Journal of Applied Measurement

Constructing Variables

Volume 7, Number 3, 2006 *ISSN 1529–7713*

Editor

Richard M. Smith ..... Data Recognition Corporation

Associate Editor

Everett V. Smith ..... University of Illinois, Chicago

Editorial Board

David Andrich ..... Murdoch University
Rita Bode ..... Rehabilitation Institute of Chicago
Trevor Bond ..... Hong Kong Institute of Education
Madhabi Chatterji ..... Teachers College, Columbia University
Karon Cook ..... University of Washington, CORR
Ayres D'Costa ..... Ohio State University
Dimiter Dimitrov ..... George Mason University
Barbara Dodd ..... University of Texas, Austin
George Engelhard, Jr. ..... Emory University
William P. Fisher, Jr. ..... Avatar International, Inc.
Anne Fisher ..... Umëa University, Sweden
Gunnar Grimby ..... University of Goteborg
Perry N. Halkitis ..... New York University
Allen Heinemann ..... Rehabilitation Institute of Chicago
Scott Hershberger ..... California State University, Long Beach
George Karabatsos ..... University of Illinois, Chicago
William Koch ..... University of Texas, Austin
Joanne Lenke ..... ETS K-12 Works
Alain LePlege ..... Institut d'Histoire et de Philosophie des Sciences, France
J. Michael Linacre ..... University of Sydney, Australia
Larry H. Ludlow ..... Boston College
Randall MacIntosh ..... California State University, Sacramento
George Marcoulides ..... California State University, Fullerton
Geofferey Masters ..... Australian Council on Educational Research
David McArthur ..... UCLA School of Medicine
Ronald J. Mead ..... Data Recognition Corporation
Carol Myford ..... University of Illinois, Chicago
Steven P. Reise ..... University of California, Los Angeles
E. Matthew Schulz ..... ACT, Inc.
Gregory Stone ..... University of Toledo
Mark H. Stone ..... Adler School of Professional Psychology
Alan Tennant ..... University of Leeds
Luigi Tesio ..... Istituto Auxologico Italiano, Milan, Italy
Craig Velozo ..... University of Florida
Mark Wilson ..... University of California, Berkeley
Edward W. Wolfe ..... Virginia Tech
Fred Wolfe ..... University of Kansas
Weimo Zhu ..... University of Illinois

JOURNAL OF APPLIED MEASUREMENT, 7(3), 307-322
Copyright© 2006

How Accurate Are Lexile Text Measures?

A. Jackson Stenner

Hal Burdick

Eleanor E. Sanford

Donald S. Burdick

Metametrics, Inc.

The Lexile Framework for Reading models comprehension as the difference between a reader measure and a text measure. Uncertainty in comprehension rates results from unreliability in reader measures and inaccuracy in text readability measures. Whole-text processing eliminates sampling error in text measures. However, Lexile text measures are imperfect due to misspecification of the Lexile theory. The standard deviation component associated with theory misspecification is estimated at 64L for a standard-length passage (approximately 125 words). A consequence is that standard errors for longer texts (2,500 to 150,000 words) are measured on the Lexile scale with uncertainties in the single digits. Uncertainties in expected comprehension rates are largely due to imprecision in reader ability and not inaccuracies in text readabilities.

Requests for reprints should be sent to A. Jackson Stenner, MetaMetrics, Inc., 1000 Park Forty Plaza Drive, Suite 120, Durham, NC 27713, USA, e-mail: jstenner@lexile.com.

The Lexile Framework for Reading exploits the conjoint measurement properties of the Rasch model to put reader ability and text readability on the same developmental scale. Reading comprehension—that is, the rate at which a reader makes new meaning—is modeled as the exponential difference between reader ability and text readability.

The Framework works by anchoring reading item difficulties to a Rasch scale via theory instead of the common practice of finding item difficulties empirically. The precision with which we predict these item difficulties is one index of the quality of our theory. This paper describes in detail the idea of a construct theory, why we care about it, how we use the theory, and finally how we conceptualize and estimate text measurement error.

Theory-Referenced Measurement

A construct theory is the story we tell about what it means to move up and down the scale for a variable of interest (e.g., temperature, reading ability, short-term memory). Why is it, for example, that items are ordered as they are on the item map? The story evolves as knowledge increases regarding the construct. We call both the process and the product of this evolutionary unfolding "construct definition." Advanced stages of construct definition are characterized by calibration equations (or specification equations) that operationalize and formalize a construct theory. These equations make point predictions about item behavior. The more closely theoretical calibrations coincide with empirical item difficulties, the more useful the construct theory and the more interesting the story (Stenner, Smith, and Burdick, 1983).

Reflection of 25 years of experience in developing the Lexile Framework for Reading (described in detail later in this paper) enables us to distinguish five stages in our thinking. Each subsequent stage can be characterized by an increasingly sophisticated use of substantive theory. Evidence that a construct theory and its associated technologies have reached a given stage or level can be found in the artifacts and instruments that are realizable at each level.

Level 1

At this stage there is no explicit theory as to why items are ordered as they are on the item map. Data are used to estimate both person measures and item difficulties. Just as with other actuarial sciences, empirically determined probabilities are of paramount importance. When data are found to fit a Rasch Model, relative differences among persons are independent of which items or occasions of measurement are used to make the measures. A familiar artifact of this stage is the scale annotated with empirical item difficulties. Most educational and psychological instruments in use today are level-one technologies.

Level 2

A construct theory can be formalized in a specification equation used to explain variation in item difficulties. If what causes variation in item difficulties can be reduced to an equation, then a vital piece of construct validity evidence has been secured. We argue elsewhere that the single most compelling piece of evidence for an instrument's construct validity is a specification equation that can account for a high proportion of observed variance in item difficulties (Stenner, Smith, and Burdick, 1983). Without such evidence, only very weak correlational evidence can be marshaled for claims that "we know what we are measuring" and "we know how to build an indefinitely large number of theoretically parallel instruments that measure the same construct with the same precision of measurement."

Note that we test the causal status of a specification by experimentally manipulating the variables in the specification equation and checking to see whether the expected changes in item difficulty are, in fact, observed. Stone (2002) performed just such an experimental confirmation of the specification equation for the Knox Cube Test-Revised, when he designed new items to fill in holes in the item map and found that the theoretical predictions coincided closely with observed difficulties. Can we imagine a more convincing demonstration that the construct theory and its associated specification equation accord well with observations (Stenner and Smith, 1982;

Stone and Wright, 1983; Stenner and Stone, 2003)?

