Measuring Reading Comprehension with the Lexile Framework

A. Jackson Stenner
MetaMetrics, Inc.

Paper Presented at the
California Comparability Symposium
Burlingame, CA
October 31, 1996

METAMETRICS, INC.
2327 Englert Drive, Suite 300
Durham, NC 27713
PHONE (919) 547-3400 FAX (919) 547-3401

Objectivity and the Idea of Measurement

Implicit in the idea of measurement is the concept of objectivity. When we measure the temperature using a thermometer, we assume that the measurement we obtain is not dependent on the conditions of measurement, such as which thermometer we use. Any functioning thermometer should give us the same reading of, for example, 75 degrees Fahrenheit. If one thermometer measured 40 degrees, another 250 and a third 150, then the lack of objectivity would invalidate the very idea of accurately measuring temperature.

It is this general objectivity that distinguishes physical science measurement from behavioral science measurement. General objectivity requires a construct theory embodied in a specification equation that is capable of estimating indicant calibrations. When these theory-based calibrations are employed in the Rasch model, observations (i.e., raw scores) can be converted into measures without relying on individual or group data on indicants (e.g., items) or objects of measurement (e.g., persons). The benefits of these methods include: (1) the construct theory is exposed to falsification, (2) it is possible to build correspondence tables between observations and measures with recourse only to theory, (3) a generalized linking solution is available for placing observations of all kinds on a common scale, (4) a reproducible unit of measurement can be developed, (5) the framework for fit statistics that is sample-dependent under the Rasch model becomes sample-independent, and (6) a complete frame of reference for measure interpretation can be constructed.

This paper shows how the concept of general objectivity can be used to improve behavioral science measurement, particularly as it applies to the Lexile Framework, a tool for objectively measuring reading comprehension. We begin with a dialogue between a physicist and a psychometrician that details some of the differences between physical science and behavioral science measurement. Building on these distinctions, we offer a definition of measurement that describes what goes on in the physical sciences and that represents an attainable ideal of what should go on in the behavioral sciences. This definition of measurement is formalized in an equation that turns out to be the Rasch model, with the important difference that indicant calibrations are obtained via theory, not data. Through the use of theory-based calibrations, we achieve a generally objective estimation of the measure parameter in the Rasch model. The next section of the paper examines the differences between local objectivity obtained with the Rasch model and general objectivity obtained with a theory-enhanced version of that model. Next, we report on a 10-year study of reading comprehension measurement that implemented the concept of general objectivity through the development of the Lexile Framework. Finally, we summarize several of the benefits of objective measurement and general objectivity as they might be realized in the measurement of constructs other than reading comprehension.

The Problem of General Objectivity: A Dialog Between the Behavioral and Physical Sciences

Psychometrician:

As I understand our purpose here today, each of us will solve a measurement problem specific to our discipline, and then discuss our respective methods. Since I opened the dialog, I will present you with your problem first.

In the box in front of you is a long thin glass tube open at one end, some mercury, and a ruler. Your task is to answer the following question: 'What change in temperature is indicated by an increase of five centimeters in column height of the mercury?'

Physicist:

I have completed my analysis and found that the answer is 18 degrees Fahrenheit.

My method was straightforward. I calculated the volume of mercury by first calculating the capacity of the glass tube and then estimating the fraction of that capacity occupied by the mercury. Next, I consulted the table of expansion coefficients and found the equation for mercury. Given the observation (column height of mercury), I set up an equation and solved for the measure (i.e., temperature) corresponding to each of two volumes of mercury separated by the indicated five centimeters. For those of us trained in physics and chemistry, this is a standard measurement problem. Through theory, a context is created for expressing an observation as a measure. The theory enables me to extract the relevant information from the context surrounding an observation (what some call an observation model) and use this information to convert observations to measures. The theory specifies what is essential to record about the context of an observation. All other features are either ignored as irrelevant (e.g., the time of day the thermometer is used) or are considered during the process of observation (e.g., removing the potential contaminating effects of barometric pressure by sealing the top of the thermometer).

Now here is your problem: In front of you is a newspaper, *USA Today*. Your task is to construct a 50-item reading comprehension test and compute the increment in reading comprehension that would be reflected in an increase from thirty to forty correct responses on the test. You may consult any other references you choose.

Psychometrician:

I spent the first few hours generating the 50-item measure and the last two hours trying to understand how you were able to solve the temperature problem while I cannot begin to imagine how to solve the reading comprehension problem. If given more time, I would administer the test to, say, 200 high school seniors, analyze the data with an appropriate measurement model (e.g.,

the Rasch model) and report the difference between 60 percent correct and 80 percent correct on some appropriately transformed logit scale. I can see, however, that this is a very different process from the one you used in solving the temperature problem.

