1968). Sample-free test calibration and person measurement. In proceedings of the 1967 *Invitational Conference on Testing Problems*. Princeton, NJ: Educational Testing Service.

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

Wright, B. D. and Stone, M. (1991). *Objectivity: Measurement Primer No. 2*. Wilmington, DE: Jastak Associates.

Zwick, Rebecca (1983-84). *The NAEP 1983-84 Technical Report.* National Assessment of Educational Progress.