Similar demonstrations have now been realized for hearing vocabulary (Stenner, Smith, and Burdick, 1983), reading (Stenner and Wright, 2002), quantitative reasoning (Enright and Sheehan, 2002), and abstract reasoning (Embretson, 2002). Artifacts that signal level-2 use of theory are specification equations, RMSEs from regressions of observed item difficulties on theory, and evidence for causal status based on experimental manipulation of item design features.

Level 3

The next stage in the evolving use of theory involves application of the specification equation to enrich scale annotations. One example of this use of the specification equation is in the measurement of text readability in the Lexile Framework for Reading. In this application, we conceptualize a book as a test made up of as many items as there are paragraphs in the book. We then use the specification equation to generate theoretical calibrations for each paragraph, which stands in for empirical item difficulties (Stone, Wright, and Stenner, 1999).

For instance, the text measure for a book is the Lexile reader measure needed to produce a sum of the modeled probabilities over paragraphs, qua items, equal to a relative raw score of 75%. We can imagine a thought experiment in which every paragraph (say 900 in total) in a Harry Potter novel is turned into a reading test item. Each item is then administered to 1000 targeted readers and empirical item difficulties are computed from a hugely complex connected data collection effort. The text measure for the Harry Potter novel (880L) is the amount of reading ability needed to get a raw score of 675/900 items correct, or a relative raw score of 75%. The choice of 75% is arbitrary, but it has proved to have useful behavioral consequences.

The specification equation is used in place of the tremendously complicated and expensive realization of the thought experiment for every book or article we want to measure. The machinery described above can also be applied to text collections (book bags or briefcases) to enable scale annotation with real world text demands (college, workplace, etc.).

Artifacts of a level-3 use of theory include construct maps that annotate the reading scale with texts that, thanks to theory, can be imagined to be tests with theoretically derived item calibrations.

Level 4

In biochemistry, when a substance is successfully synthesized using amino acids and other building blocks, the structure of the purified entity is then commonly considered to be understood. That is, when the action of a natural substance can be matched by that of a synthetic counterpart, we argue that we understand the structure of the natural substance. Similarly, we argue that when a clone for an instrument can be built and the clone produces measures indistinguishable from those produced by the original instrument, then we can claim that we understand the construct under study. What is cumulative in the history of science is the gradual refinement of instrumentation (Ackerman, 1985).

In a level-4 use of theory, there is enough confidence in the construct theory and associated specification equation that a theoretical calibration takes the place of an empirical item difficulty for every item in the instrument or item bank. There are now numerous reading tests (e.g., Scholastic Reading Inventory-Interactive, 1999, and the Pearson PASeries Reading Test, 2004) that use only theoretical calibrations. Evidence abounds that the reader measures produced by these theoretically calibrated instruments are indistinguishable from measures made using the more familiar empirically scaled instruments.

Level 5

Level-5 use of theory builds on level 4 to handle the case in which theory provides not individual item calibrations but rather a distribution of "potential" item calibrations. Again, the Lexile Framework has been used to build reading tests that incorporate this more advanced use of theory. Imagine a *Time* magazine article that

is 1500 words in length. Imagine a software program that can generate a large number of "cloze" items for this article. A sample from this collection is served up to readers when they choose to read this article. During reading, the reader fills in the blanks (missing words) that are distributed throughout the article. How can counts correct in such an experience produce Lexile reader measures, when it is impossible to effect a one-to-one correspondence between a reader response to an item and a theoretical calibration, specific to that particular item? The answer is that the theory provides a distribution of possible item calibrations (specifically, a mean and standard deviation), and a particular count correct is converted into a Lexile reader measure by integrating over the theoretical distribution.

Text Readability

The "old" method of computing readability on a text involved sampling 100 words each from the beginning, middle, and end of a book and applying, by hand, a readability formula to this 300-word sample. Repeating the process on different 100-word samples from the same text results in a distribution of values, and the standard deviation of this distribution is the standard error of measurement (SEM). The "old" method produced uncomfortably large SEMs, which caused some researchers to despair regarding the value of text measures in research and practice. Noe and Standal (1984) wrote, "The problem with using a readability formula to make judgments about text difficulty is that formula validity is low, and precision is limited" (p. 673). Duffelmeyer (1985) concurred, stating, "Readability formulas can be of assistance, but there is the danger of granting them more precision than they merit. . . . Most reading educators would grant them no more discriminating power than one grade level" (p. 392). This paper contrasts the "old" method of estimating text readability with a new method and shows that this new method greatly increases the precision of text difficulty estimates. In addition, this article introduces the concept of the ensemble interpretation, which directly addresses the issue of formula validity. Finally, we compute SEMs for the new method and discuss implications of the method for theory and practice. We begin by detailing a readability system that uses this new method: The Lexile Framework for Reading.

The Lexile Framework for Reading

The Lexile Framework for Reading, in short, is the blending of the one-parameter Rasch model with a readability formula, the Lexile specification equation.

Reading is the most tested construct in education. It is probable that reading ability is measured more frequently than temperature, height, or weight among students aged 6 to 18. Reading ability is widely recognized as the best predictor of success in higher education and on-the-job performance. Economists and educators have joined in identifying low literacy rates as a primary causal factor in the United States' dwindling economic productivity (Snow, 2002). In an information age, reading is a survival skill.

The *Ninth Mental Measurements Yearbook* (Mitchell, 1985) reviewed 97 reading tests. Associated with each of these tests is a conceptual rationale (however primitive) and a scale. Thus, there are 97 different nonexchangeable reading metrics. The incommensurable diversity of the status of reading measurement is reminiscent of late 17th-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. Similarly, the presence of dozens of competing reading scales produces confusion among educators, researchers, policy makers, and parents. The Lexile Framework represents an attempt to unify the measurement of reading by offering a common, supplemental metric shared by test publishers, book publishers, and text aggregators (e.g., EBSCO and Proquest).

Human beings communicate using various symbol systems such as those of mathematics, music, and language. Many symbol systems can be seen as having two components: a semantic one and a syntactic one. In mathematics, the semantic units are numbers and operators, which

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 readability of a text passage is determined by the familiarity of the semantic units and by the complexity of the syntactic structures used in constructing the message.