What I concluded is that I lack a theory and associated equations for transforming observations into measures. Where possible, I attempt to make up for this lack of theory by basing instrument calibrations on data instead of theory. Once I frame the problem in this way, it is clear that all I would need to accomplish what you did with the temperature problem is a good reading comprehension theory and associated equations for calibrating the items. Then I could solve the reading comprehension problem in a manner identical to the one you used in solving the temperature problem. What is your view?

Physicist:

First of all, I don't think that most of my colleagues would agree that your data-based calibration procedure is "measurement". If you don't know enough about the items you use to specify their calibrations independent of data, how in the world can you be sure that you end up measuring what you intended to measure?

Furthermore, this data-based calibration procedure–used in lieu of theory–seems to impose a local boundedness on the measures you compute. It seems clear that a new 50-item reading comprehension test administered to a different group of high school seniors would yield measures on a different scale than did the first test. Each set of measures is bound to the particulars of the instrument and local context of measurement. As an aside, might this lack of objectivity in your measurement procedures partly explain the behavioral sciences' reliance on significance testing, correlation coefficients, meta-analytic methods and other metric-insensitive procedures? Could the excessive use of these methods be the inevitable consequence of working with measures that lack complete objectivity?

Psychometrician:

In response to your first observation, we have elaborate procedures for validating inferences from behavioral science measurements. In general, the most persuasive evidence that we know what we are measuring comes from high correlations among instruments purporting to measure the same construct.

I will grant you that a construct theory capable of supporting theory-based item calibrations might prove more persuasive, but at present, we have few such theories.

As for the second observation about lack of objectivity, we attempt to solve this problem, first, by employing the Rasch model which yields "local objectivity" and second, by using either a common-persons or common-items linking design, thus bringing all measures onto one scale. These procedures may appear primitive and cumbersome, but they do yield a kind of objectivity that is superior to what our measures have enjoyed in the past.

Finally, from the new perspective you have given me, I think it likely that, since we have no standard metrics for major behavioral science constructs, it is not surprising that we would gravitate towards measures of association and inference that ignore this failing.

Physicist:

Let me try to end on a positive note, for I do believe there is hope.

It should be clear from the temperature problem that measures of temperature possess a special kind of objectivity (i.e., absolute measures are separated from the conditions of measurement). The unavoidable consequence of using data-based calibrations rather than theory-based calibrations when converting observations to measures is that the resulting measures are expressed in units which, although of equal intervals, possess a kind of location indeterminacy. The only means I know of to remove this indeterminacy is to use theory rather than data to calibrate instruments. If you can develop construct theories and associated equations capable of calibrating your instruments, it seems possible that your measures can aspire to the kind of objectivity enjoyed by temperature measurement.

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, such as the height of a column of mercury in a thermometer. The "quantity", or measure, is the number assigned to the attribute of the object being measured (e.g., a person). Quantities, unlike other collections of numbers, possess an additive conjoint property (Luce and Tukey, 1964). 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, which, in its more colloquial form, is just a story about what it means to move up and down a scale, 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 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.

Measurement is a process of which the product is 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 actual 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 measures.

Local and General Objectivity

Objectivity is the foundation of valid measurement. Indeed, it is central to the whole idea of measurement. When I report a number to be used 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 I can ignore these conditions when I report the measure. As noted earlier, if I report that it is 75 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.

As seen in the previous dialog, 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 been sought after beginning with 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 all his models result in an approximation of specific objectivity only. 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, however, 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, p. x).
>
> "*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, irrespective 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 used a distinction between specific or local and general objectivity is not clear.

On the one hand, he 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. On the other hand, 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).

Wright (1991) continues a quarter-century of exploration of "objectivity" as a fundamental requirement of measurement. He states:

"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)

Thus, objectivity is clearly the cornerstone of all measurement. Measures must be completely independent of the particular instruments used and the particular conditions of measurement surrounding their use. A critical distinction exists between specific or local objectivity, as achieved by Rasch, and the general objectivity that is inherent in the concept of measurement used in the physical sciences.

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

An ideal, approximated by measures in physics and chemistry (e.g., thermometers), is that absolute location of an object on, for example, the Fahrenheit 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 even 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 only to theory.

By operating with a construct theory and associated calibration equations, we achieve general objectivity. 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.

Under a generally objective measurement framework, however, thanks to the sufficiency of raw scores, object measures are also entirely free of any reference to individual or group data.

In short, specific or local objectivity as achieved with the Rasch model ensures only that relative measures, that is, the differences between people, 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 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.

