DOCUMENT RESUME

ED 307 577 CS 009 612

AUTHOR Smith, Dean R.; And Others

TITLE The Lexile Scale in Theory and Practice. Final Report.

INSTITUTION MetaMetrics, Inc., Washington, D.C.

SPONS AGENCY National Institutes of Health (DHHS), Bethesda, Md.
PUB DATE 89
GRANT NIH-HD-19448
NOTE 46p.
PUB TYPE Reports - Research/Technical (143)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Analysis of Variance; Correlation; Criterion Referenced Tests; Difficulty Level; *Measurement Techniques; *Readability Formulas; Reading Comprehension; Reading Research; Reading Tests; Skill Development; Test Reliability; Theory Practice Relationship
IDENTIFIERS *Lexile Scale

ABSTRACT

A three-part correlational study examined the explanatory power of the Lexile theory of reading comprehension, which was based on the semantic and syntactic components of prose. Correlations were performed between the item difficulties of nine nationally normed reading comprehension tests and computer generated difficulties which were reported in Lexiles. A correlation of 0.93 was obtained between observed and theoretical scores. A second test was performed in which the rank order of basal series units were correlated with Lexile ratings of text difficulty. A correlation of 0.99 was obtained. A third test was performed in which the correlations between test item difficulties and Lexile ratings were compared with correlations derived from nine measures of readability. Results indicated that while the Lexile equation produced better correlations on average, analysis of variance revealed that the Lexile ratings did not provide a significantly better explanation of the test item difficulties than the readability formulas. Results indicated that the Lexile theory does account for a significant portion of the difficulty of continuous prose and can be used to generate normative and criterion interpretations of a score which would facilitate the direct matching of student abilities with reading materials of appropriate difficulty. (Two figures and 10 tables of data are included; 42 references are attached.) (RS)

The Lexile Scale in Theory and Practice:

Final Report for NIH Grant HD-19448

Dean R. Smith, A. Jackson Stenner, Ivan Horabin, and Malbert Smith

MetaMetrics, Durham, North Carolina

Abstract

As part of a three year NIH initiative, the Lexile theory of reading comprehension was developed based upon the semantic and syntactic components of prose. In order to test the explanatory power of the theory, correlations were performed between the item difficulties of nine nationally normed reading comprehension tests and computer generated difficulties which were reported in Lexiles. After correction for range restriction and measurement error, the mean correlation obtained between observed and theoretical scores was .93. A second test was performed in which the rank order of basal series units were correlated with Lexile ratings of text difficulty. After correction for range restriction and measurement error, the mean correlation was .99. A third study was performed in which the correlations between test item difficulties and Lexile ratings were compared with correlations derived from nine measures of readability. Although the Lexile equation produced better correlations on average, an analysis of variance revealed that the Lexile ratings did not provide a significantly better explanation of the test item difficulties than the readability formulas (F = .233). The correlations among test item difficulties and rank order basal units suggest that the Lexile theory does account for a significant portion of the difficulty of continuous prose. The fact that the theoretical values from the Lexile theory can be used to generate individual ability scores and text difficulty ratings provides the means for developing both normative and criterion interpretations of a score. Such a juxtaposed scale

would facilitate the direct matching of student abilities with reading materials of appropriate difficulty.

THE LEXILE SCALE IN THEORY AND PRACTICE:

Final Report for NIH Grant HD-19448

This symposium is designed to report the results of a three year study of reading comprehension funded under an NIH initiative (grant #HD-19448). The specific purposes of this paper are to 1) discuss the process of construct definition, 2) introduce the Lexile theory as a construct definition of reading comprehension, 3) examine the evidence supporting the explanatory power of the Lexile theory, 4) discuss the impact of the Lexile theory upon testing and instruction, and 5) explore how the Lexile scale can be used to provide an operational definition of adult literacy.

Construct Definition

A test is a collection of items sampled from a specified universe. The items are developed in order to differentiate between people who possess varying degrees of an ability or a trait. The ability or trait being measured is a construct.

When test items are administered, the items and people order themselves according to difficulty and ability respectively. Some items are more difficult than others, and some people possess a higher degree of ability than others.