In the case of reading, it is clear that most operationalizations of the semantic component are proxies for the probability that an individual will encounter a word in a familiar context and thus be able to infer its meaning (Bormuth, 1966). This is the basis of exposure theory, which explains the development of receptive or hearing vocabulary (Miller and Gildea, 1987; Stenner, Smith and Burdick, 1983). Klare (1963) built the case for the semantic component varying along a familiarity-to-rarity continuum, a concept that was further developed by Carroll, Davies, and Richman (1971), whose word-frequency study examined the reoccurrence of words in a 5-million-word corpus of running text. Knowing the frequency of words as they are used in written and oral communication provides the best means by which to infer the likelihood that a word will be encountered and thus become a part of an individual's receptive vocabulary.

Sentence length is a proxy for the syntactic complexity of published text and is highly correlated with the difficulty that readers have in comprehending text. One important caveat is that sentence length is not the underlying causal influence (Chall and Dale, 1995). Researchers sometimes incorrectly assume that manipulation of sentence length will have a predictable effect on passage difficulty. Davison and Kantor (1982) showed that sentence length can be reduced and difficulty increased and vice versa. One suggestive approach to explaining these results has been offered by recent studies of hierarchical complexity that focus on variation in levels of conceptual meaning (Bond, 1994; Dawson, 2002). Future research will need to explore the extent to which accounting for variation in hierarchical complexity might also contribute to reduced error in readability measures.

Klare (1963) provided a possible interpretation for the impact of sentence length in predicting passage difficulty. He speculated that the syntactic component varies with the load placed on short-term memory. This explanation was supported by Crain and Shankweiler (1988), whose work provided evidence that sentence length is a good proxy for the demands that structural complexity places upon verbal short-term memory.

The Lexile Framework uses a 2-variable equation to forecast text difficulty. The word frequency and sentence length measures described above combine to produce a regression equation that explains a high proportion of the observed variance of reading comprehension task difficulties (80%+). The equation has been shown to explain variation in item difficulties on a wide range of tests with varying item formats (Stenner and Wright, 2002).

A reader with a measure of 600L who is given a text measured at 600L is expected to have a 75% comprehension rate. This comprehension rate is the basis for selecting text that is matched to a student's reading ability, but what exactly does that mean? In addition, what would the comprehension rate be if this same student were given a text measured at 350L or one at 850L?

The 75% comprehension rate for a reader-text encounter can be given an operational meaning by imagining the text to be carved into item-sized slices of approximately 125 words each with a question embedded in the slice. A reader who reads these slices and answers three fourths of the questions correctly has a 75% comprehension rate. Thus, text is imagined to be a virtual test comprising n slices/items and associated n slice/item calibrations.

A Rasch equation is solved for the measure that, when evaluated in the context of the slice/item calibrations (given by theory), produces as a sum of the modeled probabilities of a correct

answer a relative raw score of 75%. This measure is defined as the text readability or the text measure. The difference in Lexiles between the reader and the text governs comprehension. If the reader measure is greater than the text measure, the comprehension rate will exceed 75%. If the reader measure is less than the text measure, the comprehension rate will be less than 75%. The exact relationship of the Lexile difference between reader, text, and comprehension rate depends on the spread of the text difficulties of the slices within an individual text just as the relationship between raw score and measure depends, in part, on the spread of the item calibration. This relationship can be expressed as a raw score given the following equation:

$$\text{Raw Score} = \sum_{i=1}^{n} \frac{e^{(b-d_i+1.1)}}{1+e^{(b-d_i+1.1)}}. \quad (1)$$

The result of Equation 1 can then be converted to a comprehension rate (percentage) by dividing by the number of slices and multiplying by 100. The Lexile Framework uses a modification of the Rasch model in which a match between reader ability and text difficulty on a single slice results in an expected comprehension rate of 75% instead of 50% as in a typical Rasch model application.

The Lexile Analyzer

To assess the accuracy of a readability formula, a process for producing text measures must first be designed and implemented. This section details the process used by the Lexile Framework to determine text difficulty.

Text measures are realized by submitting a text file to the Lexile Analyzer. The software is available (without charge for noncommercial use such as educational research) at www.lexile.com. The steps completed by the software to measure a text are as follows:

1. An auto-edit routine is performed on the text to remove unfamiliar characters, figures, tables, and other nontext features;
2. The text file is "sliced" into standard-sized paragraphs of 125 words;
3. Each word in the slice is looked up in a frequency dictionary based on a 550-million-word corpus and the mean of the log word frequencies is computed for the slice;
4. The log of the mean sentence length is computed for the slice;
5. The two variables (from steps 3 and 4 above) are entered into an equation that returns a Lexile calibration for the slice;
6. This process is repeated for each slice in the text file;
7. The text is then treated as a virtual test with the number of test items equal to the number of slices and the item calibrations equal to the slice calibrations;
8. A measure is then returned that answers the question, "How well would a reader have to read (in Lexiles) to answer correctly 75% of the imagined test items comprising this text?";
9. The answer to the above question is the text measure assigned to the text.

Thus, text readability is reinterpreted as a question about how well a reader must read to produce a relative raw score of 75% on a virtual test. The raw score is modeled as the sum of the modeled probabilities, and, because item calibrations and the raw score are known, the Rasch model is iterated and solved for the reader measure needed to produce a relative raw score of 75%. Text readability is seen as a function of a relative raw score set by convention (75%) and a set of slice calibrations, provided by a text theory, treated as if they were test item calibrations. Treating a text as if it were a test makes it possible to use the same psychometric model for computing both reader measures and text measures. One equation knits together and permits separation of three important reading concepts: text readability, reader ability, and expected comprehension rate.

Uncertainty in the Lexile measure assigned to a text comes from the fact that each slice calibration is imperfect, due to theory misspecification. The Lexile Theory embodied in the specification equation that is used to calibrate text slices does not provide a perfect prediction of how difficult it

will be to comprehend a slice (*qua* paragraph) of text. This imperfection in the Lexile Theory and its calibrations translates directly into uncertainty in the Lexile text measure that is assigned to a book or article. The primary purpose of the present study is to estimate the standard deviation component (the square of which is a variance component) that is associated with Lexile theory misspecification when predicting ensemble means. The Lexile Theory purports to explain the vast majority of variation in slice difficulties when those difficulties are expressed as means over all allowable ways of response-illustrating each text slice (i.e., the ensemble). The following section discusses the implications of theory misspecification for computing standard errors in text measure and contrasts the accuracy of this method for computing text measures with older methods.

