1360L | Lexile: Matching readers to text

THE NEED FOR OBJECTIVE MEASUREMENT UNDER THE NO CHILD LEFT BEHIND ACT

by Malbert Smith III, Ph.D.

A white paper from The Lexile Framework for Reading

April 2004

Table of Contents

Overview

With the passage of the No Child Left Behind Act, Congress reauthorized the Elementary and Secondary Education Act (ESEA) — the principal federal law affecting education from kindergarten through high school. In amending ESEA (commonly referred to as No Child Left Behind, or NCLB), the new law represents a sweeping overhaul of federal efforts to support elementary and secondary education in the United States.

Some of the major provisions of NCLB include:

- Accountability for results
- Expanded local control and flexibility
- Requiring every child to be on-grade level/proficient by the end of the 2013-2014 school year
- Expanded parental options
- Ensuring every child can read
- Adequate yearly progress (AYP) standards
- Promoting English proficiency

One of the major weaknesses of reading education today is the lack of meaningful measurement systems. The key in the "hard" sciences is unification of measurement. In the case of the measurement of temperature in the 1600s, there were literally dozens of instrument makers with their own scale. However, once a theory of temperature had been developed and accepted, measurement unification was possible. Today, it is inconsequential whether a temperature is taken with a thermometer purchased at CVS or K-Mart — the scale is independent of the manufacturer of the instrument.

As stated in the 2003 report from the Northwest Evaluation Association (NWEA):

> In some states lack of consistency between grade levels poses a serious problem that can be avoided if standards are calibrated. Standards that are not calibrated give students, parents, and educators an inaccurate perception about the child's standing relative to the expected level of performance. Students are reported as proficient in one grade who may not remain proficient in later grades even if they show normal growth.
>
> Assume Xavier, for example, is a 3rd grader living with his family in Yuma, Arizona. Xavier scores at about the 46th percentile in mathematics on his

state assessment, which is the minimum score for a rating of proficient, or meets standard. Xavier's teacher and parents believe he is performing at a level that is satisfactory relative to grade level standards.

Now let's move forward to 8th grade. Xavier takes the state level assessment again and achieves at the same level of performance, a 46th percentile score, relative to other students. This does not put Xavier anywhere close to the level required to meet the standard. Xavier's parents are alarmed that he is no longer meeting grade level standards and his teachers may come in for criticism because Xavier's performance "slipped." But Xavier's performance didn't slip. Instead he was the victim of a poorly calibrated standard that was too low at 3rd grade to reflect the performance that would be needed by 8th grade (Kingsbury, et. al.).

The Lexile Framework® for Reading is a scientific approach to measuring reading ability and text difficulty. All of the major test publishers have linked their norm-referenced tests so that they can report out Lexile® measures to students and parents. Each year, millions of students receive a Lexile measure from one of the "instrument" makers. Tens of thousands of trade and textbooks have Lexile measures, and tens of millions of articles have Lexile measures through library database-service providers.

MetaMetrics® believes that its unifying efforts for the measurement of reading can address the issues in NCLB.

Measurement and Accountability

Many of the assessment issues and concerns that have typically been of interest only within the psychometric community have now become more visible with the high-stakes assessments of NCLB. For example, the fairly standard professional and industry guidance in not relying exclusively on a single test score has become memorialized in standard 13 in Standards for Educational and Psychological Testing (National Council on Measurement in Education, 1999). Simply stated, "in educational settings a decision or characterization that will have a major impact on a student should not be made on a single test score." Recognizing this standard and principal, Section 111 of NCLB requires that the mandated assessments in grades three through eight should employ "multiple, up-to-date measures of student academic achievement" (Koretz, 2003).

A number of researchers have attempted to provide educational practitioners with advice on how to handle the multiple-measure requirements within their accountability models. As these researchers currently point out, the need for multiple measures arises out of the recognition that measurement instruments are not infallible, and thus, should be interpreted within a range of uncertainty that our reliability estimates indicate. While we should view any score from an

assessment through the prism of reliability theory, Baker (2003) rightly points out that multiple measures should be viewed from a validity perspective as well. As Baker states, "despite multiple measures' paternity in reliability arguments, the mother of multiple measures is validity and should exert full sway on the design and continuing evaluation of assessment and accountability systems" (Baker, Linn, Herman and Koretz, 2002).