An Application To Reading Comprehension

Reading comprehension is the most tested construct in education. Among students aged six to 18, reading comprehension ability probably is measured more frequently than temperature, height, or weight. It 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 rates as a causal factor in the United States' dwindling economic prowess.

The importance of reading comprehension is underscored in today's "information age", in which the ability to read easily and well has become a survival skill. Even in production jobs, workers must be able to read complex operational and safety manuals in order to run the computerized equipment on which the modern factory depends. Strong reading skills also are necessary for the continuing education that rapidly changing technology and economic conditions demand, as well as for the requirements of citizenship, such as keeping up with political issues and current events.

Sadly, as many as one-third of Americans are functionally illiterate, unable to read standard adult-level text, such as employee manuals and newspaper articles, with reasonable comprehension. As the need for strong reading skills for work, continuing education and citizenship rises, these Americans are at an increasing disadvantage. Hence, the importance that educators and society attach to the construct of reading comprehension.

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. With no unifying theory for reading comprehension, it is impossible to convert a raw score (i.e., count correct) on one test into a scale score on another test. 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 some 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 results in confusion for educators, researchers, policy makers, and parents.

The Lexile Theory

People 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 and operators that are combined according to rules of syntax into mathematical expressions. In music, the semantic unit is the note, 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 governed largely by the familiarity of the semantic units and by the complexity of the syntactic structures used in constructing the message.

The Semantic Component. As far as the semantic component is concerned, it is clear that most operationalizations are proxies for the probability that a person will encounter a word in context and thus 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 are actually proxies for word frequency. They capitalize on the 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 hopes of identifying 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 were part 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 numerous algebraic transformations of these measures. We then ran correlations between the logit difficulties of the test items and each predictor variable. We found that the best operationalization of the semantic component of reading was 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). Through exploratory data analysis, we tested 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 then run 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 the observed item difficulty. The mean of the log word frequencies provided the highest correlation with item rank order.

The Syntactic Component. Sentence length is a powerful proxy for the syntactic complexity of a passage. One 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 visa versa.

Klare (1963) provides a possible interpretation for how sentence length works in predicting passage difficulty. He speculates that the syntactic component varies in the load placed on short-term memory. This explanation also 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.

Again, we correlated algebraic transformations of the mean sentence length for the 66 *Peabody Individual Achievement Test* (PIAT) reading comprehension items with item rank order. We found that the log of the mean sentence length was the best predictor of passage difficulty.

The Calibration Equation

We then combined the word-frequency and sentence-length measures in hopes of producing a regression equation that could 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.

Using the regression equation produced by this analysis, we assigned theoretical difficulties to 400 pilot test items (see Figure 1). The pilot items were ordered by difficulty and administered to approximately 3,000 students ranging from grades two to 12. Misfitting items were removed, leaving a total of 262 test items for which observed logit difficulties were computed using **M-scale** (Wright, Rossner, and Congdon, 1985).

FIGURE 1.

The mesa plain had an appearance of great antiquity, and of incompleteness; as if, with all the materials of world-making assembled, the Creator had desisted, gone away and left everything on the point of being brought together, on the eve of being arranged into mountain, plain, plateau. The country was still waiting to be made into a landscape. **It looked** _____.

A. arid
B. deserted
C. fertile
D. unfinished

The final specification equation was based upon the observed logit difficulties for the remaining 262 pilot test items. Again, the sentence length and word frequency variables were entered into a regression analysis of these logit difficulties. The resulting correlation between the observed logit difficulties and the theoretical difficulties was .97 after correction for range restriction and measurement error. The respective weights produced by the regression run formulated the following equation:

$$(9.82247 * LMSL) - (2.14634 * MLWF) - constant = Theoretical\ Logit \qquad (2)$$

Where: LMSL = Log of the Mean Sentence Length
MLWF = Mean of the Log Word Frequencies

The Lexile Scale

Once we established this equation, we re-scaled the theoretical logit difficulties. The logit scale is limited in that it has no fixed zero and, therefore, comparisons among different items or different populations are difficult (i.e., measures lack general objectivity). The method of imposing such a scale is quite simple.

For example, when a set of test items from a generic achievement test is given to fifth-graders from Podunk Primary, item difficulties will range from -4 to +4 logits, centered around the average item difficulty. When the same items are given to fifth-graders from Excel Elementary, the item difficulties also will be in logits from -4 to +4, again centered around the average difficulty. The average zero, however, floats depending upon the population taking the items. The students from Excel have, on average, a higher ability, and so the logit values will be lower (i.e., the items will appear to be easier). The logit values from the Podunk students will be higher (i.e., the items will appear to be more difficult) because the students have less ability.