The Ensemble Interpretation

The validation of readability formulae generally proceeds in the following fashion. Using an item-development protocol (i.e., a set of rules for writing reading test items), writers generate a set of reading items, administer these items to a sample of readers, compute the difficulty of each item, and regress the observed item difficulties on theory-based measures of text readability. Researchers then report the correlation, or its square, as a validity coefficient. When unsquared validity coefficients like these are computed for familiar readability formulae, they are found to be in the range of .70 to .94 (Stenner and Wright, 2002). In the case of the Lexile equation, the root-mean squared error (RMSE) is 150L (or equivalently, .833 logits)[1] when regressing observed item difficulties on theoretical calibrations.

Authors in several publications have asserted that individual item difficulties are what text readability theories and equations should predict (Stenner and Burdick, 1997; Stenner, Horabin, Smith, and Smith, 1988; Stenner and Smith, 1982; Stenner, Smith, and Burdick, 1987). This body of research has noted that most text readability equations are good to excellent predictors of the difficulty of published text and attributed the RMSE of 150L to an elusive third variable beyond semantics and syntax (e.g., paragraph coherence or conceptual complexity).

An alternative explanation for the error of 150L begins with the recognition that a passage of text can be understood as entailing a large number of macro propositions (Kintsch and Van Dijk, 1978), each of which can be used as the basis for building a reading comprehension item. The set of all allowable items and their contexts of use (ways they are used to build tests) is the ensemble for a particular text. Each ensemble member (item) has an associated difficulty, and the average over the distribution of these difficulties is the ensemble mean. The Lexile Theory claims that these ensemble means are predictable from knowledge of the semantic and syntactic features of text passages. The ensemble interpretation eliminates the myriad details attached to each ensemble member not by explaining variation in these details (although this course remains a theoretically interesting and practically useful direction), but rather by averaging over the details. The result of the averaging is a new concept (ensemble mean) removed from the particulars of its creation and is the unit of text readability predicted by the Lexile Theory.

The ensemble interpretation shifts the focus from the individual item difficulties to the distribution of all allowable ways of turning a text passage into a reading item (i.e., the ensemble[2]). It can now be argued that the relevant datum for testing theories about text difficulty should be the ensemble mean. Item writers unwittingly add easiness and hardness to the theoretical text calibration when writing items. Similarly, the position of an item in a test adds easiness or hardness, which produces variation that is not accessible to the readability formula because that formula is applied only to the text passage. Another way to examine this idea is to realize that the theoretical calibration for a text passage is the same no matter how many item writers make different reading items out of that passage or in how many different contexts the items might be used. Although item writers produce different reading items with different observed item difficulties for a given passage, the theoretical text calibration remains the same (Figure 1).

Passage (430L)	Observed Difficulty	Embedded Completion Items				
She disappeared through	269L	I am glad she is _____.	gone	first	best	sitting
the trees. "Fine with me,"	632L	I was _____.	upset	happy	polite	hungry
I thought angrily. It would	704L	I _____ her.	disliked	forgot	told	chased
be fine with me if I never						
saw her again.						

Figure 1. Illustration of a 3-item ensemble.

In the case of the Lexile Framework, the ensemble interpretation has made it possible to pass from a level-of-item and item-writer dependence to a theory of text that is no longer so dependent. Ensemble means do not depend on who writes an item, how foils are chosen, the context in which the item is used, or any other initial conditions. The Lexile Theory replaces statements about individual items with statements about ensembles. The ensemble interpretation enables the elimination of irrelevant details. The extra-theoretical details are taken into account jointly, not individually, and, via averaging, are removed from the data text explained by the theory. A consequence of the ensemble interpretation working in cooperation with whole-text processing (every word in a book is analyzed—there is no sampling) is a reduction in the uncertainty associated with text measures when compared to older methods of computing text readability.

An embellishment to this approach recognizes that there is a finitely large set of possible virtual tests that could be built for a particular text. Comprehension is defined as an expectation over this distribution of virtual tests. Note that there is a repeatable process for "virtually" turning each slice in a text into a native-Lexile item and collecting these items into a virtual test. Each repetition of this process yields another virtual test. However, to produce an observed item difficulty for one member of an ensemble for a passage, that item must be "contextualized." That is, it must be located on an actual test and administered to a sample of readers. Therefore, variability in a single item's difficulty arises from a 3-stage process. First, an item writer decides on one of many macro propositions as the focus of the item and chooses a correct answer and three foils. Different item writers, or the same item writer, reusing the text slice and repeating the item writing protocol, produce different items nested within each slice. Second, these items are contextualized by being located on a test form (early in the test or late in the test). Third, the difficulties of the items are estimated for a sample of examinees (small, large) using a psychometric model. As each stage in the process is executed and then repeated for each new item, ensemble variance is produced. This is called "within-ensemble variance" (engineering variance) because it includes all of the sources associated with building and using items. Control over the item-writing process involves both a construct theory that is capable of explaining variation in true ensemble means and an understanding of how much ensemble variance is generated in the process of constructing and using an item and the sources of this variation.

Expected comprehension rate is an average success rate on a specific item type (see Figure 1). The triple averaging that produces the ensemble mean and variance is taken over persons (conditioning on ability), items, and contextualizations. One ensemble member from each slice ensemble could be selected and collected into a virtual test. However, text readability should not be defined in terms of a comprehension rate on this one test. Rather, the reader measure associated with a 75% raw score when averaged over all allowable such tests should be used. This reader measure is by definition the readability of the text. The Lexile slice calibrations produced by the Lexile Analyzer can be used because these calibrations predict the ensemble mean. Stated differently, the ensemble mean taken over all possible persons, items, and

contextualizations is seen as a function of the semantic and syntactic features of the text, as operationalized in the Lexile Analyzer.

The purpose of this paper is to assess the uncertainty that should be assigned to text measures. If the measures contain a lot of uncertainty, then users must apply caution when targeting a text to a reader. If, on the other hand, the Lexile text measures are highly accurate, then most applications that employ these measures might ignore this small amount of uncertainty. When using common instrumentation (e.g., thermometers, rulers, clocks, barometers, odometers) for everyday applications, the reality of measurement error is ignored because it is small compared to other differences. It would be useful if the uncertainty in Lexile text measures were likewise small enough, relative to other detectable differences, that these errors of measurement could similarly be ignored.