The Problem of Multiple Measures

In essence, multiple measures within NCLB have provoked both reliability and validity discussions that should force the psychometric community to re-examine the way we measure basic constructs, such as reading ability. As states grapple with how and what they will include as multiple measures, perhaps a thought experiment might be helpful.

Consider for a moment how an assessment system with multiple measures might be designed if NCLB was focused on physical (*health*) outcomes as opposed to cognitive constructs. For example, if NCLB had been enacted to eradicate obesity in the K-12 population, our outcome measures would focus on weight (*pounds, grams*). Since weight and height are positively correlated, we would have to control for height. Perhaps, at the end of this approach, there would be three cut points and four groups: ectomorph, mesomorph, endomorph and below endomorph.

In the reading world, one can think of these three cuts as advanced, proficient and basic readers. Unlike the reading example, however, the cut points for physical assessment would consist of uniform definitions and exchangeable scales upon which all could agree. The physical assessment measures would still be left with the reliability concerns and thus, still in need of multiple measures. Nonetheless, because of the uniform metrics (pounds and grams, inches and meters) the assessment framework would be uncomplicated and the comparisons across states would be straightforward.

Unfortunately, many of the cognitive constructs that are of most interest to educators and policy makers are on nonexchangeable, proprietary scales. Thus, we cannot move easily from one measure to another like we can with constructs in the hard sciences. For example, because of measurement unification in temperature, height and weight, conversions from Celsius to Fahrenheit, inches to meters and pounds to grams are actively used. These scales are exchangeable and data collected from different instruments can be placed on a common scale.

The serendipitous benefit of the high-stakes consequences of NCLB is that it will expose one of the most profound limitations of measurement in the social sciences: The lack of unification of metrics (universal and standard scales). Without universal, exchangeable scales in the social sciences, our assessment systems across states may employ the same labels (advanced, proficient, basic and below basic), but may vary

dramatically in the achievement implied by these labels. An NWEA study proposes that, ÒStates have set proficiency levels using different definitions of Ôproficiency.Õ These standards are now being pressed into service as proficiency indicators under the No Child Left Behind. It is not surprising that the proficiency levels differ, but the degree to which they differ and the potential for misinterpretation are surprising.Ó A third-grade student labeled ÒproficientÓ in State A may differ dramatically from a third-grade student in State B as demonstrated in the following table (for a more detailed description of this conundrum, see Education Week, May 2003).

Table 1. Cut scores representing "proficient" or "meets standards" level of performance on 14 state assessments

Reading

Grade 3			Grade 4			Grade 5			Grade 6			Grade 7			Grade 8			Grade 9			Grade 10		
State	Cut Score	%ile	State	Cut Score	%ile	State	Cut Score	%ile	State	Cut Score	%ile	State	Cut Score	%ile	State	Cut Score	%ile	State	Cut Score	%ile	State	Cut Score	%ile
SC	205	67	WY	214	73	SC	220	73	SC	221	63	SC	227	70	WY	232	74	MT	224	43	OR	236	77
CA	200	51	SC	213	70	CA	214	54	CA	216	46	WA	226	67	SC	230	68	IA	224	43	WA	227	51
MN	193	35	WA	207	53	AZ	210	45	MT	211	35	CA	221	50	OR	227	58	ID	221	37	ID	224	44
OR	193	35	CA	205	46	OR	209	42	ID	211	35	MT	218	43	CA	226	54	CO	204	9	MT	224	44
ID	193	35	ID	200	34	IL	207	37	IN	210	32	IA	216	37	AZ	224	49				IA	223	42
MT	193	35	MT	196	26	MT	206	35	IA	209	30	ID	215	35	IN	219	35				CO	209	15
IL	193	35	IA	196	26	ID	206	35	TX	208	28	TX	210	24	MT	219	35				CA	208	14
IN	192	32	CO	191	18	IA	205	32	CO	197	11	CO	206	18	IA	219	35						
IA	191	31				MN	204	30							ID	218	32						
AZ	190	29				TX	204	30							IL	218	32						
TX	179	13				CO	197	18							MN	218	32						
CO	179	13													CO	206	12						