Construct definition (Stenner and Smith; 1982) is a process whereby an ability is operationalized as quantifiable attributes of test items. These measured attributes are then combined into a regression equation designed to explain variation in item difficulties. By explaning what makes some items more difficult than others, it is hoped that the causes of variation in person

ability can be identified. In short, test item variation is the window used to understand the cognitive processess associated with the construct. A method known as construct generalization (Stenner, Smith, and Burdick; 1983) can be used to test how well a particular construct theory can be generalized. This method involves the following steps:

1. Collect a sample of tests which were designed to measure the targeted construct.
2. Obtain Rasch difficulties for each of the test items.
3. After examining the literature related to the targeted construct, identify and quantify variables which may account for variation in item difficulties (i.e. explain why some test items are more difficult than others).
4. Use a regression analysis to develop an equation that can generate theoretical difficulties for any given test item.
5. Obtain theoretical difficulties by applying the regression equation to each of the test items.
6. Correlate the theoretical difficulties and the observed Rasch difficulties.
7. Correct the correlations for range restriction and measurement error.
8. Test the causality of the variables in the regression equation by systematically manipulating the variables and checking for predicted results in observed difficulty.

The process of construct generalization has been applied to short term memory (Stenner and Smith, 1982) and receptive vocabulary (Stenner, Smith, and Burdick; 1983), and was expanded to reading comprehension in this study.

A Construct Definition of Reading Comprehension

We communicate using various symbol systems including mathematics, music, and language. All symbol systems share two features; each possesses a semantic 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. The semantic units in language 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 largely governed by the familiarity of the semantic units and by the complexity of the syntactic structures used in constructing the message.

Applied to language, the difficulty of a passage is governed by the vocabulary and sentence structures used. The readability literature provides a rich source of quantified variables used to measure these elements of prose material.

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 context and thus infer its mearing (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 which is further developed by Carroll, Davies, and Richman (1971) whose word frequency study examined the reoccurrance 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 of their being encountered and thus becoming a part of an individual's receptive vocabulary.

Variables such as the average number of letters or syllables per word are in actuality proxies for word frequency. They capitalize on the high negative correlation between the length of words and the frequency of word usage. Long words and polysyllabic words are used less frequently than short monosyllabic words making word length a good proxy for the likelihood of an individual being exposed to them.

Stenner, Smith, and Burdick (1983) analyzed over fifty semantic variables in hopes of identifying those elements which 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, the 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. Correlations were then run between the logit difficulties of the test items and each targeted 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). In order to test the explanatory power of this variable, exploratory data analysis was performed. This 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 frequency provided the highest correlation between the theoretical difficulties (word frequency) and observed difficulties (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 is also supported by Crain and Shankweiler (1988), Shankweiler and Crain (1986), Liberman, Mann, Shankweiler and Westelman (1982), whose work has provided evidence that sentence length is a good proxy for the demands that structural complexity place upon verbal short term memory.

Hitch and Baddeley (1974) provide further insight into the impact of sentence length upon the comprehension process. They speculate that the critical facility is the executive control component, not just working memory capacity. This notion is also discussed by Shankweiler and Crain (1986). If this speculation is correct, then low scoring students on listening and reading comprehension tests are distinguished from high scoring students not in the number of working memory registers but rather in the management of available memory.

Exploratory data analysis was also performed upon sentence length in order to find the best fitting operationalization of the syntactic variable. Algebraic transformations of the mean sentence length for the 66 Peabody Individual Achievement Test (PIAT) reading comprehension items were again correlated with the respective rank order. It was found that the log of the mean sentence length was the best predictor of passage difficulty.

The Regression Equation

The word frequency and sentence length measures were then combined in hopes of producing a regression equation that could explain most of the variance found in any set of reading comprehension test items. 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 of the variance (r = _) in FIAT 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 3000 students ranging from grade 2 to grade 12. Misfitting items were removed leaving a total of 262 test items for which observed logit difficulties were obtained using M-Scale (Wright, Rossner, and Congdon; 1985).

The final regression 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) - 3.23274 = Theoretical Logit

Where LMSL = Log of the Mean Sentence Length

Where MLWF = Mean of the Log Word Frequencies

FIGURE 1

An Example Of A Lexile Test Item

Wilbur liked 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 Lexile Scale

Once the equation was established, a developmental scale was imposed which would provide a fixed zero point. The logit scale is limited in that it has no fixed zero and therefore, comparisons among different items or different populations are impossible.

For example, when a set of test items from the "Generic Achievement Test" are given to 5th graders from Podunk Primary, item difficulties will be obtained which range from -4 to +4 logits centered around zero. When the same items are given to 5th graders from Excel Elementary, the item difficulties will also be in logits from -4 to +4 centered around zero. But the zero floats depending upon the population taking the items. The students from Excel have higher ability on average, and so the logit values will be lower (the items will appear easier). The logit values obtained from the Podunk students will be higher (the items will appear to have more difficulty) because the students have less ability.