Relative to a construct theory, though, test items have a fixed difficulty. The observed variation in difficulty occurs when the test item is given to people of different ability. Unless the logit scores obtained from a test administration are tied to a fixed zero, there is no way to compare the results of these test items given to two different populations.

The method of imposing a scale with a fixed zero point is simple. First, identify two anchor points for the scale. They should be intuitive, easily reproduced, and widely recognized. For thermometers, the anchor points are the freezing and boiling points of water. For the Lexile Scale, the anchor points are the text from seven basal primers for the low end and text from the *Electronic Encyclopedia* (Grolier, 1986) for the high end.

Second, using the regression equation, obtain the logit difficulty of the two anchors. For the Lexile scale, the mean logit difficulty of the primer material was -3.3 and the mean logit difficulty for the encyclopedia samples was +2.3.

Third, decide what the unit size should be. 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. To minimize the occurrence of negative Lexile values, we did not use zero as the low-end value, but instead assigned a value of 200.

Finally, we developed an equation that converts logit difficulties to Lexile calibrations. When the regression equation is used to analyze the anchors, the resulting difficulties are -3.3 logits for the primers and 2.3 logits for the encyclopedia. In order to set the -3.3 logits for the primer anchor equal to 200, we used the following equation.

$$(-3.3 + 3.3) + 200 = 200 \text{ Lexiles} \qquad (3)$$

The 3.3 that offsets the negative difficulty of the primer now becomes one of the two constants in the final formula. The second constant is determined when this equation is made to equal 1200 Lexiles, which is where the encyclopedia has been located.

$$[(2.26 + 3.3) * \text{Constant}] + 200 = 1200 \text{ Lexiles} \qquad (4)$$

The second constant turns out to be 180, which is the amount needed to convert the logit difficulty of the encyclopedia to 1200 Lexile units. The final equation that converts theoretical logit difficulties produced by the Lexile equation into Lexile units is as follows:

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

Measurements for persons and text are now reportable in Lexiles, which are similar to the degree calibrations on a thermometer. Essentially, the higher the Lexile measure for a text, the more difficult the material and the more 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 forecasted to achieve 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 directly matching a person's measure with a measure for a text. The difference between these two measures is used to forecast the comprehension that the person will have with that text. Comprehension is always relative to the difference between person measure and text measure.

The attached Lexile map can be used to bring meaning to these Lexile measures of text and persons. This richly annotated four color, poster-sized graphic provides an extensive list of texts, from novels and non-fiction books to newspapers and magazines, at various levels of Lexile measurement. The map makes it easy to "see" how reading develops and to select other reading materials as students progress in their reading development.

Testing the Lexile Equation

A computer program incorporating the Lexile equation is available that analyzes continuous prose and reports the difficulty in Lexiles (MetaMetrics, 1995). In order to test the power of the theory, we analyzed 1,780 reading-comprehension test items appearing on nine nationally normed tests (Stenner, Smith, Horabin, and Smith, 1987). The study involved correlating Rasch item difficulties provided by the publisher with the Lexile calibrations generated from the computer analysis of the text of each item. In those cases where multiple questions were asked about a single passage, we averaged the reported item difficulties to yield a single observed difficulty for the passage.

We obtained the observed difficulties 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 based upon item p-values and raw score means and standard deviations (e.g., CAT). To obtain these logit difficulties, we used TestCalc (Horabin, 1989), a computer program for analyzing test data. 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 observed difficulty was approximated by the difficulty rank order of the item.

Once theory-based calibrations and data-based item difficulties were computed, we correlated the two arrays and plotted them separately for each test. The plots were checked for unusual residual distributions and curvature, and we 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, which still accounts for a majority of reading material. The poetry and non-continuous prose items were removed and correlations were again obtained and used to describe the fit of observation to theory.

Two major influences other than model misspecification operate to artificially deflate the relationship between theory and observation. The first is range restriction in the item difficulties. Some tests purposely do not cover the full developmental continuum for reading comprehension. The NAEP (1983), for example, is administered to grades four, eight, and 11. As might be expected, the resulting restriction in the range of item difficulties tends to attenuate the relationship between theory and data. Thorndike (1949) gives the procedure for correcting a

correlation for restriction in range where the range of the variable in the unrestricted group is known.

A second influence that operates to reduce the correlation between theory and data is unreliability in the theory-based item calibrations. Theories are rarely perfectly operationalized. As we have noted, the Lexile equation contains two terms that are both proxies for the presumed underlying causes of item difficulty. Proxies are imperfect substitutes for the theoretical causes and, as such, act to attenuate correlations. The data-based difficulties, on the other hand, are so well estimated that the reliabilities are typically near .99. Stanley (1971) gives the procedure for disattenuating a correlation for unreliability in one of the variables.