Method

Participants

Participants in this study were students from four school districts in a large southwestern state. These students were participating in a larger study that was designed to assess reading ability in the Lexile metric. The total sample included 1,186 grade 3 students, 893 grade 5 students, and 1,531 grade 8 students. The mean tested abilities of the three samples were similar to the mean tested abilities of all students in each grade on the state reading assessment. Though 3,610 students participated in the linking study, the data records for only 2,867 of these students were used to determine the ensemble item difficulties presented in this paper. The students were administered 1 of 4 forms at each grade level. The reduction in sample size is because 1 of the 4 forms was created using the same ensemble items as another form. For consistency of sample size across forms, the data records from this fourth form were not included in the ensemble study.

Instrument

Thirty text passages (see Figure 1) were response-illustrated by three different item-writing teams, resulting in three items nested within each of 30 passages for a total of 90 items. All three teams employed the same item-writing protocol. The ensemble items were spiraled into test forms at the grade level (3, 5, or 8) that most closely corresponded with the item's theoretical calibration.

Winsteps (Wright and Linacre, 2003) was used to estimate item difficulties for the 90 ensemble study items. Of primary interest in this study was the correspondence between theoretical text calibrations and the 30 ensemble means and the consequences of theory misspecification on text measure standard errors.

Results

Table 1 presents results from an earlier study in which well-estimated item difficulties for 700 native-Lexile items were regressed on theoretical calibrations produced by the Lexile Analyzer. The RMSE resulting from predicting observed item difficulties from theory is in the range of 150 to 160 Lexiles. These data are included as background for interpreting the ensemble data that follows. The RMSEs in Table 1 do not answer the question about the predictive power of the Lexile equation. As mentioned previously, raw item difficulties are not the right dependent variable for checking the predictive power of construct theories despite earlier claims to the contrary (Stenner, Smith, and Burdick, 1983).

Table 2 presents the ensemble study data in which three independent teams wrote one item for each of 30 passages for 90 items. Observed ensemble means taken over the three ensemble item difficulties for each passage are given along with an estimate of the within ensemble standard deviation for each passage.

The difference between passage text calibration and observed ensemble mean is provided in the last column. The RMSE from regressing observed ensemble means on text calibrations is 110L. Figure 2 shows a plot of observed ensemble means against theoretical text calibrations.

Note that some of the deviations about the identity line are because ensemble means are poorly estimated given that each mean is based on only three items. The bottom panel in Figure

Table 1

Fit of the Lexile Theory to 700 Item Difficulties

Item Type	Grade	*N* of Items	RMSE
Study 1	4	40	154L
	5	40	151L
	6	40	160L
	7	60	150L
	8	60	151L
	9	60	137L
Study 2	1	40	163L
	2	40	173L
	3	45	170L
	4	54	150L
	5	54	150L
	6	54	198L
	7	70	172L
	8	70	191L

Note: The observed difficulties for the 700 items were based on over 10,000 responses per item.

2 depicts simulated data when an error term [distributed ~$N(0, \sigma = 64L)$] is added to each theoretical value. Contrasting the two plots in Figure 2 provides a visual depiction of the difference between regressing observed ensemble means on theory and regressing "true" ensemble means on theory. An estimate of the RMSE when "true" ensemble means are regressed on the Lexile Theory is 64L ($110^2 - 89^2 = \sqrt{4{,}308} = 64$). This is the average error at the passage level when predicting "true" ensemble means from the Lexile Theory.

Because the RMSE of 64L applies to the expected error at the passage/slice level, a text made up of n_i slices would have an expected error of $64 \div \sqrt{n_i}$. Thus, a short periodical article of 500 words ($n_i = 4$) would have an SEM = $64 \div \sqrt{4} = 32$L, whereas a much longer text, such as a Harry Potter novel composed of 900 slices, would have an SEM = $64 \div \sqrt{900} = 2$L. Table 3 contrasts the SEMs computed using the old method with SEMs computed using the Lexile Framework for several books across a broad range of Lexile measures.

Discussion

Measurement is the process of converting observations (counts) into quantities (linear amounts) via a substantive theory (Stenner and Burdick, 1997). In the case of Lexile text measures, word frequency counts and sentence length counts are transformed and averaged to produce proxies for the semantic and syntactic constructs in the Lexile Theory. These variables are combined in a linear equation to produce slice calibrations (quantities) denominated in Lexiles. These slice calibrations are then used as the substantive theory for converting a relative raw score (75%) into a Lexile measure, which is taken to be the text readability. The results suggest that text readabilities that are computed using newer methods and the reconceptualization of the contribution of theory misspecification to uncertainty in text measures combine to produce more than an order of magnitude reduction in text measure standard errors. Single-digit standard errors (reported in Lexiles) are sufficiently small that text measure uncertainty can be effectively ignored in many applications of the Lexile Framework. The difference between text readability and reader ability is used to forecast comprehension. Uncertainty in reader abilities, as opposed to text readabilities, is by far the major source of error in expected comprehension.

Ordering text along a single continuum of readability requires that we ignore a legion of common sense qualifications, any one of which seemingly disrupts and discredits the simplicity

Table 2

Analysis of 30-Item Ensembles Providing an Estimate of the Theory Misspecification Error