However, test items have a fixed difficulty. The variation occurs when the same 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 problem also exists on the person-face of the matrix. If the same population takes two different tests, two different logit ability estimates will be obtained. Again, these logits cannot be compared until they are placed on a scale with a fixed

zero point. The method of imposing such a scale is quite simple.

First, identify two anchor points for the scale. They should be intuitive, easily reproduced, and widely recognized. For thermometers, the anchor points were the freezing and boiling points of water. For the Lexile scale, the anchor points were the text from seven basal primers for the low end and text from the Electronic Encyclopedia (Gollier, 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 of the encyclopedia samples was +2.256.

Third, decide what the unit size should be. For the Fahrenheit thermometer, the unit size (a degree) is 1/180 or the difference between the freezing (32 degrees) and boiling points (212 degrees) of 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 difficulty of the primers and the encyclopedia.

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

Finally, an equation needed to be developed which converted logit difficulties to Lexile scale scores. When the regression equation was used to analyze the anchors, the resulting difficulties were -3.3 logits for the primers and 2.256 logits for the encyclopedia. In order to set the -3.3 logits for the

primer anchor equal to 200, the following equation was used:

(-3.3 + 3.3) + 200 = 200 Lexiles

The 3.3 which 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:

[(2.256 + 3.3) * Constant] + 200 = 1200 Lexiles

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 which converts logit difficulties to Lexile units is as follows:

[(Logit + 3.3) * 180] + 200 = Lexile difficulty

The measurement of student ability and text difficulty are now reported in Lexiles which are similar to the degree calibrations on a thermometer. Essentially, the higher the Lexile score, the more difficult the material or the more ability a student possesses for comprehending a prose selection. The scale can be used to juxtapose the difficulty of test items or reading materials to the reading ability of students. Prose materials are rated using the Lexile equation and are placed on the scale according to their difficulty. The abilities of students are placed on the scale by analyzing their performance upon Lexile rated test items. This provides the means for directly matching a student's ability with reading materials of appropriate difficulty. Figure 2 is a graphic which can be used to facilitate this matching and is described below from left to right:

a. **Instructional Level Novels** are titles of books that are found in a ranges of 100 Lexiles. The list could be composed of award winning novels for children and adolescents or any basal series or any texts used in a given school district. Such a list would provide an informal way for the viewer to target books at his or her Lexile reading ability level.

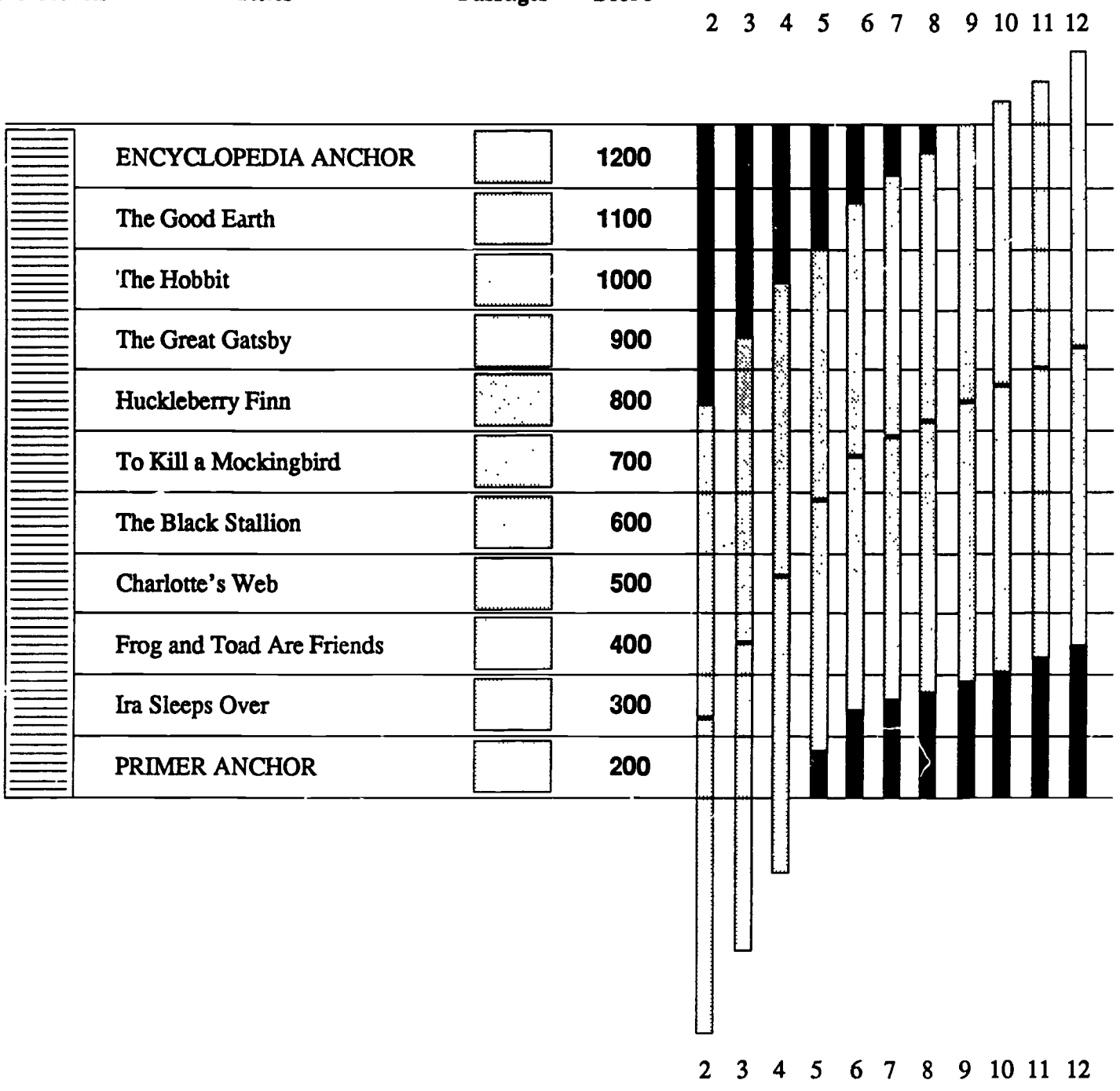
b. **Sample Titles** are novels easily recognized by a majority of people and will give viewers a good example of what a 1000 Lexile novel might be. The anchor points at 200 and 1200 Lexiles are also included.