Finally, note that the Lexile analysis was applied only to the passages and did not include the questions and their respective answers. This decision most likely introduced error, since it has long been recognized that the questions themselves add to the overall difficulty of a test item. The magnitude of these influences is difficult to estimate, but we can safely assume that some of the remaining differences between theoretical calibrations and data-based difficulties are due to these factors.

Table 1 presents the results of correlating the theoretical calibrations and observed difficulties. The last three columns of the table show the raw correlation between observed (O) item difficulties and theoretical (T) item calibrations, with the correlations corrected for restriction in range and measurement error. The mean of the raw correlations is $r_{(OT)} = .84$. When corrections are made for range restriction and measurement error, the average disattenuated correlation between theory-based calibration and data-based difficulty in an unrestricted group of reading comprehension items is $R'_{(OT)} = .93$.

TABLE 1

Correlation Between Theory-Based Calibrations Produced by the Lexile Equation and Data-Based Item Difficulties

Test	Number of Questions	Number of Passages	$r_{(OT)}$	$R_{(OT)}$	$R'_{(OT)}$
SRA	235	46	.95	.97	1.00
CAT-E	418	74	.91	.95	.98
Lexile	262	262	.93	.95	.97
PIAT	66	66	.93	.94	.97
CAT-C	253	43	.83	.93	.96
CTBS-U	246	50	.74	.92	.95
NAEP	189	70	.65	.92	.94
Battery	26	26	.88	.84	.87
Mastery	85	85	.74	.75	.77
Totals Grand Means	*1780*	*722*	*.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.

It seems reasonable to conclude from these results that most attempts to measure reading comprehension, no matter what the item form, type of skill objectives purportedly being measured, or response requirement used, all end up measuring a common comprehension factor specified by the Lexile theory (Stenner, Horabin, Smith, and Smith, 1988).

In a second study, we obtained Lexile measures for units within 11 basal series. It was assumed 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 that same book. Likewise, a fourth-grade reader is presumed to be more difficult than a third-grade reader. We estimated observed difficulties for each unit in a basal series 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 and measurement error, the average correlation between the Lexile calibration of text comprehensibility and the rank order of the basal units was .99 (see Table 2).

TABLE 2

Correlations between the Lexile Calibration and The Rank Order of Unit

Basal Series	Number of Units	$r_{(OT)}$	$R_{(OT)}$	$R'_{(OT)}$
Ginn Rainbow Series (1985)	53	.93	.98	1.00
HBJ Eagle Series (1983)	70	.93	.98	1.00
Scott Foresman Focus Series (1985)	92	.84	.99	1.00
Riverside Reading Series (1986)	67	.87	.97	1.00
Houghton-Mifflin Reading Series (1983)	33	.88	.96	.99
Economy Reading Series (1986)	67	.86	.96	.99
Scott Foresman American Tradition (1987)	88	.85	.97	.99
HBJ odyssey Series (1986)	38	.79	.97	.99
Holt Basic Reading Series (1986)	54	.87	.96	.98
Houghton-Mifflin Reading Series (1986)	46	.81	.95	.98
Open Court Headway Program (1985)	52	.54	.94	.97
Totals	660			
Means*		.86	.97	.99

$r_{(OT)}$ = raw correlation between observed rank order (O) and theory based calibrations (T).
$R_{(OT)}$ = correlation between observed rank order (O) and theory based calibrations (T) corrected for range restriction.
$R'_{(OT)}$ = correlation between observed rank order (O) and theory based calibrations (T) corrected for range restriction and measurement error.

*Means are computed on Z transformed correlations.

The fact that the 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).

Lexile Measures and Probable Error

A fundamental concept in the use of a measure is "probable error", or how much error is attributable to that measure. A well-known feature of all measurement procedures is that repeated measurement of the same thing results in a series of non-identical amounts. The root mean square of the differences among repetitions is the standard deviation of the distribution, also called the standard error of measurement (SEM). This statistic describes how close together are the repeated measurements. The SEM also defines the prediction interval for the question, "What would happen if we measured again?"

There are many ways of defining what we mean by "measuring again". Definitions of a "repetition" differ in what facets of the measurement procedure vary from one repetition to another. As more facets vary, more variation is observed in the replicate measures and the SEM increases. A highly restricted recipe for replication may result in a smaller SEM, but may overstate the precision of measurement that results from "normal use" of the measurement procedure.