Item Number	Theory (T)	Team A	Team B	Team C	Mean[a] (O)	*SD*[b]	Within Ensemble Variance	T-O
1	400L	456	553	303	437	126	15,909	–37
2	430L	269	632	704	535	234	54,523	–105
3	460L	306	407	483	399	88	7,832	61
4	490L	553	508	670	577	84	6,993	–87
11	510L	267	602	468	446	169	28,413	64
5	540L	747	825	654	742	86	7,332	–202
6	569L	909	657	582	716	172	29,424	–147
7	580L	594	683	807	695	107	11,386	–115
8	620L	897	805	497	733	209	43,808	–113
9	720L	584	850	731	722	133	17,811	–2
12	720L	953	587	774	771	183	33,386	–51
13	745L	791	972	490	751	244	59,354	–6
14	770L	855	1017	958	944	82	6,717	–174
16	770L	1077	1095	893	1022	112	12,446	–252
15	790L	866	557	553	659	180	32,327	131
21	812L	902	1133	715	917	209	43,753	–105
10	820L	967	740	675	794	153	23,445	26
17	850L	747	864	674	762	96	9,257	88
22	866L	819	809	780	803	20	419	63
18	870L	974	1197	870	1014	167	28,007	–144
19	880L	1093	733	692	839	221	48,739	41
23	940L	945	1057	965	989	60	3,546	–49
24	960L	1124	1205	1170	1166	41	1,653	–206
25	1010L	926	1172	899	999	151	22,733	11
20	1020L	888	1372	863	1041	287	82,429	–21
26	1020L	1260	987	881	1043	196	38,397	–23
27	1040L	1503	1361	1239	1368	132	17,536	–328
28	1060L	1109	1091	981	1061	69	4,785	–1
29	1150L	1014	1104	1055	1058	45	2,029	92
30	1210L	1275	1291	1014	1193	156	24,204	17

Total MSE = Average of $(T - O)^2$ = 12022; Pooled within variance for ensembles = 7984; Remaining between ensemble variance = 4038; Theory misspecification error = 64L

Barlett's test for homogeneity of variance produced an approximate chi-square statistic of 24.6 on 29 degrees of freedom and sustained the null hypothesis that the variances are equal across ensembles.

Note: All data is reported in Lexiles.

a. Mean (O) is the observed ensemble mean.

b. *SD* is the standard deviation within ensemble.

of a unidimensional representation of text readability. Yet, this is precisely what has been asserted: text can be ordered on the basis of the difficulty encountered by readers in trying to make meaning, and this ordering can be predicted with remarkable precision from an analysis of two key text characteristics—word familiarity and sentence length. This ordering of text difficulty then has important implications for reading instruction.

When encounters between readers and text produce data that fit the Lexile Framework, then reader measures and text measures can be put on a common scale (Luce and Tukey, 1964). Additionally, differences among reader measures are seen as independent of which particular texts/items are used to measure the readers and, symmetrically, differences in readability among texts are seen as independent of which particular readers are used to estimate text difficulty. One consequence of expressing reader ability and text readability in the same unit of measurement (e.g., Lexiles) is that differences in reader ability can be traded off for differences in text readability to

hold constant the expected comprehension rate. Concretely, an 880L reader is forecasted to have 75% comprehension of a novel measured at 880L. If during the school year this reader adds 120L of reading ability, texts must be chosen at 1000L (120L of added text difficulty) to maintain a 75% comprehension rate for the reader. Thus, a difference in text readability can be substituted for a difference (change) in reader ability to hold constant the comprehension rate that is considered optimal for instructional purposes. In summary, conjoint additivity is defined by an absence of a conditional relationship (interaction) between reader and text. Claims regarding reader ability need not be conditioned by reference to a particular text or text characteristic. Similarly, claims about a text's readability need not be conditioned by reference to any particular reader characteristic(s).

The specific objectivity, realized when reading data fit the Lexile Framework, and the averaging process involved in ensemble interpreta-

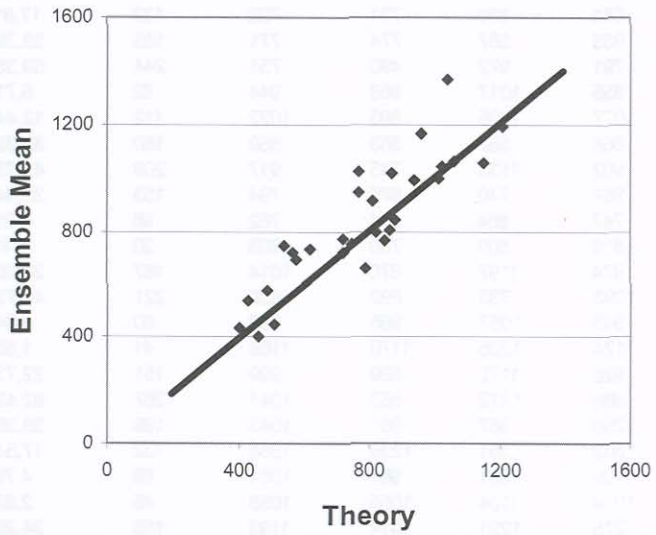

Figure 2a. Plot of observed ensemble means and theoretical calibrations (RMSE = 111L).

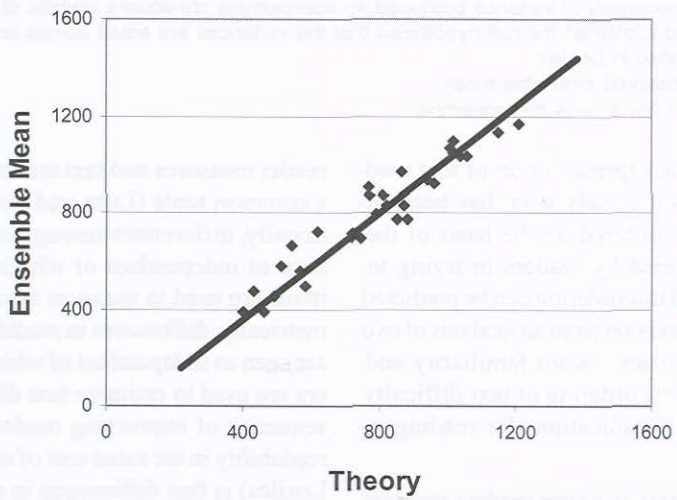

Figure 2b. Plot of simulated "true" ensemble means and theoretical calibrations (RMSE = 64L).

tions both involve parameter separation and summarization. In each case, an invariant generalization has been abstracted from the initial conditions. Specific objectivity is

> the requirement that the amount and meaning of a measure has been well enough separated from the measuring instrument and the occasion of measurement [initial conditions] so that the measure can be used as a quantity without qualification as to which was the particular instrument or what was the specific occasion. (Wright and Stone, 1999, p. 7)

In measuring reading ability, an occasion and a collection of items are necessary; the meaning of the measure depends on these specifics "disappearing from consideration." When interpreting a measurement, we routinely assume that this separation of measure and instrument has been realized and, therefore, that the measure transcends the particulars of the measuring process. Similarly, the ensemble interpretation frees the text measure from initial conditions. The ensemble interpretation states that the Lexile Framework and associated theory explain variation in true ensemble means. A mean is a summary statement about a distribution. Once computed, this summary is independent of the particulars that were averaged over (e.g., examinees and item features). Although summarizing all of the ways that a particular text passage can be turned into a reading item requires ignoring massive amounts of detail, it is exactly this summary statement that is so well predicted by the Lexile Theory.

