1280L | Lexile: Matching readers to text

WHY DO SCORES CHANGE?

by Gary L. Williamson, Ph.D.

A white paper from The Lexile Framework for Reading

April 2004

Table of Contents

Overview 3

What is Measurement? 3

What is a "Construct?" 5

Why Measurements Vary for an Individual 6

What is "Measurement Error?" 7

Consequences of Measurement Error 10

Conclusion 11

About the Author 12

About The Lexile Framework for Reading 12

References 13

Overview

Since 1990, educational accountability systems have been widely implemented in the United States. The focus on accountability recently gained new emphasis with the reauthorization of the Elementary and Secondary Education Act (ESEA) signed into law by President Bush on January 8, 2002. The law, usually called the No Child Left Behind Act of 2001 (NCLB), put in place sweeping requirements for increased accountability in the public schools of the United States. A central feature of the new law is the requirement for annual assessments of students in reading and mathematics.

Because of the new federal requirements as well as state testing programs that were already in place in many states, the academic performance of students in the United States is perhaps more widely measured now than at any previous time in history. With more frequent measurement, parents and teachers have access to more information about their students' performance than at any previous time. With the increased availability of information, parents and teachers are better informed than in the past. Ironically, they may also find themselves having more questions about the results than at any time in the past.

This white paper addresses some fundamental questions that parents or teachers may have about the scores that their students get on tests of reading or mathematics achievement, and why scores may change. As a starting point, it is useful to consider what kind of information is produced by tests of academic achievement and how that differs from familiar physical measurements that we use in our everyday lives.

What is Measurement?

Measurement is the assignment of a number to an object for the purpose of describing some attribute of the object in terms of its quantity. For example, we measure a stick to determine its length in inches. Of course, it is not quite that simple. How we assign the number and how we interpret it are important and are major themes in the history of measurement theory and practice. So, understanding some aspects of measurement can help us understand why scores vary.

Two familiar examples are the measurement of height and weight. Height and weight are physical attributes of human beings. In the case of height, our procedure

for assigning a number might be to use a standard yardstick, measure the distance from the top of a person's head to the ground, and record the number of inches measured. The number of inches tells us how tall the individual is. In this case, the attribute being measured is height and the measurement procedure involves the use of a familiar device — a yardstick. In the case of weight, we might use a standard bathroom scale, have the individual stand on the scale and read the number of pounds to quantify the attribute of weight of the individual.

These examples are simple and familiar today. However, it took many years to develop and agree on the rules for measurement of height and weight. The development of rulers, scales and similar devices for physical measurement occurred over time. Even today, the specific numerals assigned to height or weight are not universal. In the United States, we would express the measurements in inches and pounds, but in Europe we would probably use centimeters and kilograms to express our measurements of the same attributes. Still, there are simple mathematical relationships between inches and centimeters and between pounds and kilograms, so that the measurements are clear.

The measurement of height is simple to think about because we can lay a yardstick end to end to measure distances that aren't too large (like the height of a person.) The measurement process involves an operation (i.e., positioning the yardstick) that can be *concatenated* (i.e., linked together in a series) to produce the measurement. The measurement of weight has a similar property if we use a scale with a balance beam, where we simply add additional counter-weights until the beam balances to arrive at the measurement. In both cases, using procedures that can be concatenated to produce a measurement, we assume that the component operations are *additive*. That is we can count the number of counter-balancing weights or the number of times we lay the yardstick end to end and add them to arrive at the total weight or height of the object. This additive property of the measurement procedure produces scales that have *equal intervals*.

However, some physical attributes are not easy to observe directly, or to measure with an additive operation. In such cases, scientists often try to find some approach that approximates these mathematical properties so as to have a measurement process that behaves as if it is an equal interval measurement. A good example of such a physical attribute is temperature. Temperature has to do with the level of molecular activity in a substance. That is certainly not readily observable, although some physical states change at certain temperatures (e.g., the appearance of water changes when the temperature reaches the boiling point). However, in general, we cannot know its temperature from simply looking at an object. The molecular activity takes place in three dimensions beyond our powers of direct observation.