With regard to the Lexile Framework, if we were to process 20 pages of text again and again, we would see a standard deviation of 40 Lexiles. Similarly, if we were to test individuals according to a broad recipe–different items on different days–we would end up with a standard deviation of 70 Lexiles.

Application of the Lexile Scale

One of the major weaknesses of 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 all third-grade students involved in the norming study. Percentile scores on standardized reading tests, however, do not provide any information about what a student can or cannot read. What does a teacher or parent actually do with a percentile score? What kind of instruction can a teacher give a student when the only information provided is that a particular child is reading at the 65th percentile of all third-graders across the nation?

The Lexile scale is designed to provide both a normative and a criterion-referenced interpretation of a measure. Since the Lexile scale is based upon the Rasch model, the probability

of a person answering a reading item correctly is governed only by the difference between the person's measure and the task's calibration. This relationship is captured in the following equation:

$$O = \sum_{i} \frac{e^{(M_o - C_i)}}{1 + e^{(M_o - c_i)}}$$

O = observation (count correct)
P_{ni} = the probability of a correct response
M_O = the person's measure
C_i = item calibrations

If 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. If 20 such tasks were given to this person, one would expect three-fourths of the responses to be correct. If the task is more difficult than the person is able, then the probability is less than 75 percent that the response of the person to the task will be correct; similarly, if the task is easier compared to a person's measure, then the probability is greater that the response will be correct.

A person with a Lexile ability of 600L who is given a text measured at 600L will have a 75 percent comprehension rate. If the same student is given a text measured at 350L, the forecast comprehension rate improves to 90 percent.* Give the same student a 100L title, and comprehension improves to 96 percent. The more a person's Lexile measure surpasses the Lexile measure for a task, the higher the forecasted comprehension rate. The more the Lexile measure for a task surpasses a person's Lexile measure, the lower the forecasted comprehension rate. Tables 3 and 4 illustrate the relationship between person measure, text measure and the forecasted comprehension rate.

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%

Note that it is the difference in Lexiles between person and text that governs comprehension. The difference between a 200L text and a 450L reader results in the same success rate as with a 600L text and an 850L reader. Each case produces a 90 percent comprehension rate.

Empirical evidence supporting a 75 percent target comprehension rate, as opposed to, say, a 50 percent or 90 percent rate, is limited. Squires, Huitt, and Segars (1983) did find 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). It may be, however, that there is no one optimal rate, but rather a range in which individuals can operate to improve optimally their reading ability. We have found, however, that the subjective report of readers reading at 50% comprehension is frustration whereas readers reading at 75% comprehension report comfort and confidence with the text.

*If a text was comprised of slices (125-word passages) all having the same Lexile calibration, then the pivot value needed to move from 25 to 50 to 75 to 90 to 96 percent comprehension always would be 1.1 logits or 200 Lexiles. Text slices vary within a title, however. Consequently, a pivot value of 250 Lexiles is used.

Most applications of the Rasch model anchor the scale in such a way that when measure and indicant difficulty are equal, the probability of a correct response is p=.50. By adjusting the ability scale upward an arbitrary 1.1 logits, we anchor the scale in such a way that, when measure and indicant difficulty are equal, the probability of a correct response is p=.75. We have found that teachers are more comfortable with this approach.

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 a famous athlete. The teacher may not know, however, whether that biography is too difficult or too easy for the student. The Lexile Framework can provide a person measure and text measure on the same scale. Armed with this information, a teacher or parent can plan for success. This ability to provide strong linkages between test results and reading material, among students, teachers and parents, and between schools and workplaces is one of the primary advantages of the Lexile Framework (see Appendix A).

Improving student's progress in reading requires that they read properly targeted prose accompanied by frequent response requirements. Response requirements range from having a more competent reader ask occasional questions as the reader progresses through the prose to embedding questions in the text, much as is done with Lexile test items. The reason for requiring that readers do more than simply read is that unless there is some evaluation, there can be no assurance that the reader is properly targeted and comprehending the material. Students should be given text on which they can practice being a competent reader (Smith, 1973). The above approach does not represent a fully articulated instructional theory, but its prescription is straightforward. Students should read more targeted prose and teachers should 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 attendant 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 chosen level. In essence, we need to locate a reader's "edge" and then systematically expose the reader to text that plays on that edge. When this approach is followed in any domain of human experience, the edge moves and the capacities of the individual are enhanced.

What happens when the "edge" is over-estimated and repeatedly exceeded? In any kind of physical exertion, if you push beyond the edge you feel pain; if you demand even more performance on the part of 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 materials well out of their reach reduces self-confidence, stunts growth and eventually results in the individual "tuning out". With the tremendous emphasis placed on reading in daily activities, virtually every encounter with written text becomes a reconfirmation of a poor reader's inadequacy. Is it any wonder that 15 to 20 percent of U.S. high school students decide to find some other way to spend their days (Hahn, 1987)?