c. **Sample Passages** from the targeted sample novels provide the viewer with an idea of what a 300 or 400 Lexile passage looks like.

d. **The Lexile Scale** ranging from 200 to 1200 Lexiles is centered on the graphic.

e. **The Norms** depict where students rank using the traditional percentile approach to testing. The percentiles would range from 5% to 95% with a distinctive indicator at 50% which would show how the average student is reading at a given grade level. It would also provide indicators in 5% increments so that the viewer can plot any given percentile on a standardized test of reading comprehension to the corresponding level in Lexiles.

FIGURE 2



LEXILE SCALE OF COMPREHENSION

Testing the Generalizability of the Lexile Equation

Based upon the Lexile equation, a computer program has been developed that analyzes continuous prose and reports the difficulty in Lexiles (Horabin, 1987). In order to test the power of the theory, 1780 reading comprehension test items appearing on nine different tests were analyzed (Stenner, Smith, Horabin, and Smith; 1987). The study involved correlating the test item difficulties provided by published norms with the Lexile difficulties generated from the computer analysis of each test passage. In those cases where multiple questions were asked about a single passage, the reported item difficulties were averaged to yield a single observed difficulty for the passage.

The observed 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 based upon 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 obtain 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 observed difficulty was approximated by the rank order of the item.

Once theoretical values and observed item difficulties were computed, the two arrays were correlated and plotted separately for each test. The plots were checked for unusual residual distributions 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 which still accounts for a large 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 4, 8, and 11. As might be expected, the resulting restriction in the range of item difficulties tends to attenuate the relationship between theory and observation. Thorndike (1949) defines the procedure for correcting a correlation for restriction in range where the range of the theoretical variable in the unrestricted group is known.

A second influence that operates to reduce the correlation between theory and observation is unreliability in the theoretical item difficulties. Theories are rarely perfectly operationalized. As has already been noted, the Lexile equation contains two terms both of which are 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 observed difficulties on the other hand are so well estimated that the reliabilities are typically near .99. Stanley (1971) defines the procedure for

disattenuating a correlation for unreliability in one of the variables.

Finally, it should be noted 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 it is safe to assume that some of the remaining differences between theoretical difficulties and observed difficulties are due to these factors.

Table 1 presents the results of correlating the theoretical and observed difficulties. The last three columns of the table show the raw correlation between observed (O) item difficulties and theoretical (T) item difficulties; the correlations corrected for restriction in range; and 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 and observation in an unrestricted group of reading comprehension items is $R'(OT) = .93$.

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 captured by the Lexile theory.

TABLE 1

Correlations between Theoretical Difficulties

Produced by the Lexile Equation and Observed Item Difficulties

Test	# of Questions	# 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	1780	722			
GRAND MEANS			.84	.91	.93

r(OT) = raw correlation between observed difficulties (O) and theoretical Lexiles (T).