For a decade, the data text has been misread and an RMSE of 150L at the individual item level has been misinterpreted as evidence for the incompleteness of the Lexile Theory. We reasoned that there must be unmeasured characteristics of the text that, once conceptualized and measured, would reduce the RMSE to levels commensurate with a completeness argument (not necessarily RMSE = 0). Failure to find the elusive third variable prompted us to reconceptualize the problem by fusing a hierarchical theory (items and their use nested within passage) with the notion of an ensemble (Einstein, 1902; Gibbs, 1902). Failure to explain all item variation at the individual item level is now seen as a necessary outcome of the ensemble interpretation. A slice of text consists of several connected macro propositions, any one of which could become the focus of an item writer's attention. The resulting item might be contextualized as an easy item on a hard test or a hard item on an easy test. Such item-writing decisions and contextualizations generate within-ensemble variance. Once the potentiality of many allowable response illustra-

Table 3

Old Method Text Readabilities, Resampled SEMs, and New SEMs for Selected Books

Book	Number of Slices	Lexile Measure	Resampled Old *SEM*[a]	New *SEM*
The Boy Who Drank Too Much	257	447L	102	4
Leroy and the Old Man	309	647L	119	4
Angela and the Broken Heart	157	555L	118	5
The Horse of Her Dreams	277	768L	126	4
Little House by Boston Bay	235	852L	126	4
Marsh Cat	235	954L	125	4
The Riddle of the Rosetta Stone	49	1063L	70	9
John Tyler	223	1151L	89	4
A Clockwork Orange	419	1260L	268	3
Geometry and the Visual Arts	481	1369L	140	3
The Patriot Chiefs	790	1446L	139	2
Traitors	895	1533L	140	2

Note: Three slices selected for each replicate: one slice from the first third of the book, one from the middle third, and one from the last third. Resampled 1,000 times. *SEM* = *SD* of the resampled distribution.

tions is collapsed into one, we are on the average assured of unexplained variation because within-ensemble variation is not accessible to a theory focused on semantic and syntactic characteristics of text. Whatever choices are made by the item writer and instrument maker, the theory makes the same prediction regarding difficulty for every member of the ensemble.

Lexile text measures are an order of magnitude more accurate than those produced by older technologies that rely on text sampling instead of whole-text processing. One consequence of this increased accuracy in measuring text readability is that error in expected comprehension rates is seen as due to unreliability in Lexile reader measures and not to inaccuracies in Lexile text measures. Many generalizations about reader behavior are text-dependent generalizations (i.e., "good" readers summarize what they read better than "poor" readers), whereas, other generalizations are text independent (i.e., "good" readers evidence larger sight vocabularies than "poor" readers). In the text-independent case, generalizations about readers are not conditioned on the text the reader is reading; whatever the text, the "good" reader will evidence a larger sight vocabulary than the "poor" reader. In the text-dependent case, generalizations only make sense in the context of a particular text or text readability. "Good" readers in this second sense are those who demonstrate good comprehension, which, as we have seen, is dependent on the difference between reader ability and text readability. Good comprehenders can summarize better than poor comprehenders and thus the assertion that "good" readers (*qua* high comprehenders) summarize better than "poor" readers is a text-dependent generalization. When making text-dependent generalizations, it is comforting to know that text readability can be measured with high accuracy.

Finally, we wish to reiterate that this paper considers only one source of error in expected comprehension rates, namely, Lexile Theory misspecification. Text readability measures for texts of typical lengths (books, periodicals, articles) are measured with a high degree of accuracy. Uncertainty in expected comprehension rates for a particular reader encountering a specific text is a joint function of reader measure reliability and text measure accuracy. These uncertainties are largely due to unreliability in reader measures and not to inaccuracy in text measures. We leave for another paper the issues of reader measure reliability, the generalizability of reader growth trajectories, and the contribution made by reader-measure unreliability to the accuracy of expected comprehension rates.

Footnotes

[1] The Lexile scale is a transformed logit scale where 1 logit equals 180 Lexiles.

[2] It may have occurred to the reader that a Hierarchical Linear Model (HLM) would be a useful model for analyzing this type of data. Level 1 would be the item level and level 2 would be the ensemble level. Level 2 predictors would be dominated by variables operationalizing the construct theory and level 1 predictors would focus on item engineering variables (e.g., what foils are chosen, how is location indeterminacy resolved, where is the item located on the test). We in fact have a paper in process that develops this perspective. We wanted the current paper to have a simple message with a straightforward methodology.

References

Ackerman, R. J. (1985). *Data, instruments, and theory: A dialectical approach to understanding science.* Princeton University Press.

Bond, T. G. (1994). Piaget and measurement II: Empirical validation of the Piagetian model. *Archives de Psychologie, 63*, 155-185.

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

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

Carver, R. P. (1974). Measuring the primary effect of reading: Reading storage technique,

understanding judgments and cloze. *Journal of Reading Behavior*, *6*, 249-274.

Chall, J. S., and Dale, E. (1995). *Readability revisited: The new Dale-Chall readability formula.* Cambridge, MA: Brookline.

Crain, S., and Shankweiler, D. (1988). Syntactic complexity and reading acquisition. In A. Davidson and G. M. Green (Eds.), *Linguistic complexity and text comprehension: Readability issues reconsidered* (pp. 167-192). Hillsdale, NJ: Lawrence Erlbaum.

Davison, A., and Kantor, R. (1982). On the failure of readability formulas to define readable texts: A case study from adaptations. *Reading Research Quarterly*, *17*(2), 187-209.

Dawson, T. L. (2002). New tools, new insights: Kohlberg's moral reasoning stages revisited. *International Journal of Behavioral Development*, *26*(2), 154-166.

Duffelmeyer, F .A. (1985). Estimating readability with a computer: Beware the aura of precision. *The Reading Teacher*, *38*, 392-394.

Einstein, A. (1902). Ann. d. Phys., IV. *Folge*, *9*, 417.