For students to become competent readers, they need to be exposed to text that results in a comprehension rate of 75 percent or better. If an 850L reader is faced with an 1100L text

(resulting 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 properly targeted, they read fluently with comprehension; when improperly targeted, they struggle both with the material and with maintaining motivation and their self-esteem. In reality, there are no poor readers–only mistargeted readers who are being challenged inappropriately.

Leading researcher Jeanne S. Chall states that the Lexile Framework, while undergirded by highly sophisticated statistical procedures, stands firmly in the tradition of classic readability formulas with its emphasis on comprehension as a function of semantic and syntactic components. In contrast with cognitive-structural readability researchers, she notes that the developers of the Lexile Framework view comprehension as a unidimensional ability that subsumes different types of comprehension. Under the Lexile theory, no special consideration is given to prior or special subject knowledge.

Chall also points out several unique features of the Lexile Framework, including the use of calibrated scores to represent equal units of difficulty and the fact that the same scores measure both the difficulty of text and the ability of readers. These Lexile scores correlate well with nine classic readability formulas and with item difficulties on several tests of reading comprehension (Chall, 1995).

Benefits of Objective Measurement

What has been proposed in this paper can be characterized as a minor adjustment to a well-known model that results in profound implications for what we can do with behavioral science measures. Truly objective measurement can be achieved by simply replacing the data-based difficulties of the Rasch model with theory-based calibrations. When this is done, several benefits accrue: (1) the construct theory is exposed to falsification, (2) it is possible to build correspondence tables between observations and measures with recourse only to theory, (3) a generalized linking solution is available for placing observations of all kinds on a common scale, (4) a reproducible unit of measurement can be developed, (5) the framework for fit statistics that is sample-dependent under the Rasch model becomes sample-independent, and (6) a complete frame of reference for measure interpretation can be constructed.

A Refutable Construct Theory

A construct theory is a story about what it means to move up and down a scale. A specification equation is a regression model, based on the theory, that forecasts indicant calibrations on the scale. To the extent that the theory is adequate and the specification equation forecasts accurately, it is possible to achieve generally objective parameter estimation. Construct theories embodied in specification equations are exposed to test and falsification. Through such

attempts at falsification, construct theories are sharpened, specification equations are made more accurate and understanding of what our measures, in fact, do measure is enhanced.

Correspondence Tables Via Theory

Measurement is the process of converting observations into quantities through theory. The graphic that represents the essential correspondence between observation and measure (i.e., quantity) is termed a "correspondence table". This table is constructed via theory in physical science measurement and via data in behavioral science measurement. General objectivity means that two observers will arrive at the same numbers to assign as a measure, given that they begin with the same observation. Objectivity refers to the process of converting observations into measures and is not meant to imply that two applications of the same measurement procedure will arrive at identical measures.

A Generalized Linking Solution

Rasch measurement recognizes two linking designs: common objects and common indicants. Under the common-objects design, different instruments (sets of indicants) are administered to the same sample of objects, thus establishing calibration differences between the various measurement procedures. The common-indicants design involves administering a sub-set of indicants taken from two or more instruments to a sample of persons and using the differences in mean calibrations to equate scales. Unfortunately, the most frequently encountered situation that requires a linking procedure involves two tests that share no indicants and that have not been administered to the same group of persons. This situation has been considered intractable.

Objective measurement provides a means for linking or equating instruments in the case of uncommon objects and uncommon items. The procedure uses a common theory embodied in a common specification equation to link instruments. The specification equation provides theory-based calibrations for each indicant, thus equating all measures of a construct to a common theory-referenced scale. Under this new procedure, all reading comprehension tests developed in the last century can now be placed on a common scale (Stenner, Horabin, Smith, and Smith, 1988).

A Reproducible Unit of Measurement

Most units of measurement used in the behavioral sciences are fractions of some sample standard deviation. The unit, as such, cannot be reproduced in the same way that a degree Celsius can be reproduced, but an approximation can be gained through a common-objects or common-indicants design.

A sounder strategy is to select as a unit some fraction of the difference between two publicly reproducible states or artifacts. In the case of the Celsius scale, the unit is 1/100 of the

distance between freezing and boiling "normal" water at sea level. In the case of the Lexile framework, the unit is 1/ 1000 of the distance between the comprehensibility of first-grade primers and encyclopedias. This unit also is reproduced easily from the Lexile theory and associated specification equations.