So, what did scientists do to measure temperature? They found a substitute for direct observation. They noticed that the length of a column of mercury in a thermometer could accurately indicate the temperature of water. This effectively

allows us to assign a number for the temperature of water through direct observation of the level of mercury in the thermometer. It also allows us to reduce the three-dimensional nature of molecular motion to the one-dimensional (length) measure of the column of mercury. Additionally, because the measure is based on determining the length of the column of mercury, it has the additive property that measurements of length have.

What is a "Construct?"

Generally, measurement of physical attributes is straightforward because we can directly observe them (height, weight) or arrange to measure something that is observable (column of mercury) that indicates the property of interest (temperature.) This makes the measurement fairly obvious in most cases.

However, in education we are often interested in attributes that we cannot directly observe, such as a student's reading ability or mathematical understanding. We cannot observe these attributes directly but we believe that they exist when students can read and do mathematics. So we *construct* the attributes in theory (hence psychologists call them "constructs") to explain human behavior.

Measuring a psychological construct such as reading ability is more difficult than measuring height because we don't directly observe reading ability. It is more like measuring temperature. We cannot see it directly, but we can find a substitute that we can directly observe. We can ask students to read some text and then question them about what they have read. Then we *can* observe their responses to our questions. Achievement tests are designed to collect observable responses that we can use to infer how much a student has learned. So, a test of reading ability allows us (through a well-designed procedure) to assign a number to the attribute "reading ability" for an individual.

The measurement procedures used to assign numbers to psychological constructs were fairly simple during most of the last century. Educators simply counted the number of correct responses a student gave to a specified set of questions. However, just as with height and weight, our procedures for measuring academic constructs have improved over time. Today, there are precise procedures that mathematically model what happens when a student answers a particular question. These *psychometric* theories allow us to measure (i.e., create a *metric* to denote) the amount of a *psycho*logical attribute. They provide the measurement operation for assigning the numerals that tell us how much of an attribute characterizes an individual. Furthermore, they can do this in such a way that the measurements are additive mathematically. That is, modern psychometric procedures can produce equal interval measurements of psychological constructs.

Moreover, in the last two decades, measurement of reading has progressed beyond the point of merely having equal interval scales. Now there is a unifying theory of measurement that allows different measures of reading to be placed on a common unambiguous scale (similar to what was accomplished for the measurement of temperature after many decades). This psychometric research culminated in the development of The Lexile Framework® for Reading by MetaMetrics® Inc. Examples in the following sections will use the Lexile® scale to illustrate variability in students' scores.

Why Measurements Vary for an Individual

Parents and teachers often notice that a student's score on an achievement test differs from one occasion to the next, even when the occasions are close together in time. When scores increase, it is usually *assumed* that this is because the student has learned additional material (though this is not necessarily always true — the student could have made some lucky guesses!). However, when the scores go down there is virtually always concern. Sometimes, students must retake a test (for example because they do not score well enough the first time to meet a promotion standard or some other criterion.) When this happens the second score is usually a little different than the first one. Why? There are two fundamental reasons.

- One reason that a student's score, say in reading, can change from one occasion to the next is that the amount of reading ability the student possesses has actually changed. The student has perhaps benefited from additional instruction and experience, and so is a more competent reader. In these cases, the actual amount of the construct being measured has changed. This explanation is more plausible when the time between the two test scores is sufficient for the student to have benefited from additional instruction or experience. To conclude that a change in scores is due to improvement, one has to allow enough time between measurements for improvement to take place (e.g., if you're dieting, you don't weigh yourself every hour to see how much weight you lost since the previous weighing!).
- A second reason that a student's score in reading could change is that something else affects his or her ability to respond to the test. These might be *internal* or *external* influences, but they have nothing to do with the student's actual reading ability (except that they prevent the student from responding in a way that is indicative of the true level of reading ability.) For example, the student might have been tired on testing day and was not able to respond effectively. Perhaps he or she failed to have breakfast before leaving for school and hunger proved to be a distraction. The possibilities are many. External influences can be just as varied. Perhaps there was noise in the hall while the student was taking the test. Perhaps the light was flickering and causing annoyance. Again, the possibilities are nearly endless.

The key difference in these two cases is that in the first case, the change in the student's score is relevant to the construct being measured. The change represents an actual increase in the student's reading ability. In the second case, the change in the student's score is unrelated to the true attribute of interest. This kind of change is irrelevant to the student's true reading ability, but it keeps us from getting a true picture.