R(OT) = correlation between observed difficulties (O) and theoretical Lexiles (T) corrected for range restriction.

R'(OT) = correlation between observed difficulties (O) and theoretical Lexiles (T) corrected for range restriction and measurement error.

A second study was performed in which Lexile ratings were obtained for units within eleven major 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. 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 then run upon the ranked ordered units to the Lexile ratings of each unit. After correction for range restriction and measurement error, the average correlation produced between the Lexile theory's analysis of difficulty and the rank ordering of the basal series units was .99 (see Table 2).

The fact that the Lexile theory accounted for the unit difficulties of eleven basal series is all the more noteworthy when it is recognized that the series differ in prose selections, differ in the developmental range addressed, differ in the types of prose introduced (i.e. narrative versus expository), and differ in what 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).

TABLE 2

Correlations between the Lexile Measure Of Difficulty and the Rank Order of Units from 11 Basal Series

Basal Series	# 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			
GRAND MEANS		.83	.97	.99

r(OT) = raw correlation between observed rank order (O) and theoretical Lexiles (T).

R(OT) = correlation between observed rank order (O) and theoretical Lexiles (T) corrected for range restriction.

R'(OT) = correlation between observed rank order (O) and theoretical Lexiles (T) corrected for range restriction and measurement error.

A third study was performed which compared how well the Lexile equation did in predicting item difficulties compared to estimates from nine readability formulas. Again, correlations between Lexile ratings of passage difficulties and observed logit or rank order difficulties from five tests of reading comprehension were obtained. These results were then compared to correlations between the theoretical difficulties produced by the nine readability formulas and the reported observed item difficulties. Although the Lexile equation produced better correlations on average, an analysis of variance revealed that the Lexile ratings did not provide a significantly better explanation of the test item difficulties than the readability formulas ($F = .233$). Table 3 provides a summary of the correlations obtained from the Lexile equation and the nine readability formulas.

TABLE 3

A Comparison of the Lexile Equation and Nine Readabilty Formulas

Equations	Tests					
	Lexile	SRA	Battery	Mastery	Peabody	Mean
Lexile	.895	.921	.854	.744	.942	.871
Holmquist	.807	.913	.807	.835	.864	.845
ARI	.851	.934	.850	.705	.849	.837
Flesch-Kincaid	.851	.936	.850	.703	.845	.837
FOG	.847	.919	.726	.750	.848	.818
Powers	.816	.929	.827	.647	.741	.792
Dale-Chall	.735	.870	.747	.746	.833	.786
Coleman	.759	.916	.816	.734	.671	.779
Flesch-1	.785	.921	.809	.610	.691	.763
Flesch-2	.753	.871	.695	.521	.712	.710

The Lexile ratings of difficulty were obtained from PC-LEX (Horabin, 1987).

All other ratings of difficulty were obtained from Readability Calculations (1984).

The Accuracy of the Lexile Equation

Classical reliability coefficients such as Cronbach's alpha and retest coefficients yield inflated estimates of score reliability. When the intention is to generalize an observed score to a universe in which items and occasions are random, then the most appropriate coefficient to report is the alternate forms reliability which for the Lexile Test of Reading Comprehension (Stenner, Smith, Horabin, and Smith; 1987) is .95.

However, in order to provide information regarding a score's accuracy, we must go beyond reporting a reliability coefficient and provide information regarding a confidence interval (SEM) for a person's score or a text's difficulty rating. The size of the standard error of measurement will affect our confidence in a single score and will help us estimate just how many more measurements are needed to approximate a student's true score more precisely. As pointed out by Anderson and Davison (1988), this is especially important when an individual's ability score is to be matched with the difficulty rating for text materials.

A generalizabilty analyses for items x group was performed using GENOVA (Crick and Brennan, 1982). The results illustrated that person scores are highly generalizable over items and occasions with the average SEM being approximately 102 Lexiles.

Not only should a confidence interval be supplied for a person's score, but a confidence interval should be established for a text's rated difficulty. A generalizability study was performed in order to determine the precision of a Lexile rating and in order to determine the number of samples needed to produce a stable estimate of difficulty for novels. The reliability of

the rating increases and the SEM for text difficulty decreases as the number of samples taker is increased (Table 4).

For ten samples of 225 words each, a confidence interval of ± 60 Lexiles (59.51) is obtained. A minimum of six 225 word samples is recommended which produces a reliability of .90 with a confidence interval of ± 75 Lexiles (74.83).