Model Fit

Objective measurement shifts the focus away from fit to the model to theory and measure validation. Theory validation is tested by checking the correspondence between theory-based and data-based estimates of indicant parameters. The analysis of interest is the plot between theory-based calibrations of the indicant and data-based difficulties. The construct theory and associated specification equation is well confirmed to the extent that the above correlation approaches r=1.0 upon repeated application of the specification equation to indicants from different instruments. Measure validation corresponds to person fit in the Rasch model and uses the identical statistics and interpretive framework. The difference is that Rasch measurement does not require an a priori statement of intention. Rather, "intention" is arrived at a posteriori in the form of estimated difficulties. Fit is then checked against this data-dependent statement of intention. Surely, intention must be stated prior to data analysis and its argument must be framed in language that moves beyond the immediate materials and moment of measurement. Objective measurement through the specification equation and theory-based indicant calibrations makes an explicit a priori statement of intention. Whether our intention is realized is examined at the theory or specification equation level and at the individual object measure level. Thus, we speak of construct theory validation and of measure validation.

Elsewhere, we have argued that the conventional approach to validity confuses the question of whether an instrument measures what it is intended to measure with the question of how useful the obtained measures are in the description and prediction of phenomena (Stenner, Horabin, Smith, and Smith, 1988). Objective measurement clearly distinguishes between these two questions. It is quite possible for an instrument to generate measures that are valid in the sense discussed above, but have no known use. The absence of measure utility, however, does not mean the construct theory is invalid or that the measures produced are invalid. On the other hand, low theory validity and/or measure validity does reduce the expected utility of a measure.

Frames of Reference

For most of this century, the dominant frame of reference for score interpretation has been a normative one. Early attempts at criterion-referenced score interpretation either emphasized objectives or proposed so-called mastery or cut-scores. The former generally had little to do with what was actually being measured by an instrument and the latter proved unsatisfyingly primitive as a means of attaching meaning to various score levels.

Perhaps the best-known break with this tradition was the scale annotation procedure illustrated by Thurstone (1925), used to good advantage by Woodcock (1974) and now employed

by NAEP (1987). The Keymath test manual (Connolly, Nachtman and Pritchett, 1971) provided a graphic that annotated selected points along the scale with items taken from the test. The test user could examine these items and "see" the progression in math ability as scale scores increased. This criterion-referencing procedure represented a major advance over objective-based and mastery-level procedures. The disadvantage is that the annotation procedure is context-bound. Only items used in the test or linked to the test through a common-objects or common-indicants design can be used in the annotation process. With objective measurement, the specification equation can be used to build a rich criterion frame-of-reference that is independent of any particular set of indicants. In the reading comprehension domain, for example, a specification equation is used to generate Lexile measures for books and periodicals. These titles are then arranged along the scale to describe what readers at various levels of reading comprehension can read at selected comprehension rates. We refer to this process as scale specification. It is distinguished from scale annotation by the use of specification equations that can compute theory-based calibrations (in a common metric) for any task from the construct universe governed by the construct theory. Just as we don't need data to know what a particular hash mark on a thermometer means, we don't need data to know what a given raw count or raw score means on a theory-referenced task or measure.

Most, if not all, norms in use today reflect norm-group performance on a particular test. Rasch measurement practitioners and theoreticians have written on the prospect of norming a reference scale (Wright and Stone, 1979), but, in the absence of general objectivity, it has not been worth the effort. Objective measurement changes this state of affairs. Since absolute scale locations for persons are sample-independent, when measures are generally objective, it makes sense to norm-reference the measurement scale and not just scores produced by a single instrument. Once this has been done and a norm-maintenance program has been established, all past and future measures of the construct in question are norm-referenced.

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, *Constnuct 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.

Chall, J.S. and Dale, Edgar (1995). *Readability Revisited: The New Dale-Chall Readability Formula.* Cambridge, MA: Brookline Books.

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

Connolly, Nachtman, and Pritchett (1971). *Key Math Diagnostic Arithmetic Test.* American Guidance Service, Inc.

Crain, S., and Shankweiler, D. (1988). Synactic 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., and C. M. (1975, March). *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- 2O9.

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 theorv 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 Probalistic 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). *Probalistic 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., Testing construct theories. *Perceptual and Motor Skills,* 1982, 55, 415-426.

Stenner, A. J., Smith, M., and Burdick, D. S. (1983). Toward a theory of construct definition. *Journal of Educational Measurement*, 20, 3O5-315.

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

Stenner, A.J., Horabin, I., Smith, D.R., and Smith, M. (1988). *Most Comprehension Tests Do Measure Reading Comprehension: A Response To McLean and Goldstein.* Phi Delta Kappan, June 1988, 765-769.

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.