Unfortunately, every time a student takes a test any of the above factors can influence the result. Consequently, when a student retakes a test, the score can vary (up or down) for reasons that are relevant to the construct of interest, or for reasons that are irrelevant.

What is "Measurement Error?"

Construct-irrelevant variance in scores is called measurement "error" by psychometricians. The errors can be systematic or random, and can come from within the student or from sources external to the student. The term "error" does not mean that the test was scored incorrectly (it could have been, but that is relatively rare). Rather, it means that the score the student obtained was influenced in part by factors other than the attribute being measured. Consequently, measurement error is of critical interest to psychometricians and they do everything within their power to minimize its effect on psychological measurement.

Measurement error is not unique to the measurement of psychological constructs. For example, if we measured a person's height repeatedly several times and recorded the results, we would have slightly different answers each time depending on a variety of "errors" that might creep into the measurement. For example, variations can occur because of differences from one occasion to the next in the person's posture, whether the person wears shoes or not and what kind, the positioning of the yardstick, the ability of the reader to clearly and consistently read the markings on the yardstick, how the position of the top of the head was determined, etc.

With cognitive constructs, like reading ability, a variety of factors can influence the result. Some were mentioned in the previous section. Systematic errors of measurement are of concern because they affect the accuracy of the measure. However, because the errors are systematic, they tend to apply just as much on one occasion as on another, and so they do not contribute to the tendency for scores to change. Random errors of measurement, however, can affect both the accuracy and consistency of the scores. Because of their effect on consistency of measurement, they affect the ability to be able to get the same score on different occasions, and so directly relate to this issue. There are an amazing variety of random errors (influences) that affect the consistency of scores. Table 1 shows some potential sources of random errors of measurement that can affect measures of academic achievement.

Table 1. *Sources of Random Errors of Measurement*

INTERNAL SOURCES	EXTERNAL SOURCES
Student's Physical State	**Irregularities in Test Administration**
Fatigue	Timing
Hunger	Interruptions
Visual Acuity	Breaks
Aural Acuity	Acoustics
Sickness	Lighting
General Alertness	Noise
Student's Psychological or Emotional State	Voice of test administrator
Anxiety	Clarity of Directions
Guessing	**Test Related**
Memory	Item sampling
Excitement	Content sampling
Attentiveness	Scoring errors
Relationships with family or friends	Inter-rater reliability
Other Sources	Intra-rater consistency
Mistakes	Reliability of test
Speed	Mode of administration
Carelessness	**Other Sources**
Perceived Importance of test	Luck
Motivation	Distractions
Misreading a question	Domestic tranquility
Clerical errors	Behavior of friends
Skipping an item by mistake	Consequences of the scores
Misunderstanding instructions	Attitude of significant adults
Misreading an item	

The sources of measurement error presented in the table are examples. They do not represent an exhaustive list of all possible sources of measurement error; nor does every one of these sources of measurement error actually apply in any given instance of measurement.

For psychometricians, consistency of measurement is called *reliability*. They prefer to use tests that are reliable, so they have devised ways to characterize the reliability of measurements. One way is to calculate an index that expresses the similarity of scores on alternate forms of a test. They give alternate (equivalent) forms of the same test to a group of students and calculate the degree of similarity between the

two sets of scores. If the two sets of scores are very similar, they say the test is very reliable. The index ranges from 0 to 1 with higher numbers indicating higher reliability. For example, within a particular grade, alternate form reliabilities range from .75 to .85 for many tests used in common practice.

Knowing then that every measurement is subject to measurement error, whether it is a physical measurement or a cognitive measurement, it is important to be able to estimate how much measurement error exists. Psychometricians do this by theoretically examining the variability of repeated measurements. They imagine that the measurement process can be performed repeatedly under the same conditions with the same individual, and estimate the variability of the scores produced. By using measurement theory, it is as if they are able to repeat the measurement process many times under the same conditions with the same individual. The resulting variability in scores is due to random errors of measurement since the individual does not change.

Psychometricians routinely do this theoretical analysis of variability for tests of academic achievement. As a result, all reputable test publishers report the *standard error of measurement* (SEM) for their tests. Table 2 shows the range (over a variety of tests) of the SEM for an average student's score. In general, we can be reasonably sure that a student's true score is within one SEM of their observed score. We can be very confident it lies within two SEM.