Now that evidence exists that the Lexile theory constitutes an objective scal , provides a well specified operationalization of reading comprehension, and yields reliable and generalizable scores, the question of its utility and application becomes important.

TABLE 4

Reliability Coefficients for Lexile Ratings of Text Difficulty

Over Multiple Samples of 225 Words

Number of 225 Word Samples	Reliability Coefficient
2	.750
3	.818
4	.857
5	.882
6	.900
7	.913
8	.923
9	.931
10	.938

Applications of the Lexile Scale

One of the biggest 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 grade. A student's performance is typically reported as a percentile. A percentile of .65 for a sixth grade girl indicates that she scored better than 65 percent of all sixth grade students involved in the norming study. However, percentile scores on standardized reading tests 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 sixth graders?

The Lexile scale is designed to provide both a normative and a criterion referenced interpretation of a score. Because 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 individual's ability and the item's difficulty. This relationship is captured in the following equation:

$$p = \exp(b-d) / [1 + \exp(b-d)]$$

Where p = the probability of a correct response

Where b = the ability of an individual

Where d = the difficulty of a task or test item

If a person's ability is equal to the item's difficulty, then the Lexile scale states that the individual has a 75% chance

of getting the item correct. If twenty such items were given to this individual, one would expect three fourths of the responses to be correct. If the item is more difficult than the person is able, then the probability is less than 75% that the response of the person to the item will be correct; similarly, if the item is easier compared to a person's ability, then the probability is greater that the response will be correct.

A student with a Lexile ability of 600 who is given a test item rated at 600 Lexiles of difficulty will have a 75% chance of getting the item right. If the same student is given an item of 400 Lexile difficulty, the odds improve to a 90% chance of a correct response. Give the same student a 200 Lexile item, and the odds of success improve to 96%. The more a person's Lexile score surpasses the Lexile rating for a passage or test item, the higher the probability that the person will read the passage with understanding. The more the Lexile rating for a passage, book or item surpasses a reader's Lexile score, the lower the chances the reader will understand what is read. Tables 5 and 6 illustrate the relationship between ability and difficulty and the resulting success rates.

TABLE 5

Success Rates for the Same Individual

With Materials of Varying Difficulty

Lexile Ability	Text Difficulty	Sample Titles	Predicted Success Rate
1000	600	(Old Man and the Sea - Hemingway)	96%
1000	800	(The Time Machine - Wells)	90%
1000	1000	(Reader's Digest)	75%
1000	1200	(Encyclopedia)	50%
1000	1400	(The Washington Post)	25%
1000	1600	(New England Journal of Medicine)	10%

TABLE 6

Success Rate of Different Ability Individuals

With the Same Material

Lexile Ability	Text Rating for Sports Illustrated	Predicted Success Rate
600	1000	25%
800	1000	50%
1000	1000	75%
1200	1000	90%
1400	1000	96%

Note that it is the difference in Lexiles between the person and item that governs the probability of success, and it does not matter where on the Lexile scale the difference occurs. The difference between a 200 Lexile item and a 400 Lexile reader results in the same success rate as with a 600 Lexile passage and an 800 Lexile reader. Each case produces a 90% success rate.

Empirical evidence supporting a 75% target success rate as opposed to say a 50% or 90% rate is limited. Squires, Huitt, and Segars (1983) did find that reading achievement for second graders peaked when the success rate reached 75%. A 75% success rate is also supported by the findings of Crawford, King, Brophy, and Evertson (1975). However, it may be that there is no one optimal rate, but rather a range exists in which individuals can operate successfully and improve their reading ability.

Because the Lexile theory provides complementary procedures for measuring reading ability and assessing the difficulty of reading material, the scale can be used to match a student's level of comprehension with books that the student could read with a high success rate. Up to this time, trying to identify possible supplemental reading for students has, for the most part, been guess work. 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. However, the teacher has no way of knowing whether or not that biography is too difficult or too easy for the student. The Lexile system can provide a measure of the student's reading ability as well as a measure of the biography's difficulty. Armed with this information, a teacher or parent can insure a student's success

rate with selected books.

To improve students' success in reading requires that they read properly targeted prose accompanied by frequent response requirements. Response requirements range from asking a more competent reader occasional questions as the reader progresses through the prose to questions being embedded 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 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 embedded response requirements (Anderson, Hiebert, Scott, and Wilkinson; 1984).

As the reader becomes more and more comfortable with text at a particular level, then the text difficulty can be slowly increased to match the growing comprehension ability of the reader. 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 a 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 basically tuning out. Because of the tremendous emphasis placed on reading in daily activities, virtually every encounter with written text is a reconfirmation of the low reader's inadequacy. Is it any wonder that 15-20% of U.S. high school students decide to find some other way to spend their days (Hahn, 1987)?