Embretson, S. E. (2002). Generating abstract reasoning items with cognitive theory. In S. H. Irvine and P. C. Kyllonen (Eds.), *Item generation for test development* (pp. 219-250). Mahwah, NJ: Lawrence Erlbaum.

Enright, M. K., and Sheehan, K. M. (2002). Modeling the difficulty of quantitative reasoning items: Implications for item generation. In S. H. Irvine and P. C. Kyllonen (Eds.), *Item generation for test development* (pp. 129-157). Mahwah, NJ: Lawrence Erlbaum.

Gibbs, J. W. (1902). *Elementary principles in statistical mechanics.* Yale, CT: Yale University Press.

Kintsch, W., and Van Dijk, T. A. (1978). Toward a model of text comprehension and production. *Psychological Review*, *85*(5), 363-394.

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

Luce, R. D., and Tukey, J. W. (1964). Simultaneous conjoint measurement: A new type of 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.* Lincoln, NE: University of Nebraska Press.

Noe, K. S., and Standal, T. C. (1984). Readability: Old cautions for the new technology. *The Reading Teacher*, *37*, 673-674.

Pearson Educational Measurement (2004) *PASeries reading test.* Iowa City, IA: Author.

Scholastic (1999). *Scholastic reading inventory.* New York: Author.

Snow, C. E. (2002). *Reading for understanding toward a research and development program in reading comprehension.* Santa Monica, CA: Rand.

Stenner, A. J., and Burdick, D. S. (1997). *The objective measurement of reading comprehension: In response to technical questions raised by the California Department of Education technical study group.* Unpublished manuscript.

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*, 765-769.

Stenner, A. J., and 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*(4), 305-315.

Stenner, A. J., Smith, M., and Burdick, D. S. (1987). *Fit of the Lexile theory to item difficulties on fourteen standardized reading comprehension tests.* Durham, NC: MetaMetrics.

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

Stenner, A. J., and Wright, B. D. (2002). Readability, reading ability, and comprehension. In B. D. Wright and M. H. Stone (Eds.), *Making measures* (pp. 79-115). Chicago: Phaneron.

Stone, M. H. (2002). *Knox cube test-revised.* Itasca, IL: Stoelting.

Stone, M., and Wright, B. D. (1983). Measuring attending behavior and short-term memory with Knox's cube test. *Educational and Psychological Measurement*, *43*(3), 803-804.

Stone, M. H., Wright, B. D., and Stenner, A. J. (1999). Mapping variables. *Journal of Outcome Measurement*, *3*, 308-322.

Wright, B. D., and Linacre, J. M. (2003). *A user's guide to WINSTEPS Rasch-Model computer program, Vol. 3.38.* Chicago: Winsteps.com.

Wright, B .D., and Stone, M. H. (1999). *Measurement essentials* (2nd ed.). Wilmington, DE: Wide Range.

CONTRIBUTOR INFORMATION

Content: *Journal of Applied Measurement* publishes refereed scholarly work from all academic disciplines that relates to measurement theory and its application to developing variables. The construction and interpretation of meaningful and unambiguous variables is a salient feature of measurement. It represents the congruence of measurement theory and substantive research in a wide range of scientific endeavors. The development of variables that map the persons and items onto a common metric, operationally defined by the items, that are invariant across samples of persons and items, is a cornerstone of developing an understanding of the phenomena being measured and the construction and verification of hypotheses based on these phenomena. The journal will also publish invited articles that provide examples of methodological issues that are relevant to constructing useful variables.

Book and Software Reviews: The *Journal of Applied Measurement* publishes reviews of current books and software. These reviews permit scholarly assessment of current books and software. Suggestions for reviews are accepted. Original authors of reviewed books and software will be given the opportunity to respond to all reviews.

Manuscript Preparation: Manuscripts should be prepared according to the Publication Manual of the American Psychological Association (5th ed., 2001). Limit manuscripts to 30 pages of text, exclusive of tables and figures. Manuscripts must be double-spaced including the title page, abstract, text, quotes, acknowledgments, references, and appendices. On the cover page list author name(s), affiliation(s), address(es), telephone number(s), and electronic mail address(es). On the second page include a 100 to 150 word abstract. Place tables on separate pages. Include photocopies of all figures. Number all pages consecutively.

Authors are responsible for all statements made in their work and for obtaining permission from copyright owners to reprint or adapt a table or figure or to reprint a quotation of 500 words or more. Copies of all permissions and credit lines must be submitted.

Manuscript Submission: Submit four (4) manuscript copies to Richard M. Smith, Editor, *Journal of Applied Measurement*, P.O. Box 1283, Maple Grove, MN 55311 (e-mail: jampress@att.net). Prepare three copies of the manuscript for peer review by removing references to author(s) and institution(s). In a cover letter, authors should indicate that the manuscript includes only original material that has not been previously published and is not under review elsewhere. When a manuscript has been accepted the author(s) is asked to submit a final printed copy of the manuscript, original graphic files, and camera-ready figures, a copy of the final manuscript in WordPerfect or Word format on a 3½ in. disk for IBM-compatible personal computers, and sign and return a copyright-transfer agreement. Reprints of articles are available for purchase at the time the manuscript is published. Manuscripts are copy-edited and composed into page proofs. Authors are asked to review proofs before publication.

Peer Review of Manuscripts: Manuscripts are anonymously peer-reviewed by at least two reviewers appropriate for the topic and content of the article submitted. The editor will maintain the anonymity of the author(s) and reviewers during the review process. The review process normally requires three to four months.

SUBSCRIBER INFORMATION

Journal of Applied Measurement is published four times a year and is available on a calendar year basis. Individual subscriptions are $49.00 per year. Institutional subscriptions are available for $122.00 per year. There is an additional $28.00 charge for air mail postage outside of the United States and Canada. Funds are payable in U.S. currency. Send subscription orders, information requests, and address changes to *Journal of Applied Measurement*, P.O. Box 1283, Maple Grove, MN 55311 USA. Claims for missing issues will not be honored beyond 6 months after original mailing date. Duplicate copies cannot be sent to replace issues not delivered due to failure to notify publisher of change of address. Back issues are available at a cost of $15.00 per issue postpaid in the United States and Canada. An additional $6.00 is required for airmail postage to other locations. Please send requests to the address listed above.

Copyright© 2006, Richard M. Smith, Editor. No part of this publication may be used, in any form or by any means, without permission of the publisher. Printed in the United States of America. ISSN 1529–7713.