Table 2. *Range (Expressed in Lexiles) of Standard Errors of Measurement for Selected Reading Tests by Selected Grades*

Grade	2	4	6	8	10
	99L-125L	88L-135L	72L-133L	80L-153L	85L-127L

Notes:
- Tests included the Stanford Achievement Test (9th edition), Stanford Diagnostic Reading Test (4th edition), Metropolitan Achievement Test (8th edition), Stanford Achievement Test (10th edition), TerraNova, and the Gates-McGinite Reading Test.
- SEMs were derived from alternate forms reliabilities obtained from the respective publishers, and data used to link the target test with The Lexile Framework for Reading.

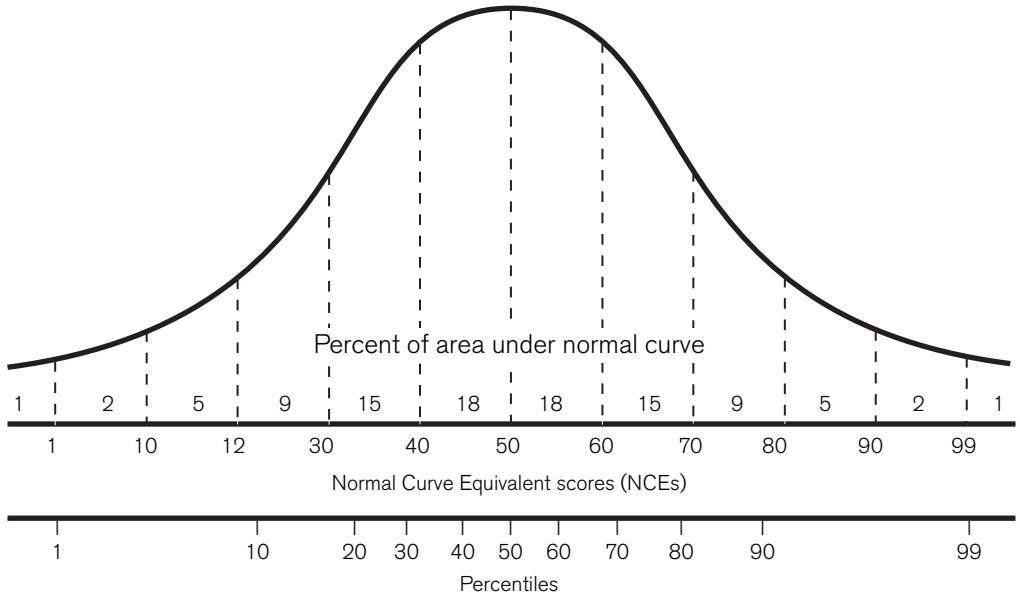
Knowing the standard error of measurement for a test is useful, but it does not answer the question, "How much measurement error is reasonable?" That requires a subjective judgment, and it depends on the consequences of measurement error in any given context.

Consequences of Measurement Error

There are at least three contexts in which measurement error is relevant for a student. These are relative comparisons (such as *norm-referenced* interpretations of a student's score — i.e., where a student ranks within a group), absolute comparisons (such as whether a student attains a certain performance level), and in making instructional decisions (e.g., deciding what level of text a student should read). The consequences of measurement error vary accordingly.

- A norm-referenced interpretation of a score describes where the student ranks in a group — usually a representative sample of students in the nation who are in the same grade as the student. Sometimes the reference group consists of the students in a given state if the state develops its own achievement tests, as is the case in North Carolina. When such interpretations are made, the direct consequence of measurement error is that the ranking attributed to the student's score may be in error. How much depends on the standard error of measurement of the score and the relative position of the score in the distribution of scores that comprise the norm group. Since percentile ranks are more tightly packed near the middle than at the extremes of the scale (see Table 3), small errors in scores near the median score can translate into large errors in the ranking attributed to the student's score.

Table 3. *Norm-referenced interpretations of scores*



- When absolute decisions are made about a student's score, the practical consequence of measurement error depends critically on where the student scores in relation to the standard (that is the "passing" score.) If the student's score is near the standard, then even small errors of measurement can result in incorrect classification of the student as meeting or not meeting the standard.
- When instructional decisions are involved, such as deciding what level of text a student should be reading, measurement error should be taken into account to project a range of appropriate material for the student. Otherwise the student may not receive sufficient challenge to stimulate interest and improvement, or conversely may be over-taxed and become discouraged and disaffected.