In order to assist individuals in becoming competent readers, they need to be exposed to text that results in a comprehension rate of 75% or better. If a 900 Lexile reader is faced with 1100 Lexile text (resulting in a 50% comprehension rate), there will be too much unfamiliar vocabulary and too much of a load placed on short term memory for the 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 with comprehension; when improperly targeted, they struggle with the material and struggle with maintaining their self-esteem. In reality, there are no poor readers -- there are only mistargeted readers who are being inappropriately "challenged."

The Lexile Scale Applied to Adult Literacy

In an information oriented society, the ability to read and comprehend adult oriented material is necessary for daily survival. Today's information needs force us to redefine what constitutes basic adult literacy. Depending upon where we draw the literacy line, a sizable proportion of students and adults are not capable of handling the demands placed upon them.

Where should we draw the literacy line? In an attempt to provide some direction, a series of real-world selections of prose were analyzed to discover what Lexile range covered adult-to-adult daily communication. Based upon the 1985 World Almanac's listing of periodical distribution, popular magazines were sampled and analyzed using a computer program developed specifically for obtaining Lexile values of continuous prose. Twenty-six of the top thirty magazines with the greatest distribution were obtained, randomly sampled, and analyzed. In addition, sixteen more magazines were randomly selected from the list, sampled and analyzed making a total of 16,796 words sampled from forty-two periodicals. Table 7 presents a list of the magazines sampled along with their respective sample size. The mean Lexile value obtained from this analysis was 1153 Lexiles with a standard deviation of 159.26.

TABLE 7

Sample Sizes of Analyzed

Magazines Ranked by Popularity

Popularity Ranking	Periodical Title	# of Words Sampled	
1	Reader's Digest	401	
2	TV Guide	319	
3	National Geographic	275	
4	Better Homes & Gardens	329	
5	Family Circle	343	
6	McCalls	421	
7	Woman's Day	304	
8	Good Housekeeping	312	
9	Time	467	
10	National Enquirer	278	
11	Playboy	377	
12	Redbook	460	
13	Star	421	
14	Penthouse	397	
15	Newsweek	439	
16	Cosmopolitan	609	
17	People	455	
18	Prevention	300	
19	Sports Illustrated	346	
21	Southern Living	427	
22	Smithsonian	371	
23	US News & World Report	368	
24	Field and Stream	307	
26	Popular Science	391	
27	Ebony	459	
30	Parents	423	
33	Popular Mechanics	296	
36	Glamour	406	
40	Outdoor Life	335	
49	Mademoiselle	397	
54	Vogue	521	
58	Rolling Stone	452	
62	Travel and Leisure	288	
67	Car and Driver	433	
68	Psychology Today	296	
75	Essence	392	
79	Motor Trend	349	
93	Esquire	402	
100	Life	349	
107	The New Yorker	532	
114	Ms.	530	
115	The Atlantic Monthly	448	
MEAN		391	1153 Lexiles
SD		76	159 Lexiles
TOTAL		16,796	

This same approach was taken with five nationally recognized newspapers (Table 8). The mean Lexile rating obtained was 1248 with a standard deviation of 96.72 Lexiles. Whether or not this same value will adequately describe the difficulty of local newspapers is still to be tested.

One further analysis was made in order to target the range of adult-to-adult communications. Various pieces of continuous prose were collected that represent materials encountered by adults on a daily basis. These included insurance forms, welfare and job applications, tax manuals, first-aid pamphlets, political advertisements, recipes, directions for assemblying a child's toy, menus, etc. Each piece of prose was assigned to one of three categories or indexes: Health/Safety Information, Consumer/Business Information, and General Information. After classification, the materials were sampled and entered into the computer as one entire piece of continuous prose. This would allow the assignment of one Lexile value to each area of interest.

The results of this analysis are reported in Table 9. The mean sample size for the three indexes was 69.2 words. The mean Lexile value obtained was 1041 with a standard deviation of 57.67 Lexiles.