As a more concrete example of a situation where an absolute decision is made, suppose we are trying to classify a student into one of four proficiency categories —Below Basic, Basic, Proficient or Advanced — based on the student's score. If the score is very close to the boundary of one of the categories, say the boundary between Basic and Proficient, then even a small amount of measurement error has consequences for our ability to accurately classify the student. The magnitude of the measurement error also affects our level of confidence in our ability to make the same decision again were the student to be tested a second time.

It is also worth noting that test-based decisions involve high stakes more often with some contexts than others. Normative decisions and absolute decisions are more often associated with higher stakes for the student — for example, admission into a program (based on a normative comparison) or passing a standard (an absolute decision). Instructional decisions usually entail lower stakes because such decisions are easy to change quickly in the classroom if a mistake is made. The consequences of measurement error are thus greater when high stakes are involved.

Conclusion

We know that measurement error can affect a student's score. Whether it does in any given instance, and how much it does, is more difficult to know. Usually, standardized tests are given in uniform, rigorously controlled environments to minimize the influence of external factors unrelated to the construct being measured. Students are typically encouraged to get sufficient rest and nourishment prior to testing to minimize internal distractions. However, ultimately it is up to students, parents and teachers to reflect critically on the scores and to view them in the context of broader knowledge about the student.

To the extent that scores are consistent with the student's performance on other related measures or on other occasions, then a given score interpretation has support. Whenever a score is inconsistent with other information, then it should be viewed cautiously. A logical analysis will often provide insight into the student's performance, and the standard error of measurement will provide a context

for exercising the proper caution in interpreting the score. In most cases, parents and teachers can also consult the directors of the school district or state testing programs for assistance in understanding the testing program, the reliability of the tests and specific results for their students.

About the Author

Gary L. Williamson, Ph.D., is senior research associate with MetaMetrics, Inc. With more than 30 years of experience in educational research on the academic, state and school district levels, Williamson's specialty is quantitative methodology encompassing psychometric, mathematical and statistical applications to educational data. He most recently was educational research and evaluation unit director at the North Carolina Department of Public Instruction. He has written and spoken extensively on the subjects of educational assessment and accountability. Williamson earned both a doctorate of philosophy in mathematical methods for educational research and a master's of science in statistics from Stanford University. He also holds a master's of education in educational research and evaluation from The University of North Carolina at Greensboro, and a bachelor's of science in mathematics from The University of North Carolina at Chapel Hill.

About The Lexile Framework for Reading

The Lexile Framework for Reading (www.Lexile.com) provides a common scale for matching reader ability and text difficulty, allowing easy monitoring of progress. Lexile measures give teachers and parents the confidence to choose materials that will improve student reading skills across the curriculum and at home. Tens of thousands of books and tens of millions of articles have Lexile measures, and all major standardized tests can report student reading scores in Lexiles. As the most widely adopted reading measure in use today, Lexiles are part of reading and testing programs at district, state and federal levels. The Lexile Framework was developed by MetaMetrics, an independent education company based in Durham, N.C., after 15 years of research funded by the National Institutes of Health.

References

Crocker, L. & Algina, J. (1986). "Introduction to classical and modern test theory." New York: Holt, Rinehart and Winston.

Nunnally, J.C. (1978). "Psychometric theory." New York: McGraw-Hill Book Company.

Stanley, J.C. (1971). "Reliability." In R.L. Thorndike (Ed.), Educational Measurement (2nd Ed.), pp. 359-442. Washington, D.C.: American Council on Education.

Stenner, A. J. (1996). "Measuring reading comprehension with the Lexile Framework." Paper presented at the California Comparability Symposium, Burlingame, CA.

Thorndike, R.L. (1951). "Reliability." In E.F. Lindquist (Ed.), Educational Measurement (pp. 560-620). Washington, DC: American Council on Education.

Torgerson, W.S. (1958). "Theory and methods of scaling." New York: John Wiley.

Wright, B.D. & Stenner, A. J. (1999). "One fish, two fish: Rasch measures reading best." Popular Measurement (Spring, pp. 34-38).

Lexile: Matching readers to text

LEXILE®

www.Lexile.com
1.888.LEXILES