TABLE 8

Sample Sizes of Five Major Daily Newspapers

Periodical Title	Sample Size	
Wall Street Journal	415	
Washington Post	301	
Christian Science Monitor	303	
New York Times	311	
USA Today	363	
MEANS	339	1248 Lexiles
SD	50	97 Lexiles
TOTAL	1693	

TABLE 9

Lexile Ratings and Sample Sizes of Three Adult Communication Indexes

Index	Lexile	Sample Size
Health/Safety Index	1004	8520
General Information Index	1025	7303
Consumer Business Index	1094	5032
MEANS	1041	6952
SD	58	2168
TOTALS		20855

By comparing these results (Table 10), it would appear as if adults communicate with one another between 1050 Lexiles and 1250 Lexiles. If the weighted average of these materials is taken as a bench mark, then the minimal reading level for an adult functioning in our society is approximately 1100 Lexiles for a 75% comprehension rate. Based on an analysis of three nationally normed tests of reading comprehension (Horabin, 1989), the fiftieth percentile graduating senior is reading at an average of 1022 Lexiles. This finding implies that a large portion of young adults cannot read adult oriented materials with a 75% success rate. Most likely many members of this initially disinfranchised group go on to acquire a reading level of 1100 Lexiles, but the vast majority do not. If a person leaves school with a reading level below 800 Lexiles, their prognosis is very poor because they will encounter very little text in the adult world that is written at 800 Lexiles. The result is very little practice reading with comprehension and thus very little improvement in their reading ability.

On the other hand, the prognosis for a 1000 Lexile reader is better because they can acquire adult to adult communications that are at or near the level that they can read with 75% comprehension. Sufficient exposure to this type of text either because their job requires it or the individual is interested in self improvement can result in these individuals attaining 1100 Lexile status.

TABLE 10

Summary of the Lexile Analysis of Adult Reading Materials

Source	Mean	SD	Range	Minimum	Maximum	N
Magazines	1153	159.26	801	727	1528	42
Newspapers	1248	96.72	223	1141	1364	5
Indexes	1041	57.67	90	1004	1094	3
Arithmetic Mean	1147.33					
Weighted Mean	1097.72					

Conclusion

Not until the development of the Lexile theory has reading comprehension been measured on an interval scale with a constructed zero point. Just as on the Farenheit scale, anchor points for the Lexile scale are reproducible and help provide meaning to the scores.

The Lexile theory and accompanying scale do not provide answers to all of the questions related to reading comprehension just as the temperature scale cannot be used to explain all phenomena associated with the weather. However, the Lexile theory does clearly identify and explain the essence of what happens when a reader interacts with a text. Both the temperature scale and the Lexile scale provide a metric for understanding basic elements associated with their respective constructs.

References

Anderson, R. C. & 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., & 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. & 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., & Shankweiler, D. (1988). Syntactic complexity and reading acquistion. In A. Davison 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., & Evertson, 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. & Brennan, R. L. (1982). GENOVA: A generalized analysis of variance system [computer program]. Dorchester, MA: University of Massachusetts at Boston.

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

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

Dunn, L. M. & 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. & 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. (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: Iowa State University Press.

Liberman, I. Y., Mann, V. A., Shankweiler, D. & Werfelman, M. (1982). Children's memory for recurring linquistic and non-linquistic 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.

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

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

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

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

Smith, F. (1973). Psycholinguistics and reading. New York.: Holt, Rinehart & Winston.

Squires, D. A., Huitt, W. G., & Segars, J. K. (1983). Effective schools and classrooms. Alexandria, VA: Association for Supervision and Curriculum Development.

Stanley, J. C. (1971). Reliability. In R.L. Thorndike (Ed.,, Educational measurement: 2nd edition. Washington, DC: American Council on Education.

Stenner, A. J., Smith, D. R., Horabin, I. and Smith, M., (1987). The Lexile Test of Reading Comprehension. Durham, NC: MetaMetrics.

Stenner, A. J., Smith, D. R., Horabin, I. and Smith, M., (1987, December). Fit of the Lexile theory to item difficulties on fourteen standardized reading comprehension tests. Paper presented at the National Reading Conference, St. Petersburg, FL.

Stenner, A. J. & Smith, M. (1982). Testing construct theories. Perceptual and Motor Skills, 55, 415-426.

Stenner, A. J., Smith, M. and Burdick, D. S. (1983). Toward a theory of construct definition. Journal of Educational Measurement, 20, 305-316.

Survey of Basic Skills: Form P (1984). Chicago: Science Research Associates.

Thorndike, R. L. Personnel selection. New York: Wiley, 1949.

Woodcock, R. W. (1973). Woodcock reading mastery test. Circle Pines, MN: American Guidance Service.

Woodcock, R. W. & Johnson, M. B. (1977). Woodcock-Johnson psycho-educational battery. Allen, TX: DLM Teaching Resources.

Wright, B. D., Rossner, M., & Congdon, R. T. (1985). M-Scale: A Rasch Program for Ordered Categories [computer program]. Chicago: Meca Press.

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