Mathematics

Grade 3			Grade 4			Grade 5			Grade 6			Grade 7			Grade 8			Grade 9			Grade 10		
State	Cut Score	%ile	State	Cut Score	%ile	State	Cut Score	%ile	State	Cut Score	%ile	State	Cut Score	%ile	State	Cut Score	%ile	State	Cut Score	%ile	State	Cut Score	%ile
SC	208	75	WY	221	83	SC	227	76	SC	235	78	SC	242	78	WY	257	89	MT	242	47	WA	257	73
CA	204	60	WA	218	76	CA	225	70	CA	230	67	WA	242	78	SC	251	80	IA	241	44	MT	247	40
IN	201	50	SC	217	74	AZ	220	59	IN	221	47	CA	238	70	AZ	248	75	ID	240	42	IA	247	40
OR	199	46	CA	212	59	OR	215	46	ID	219	42	ID	225	44	CA	240	59	CO	235	32	OR	245	33
AZ	199	46	ID	205	39	ID	213	41	IA	218	40	MT	224	42	OR	235	50				ID	242	25
MN	198	42	IA	205	39	MT	212	38	MT	218	40	IA	222	38	ID	233	46				CO	233	14
MT	197	39	MT	205	39	IA	212	38	CO	207	19	TX	221	35	MN	231	42				CA	232	13
IA	197	39				MN	210	33				CO	216	26	IN	231	42						
ID	196	36				IL	210	33							IL	230	40						
IL	193	29				TX	209	31							MT	228	36						
					CO	201	15							IA	228	36							
														CO	225	31							

¥ Indiana tests students in the fall. Their cut scores were adjusted to reflect equivalent spring performance.

¥ Colorado uses the partially proficient level of performance for NCLB reporting. To maintain consistency, NWEA reports the level each state uses for NCLB reporting here.

¥ The Texas estimate is based on the level for proficient performance that will be implemented in 2005.

(Source: NWEA report, 2003)

Consequently, the real reason that the multiple measures requirement is on such a slippery slope is that our instruments do not have exchangeable scales. Without standard objective scales, like those employed in the hard sciences, educators will be left with less-than-satisfactory methods and very confusing, complicated schemes for reporting such data.

Looking back to the late 1800s, one can find a direct analogue to the reporting dilemmas and confusion confronting the state assessment system. Before the introduction of the railroad system, our country literally had thousands of time zones, as each community would set their clocks to noon when the sun reached its zenith. This meant that every community was on their "local" system. Two neighboring communities might differ significantly, and traveling from one town to another meant that a person had to reset his watch upon arrival. With the introduction of the railroad system it was no longer feasible or practical to have all these "localized" time zones, and movement was begun by a Canadian engineer, Sir Sanford Fleming, to unify (standardize) the measurement of time (Blaise, 2000).

Just as the unification of time was borne out of a Canadian engineer's frustration in trying to figure out what time to pick up his nephew at the railroad station, the complexity and confusion of our state accountability systems may serve as the impetus to agree to standard, universal scales for the constructs of reading and math. A search of Buros Mental Measurement Yearbook (BMMY) for an instrument to measure reading ability or mathematics, yields hundreds of choices that are each on a unique, proprietary scale that is nonexchangeable across instruments.

There are countless constructs — such as temperature, time and weight — that looked like "reading ability" in terms of the number of ways to measure them in their early days of inception. Unification of these constructs was driven by two forces: First, as the underlying scientific theories were developed, there was implicit recognition that the underlying "scales" were important, not the plethora of instruments per se; and secondly, applications forced unification.

The Lexile Framework in Context

Today, at least when it comes to the measurement of reading, our theoretical understanding of the construct is sufficient for unification of scales, and perhaps the application that will accelerate unification is the legislation of NCLB.

A promising candidate in the unification of measurement of reading is The Lexile Framework for Reading. The Lexile Framework is an approach that makes it possible to place readers and text on the same scale (the Lexile scale). The Lexile Framework systemizes two common measurement assumptions:

- Text can be ordered as to difficulty (see Chall, 1996, for a thorough review of readability and the Lexile Framework)
- Readers can be ordered as to reading ability

By placing readers and text on the same scale (the Lexile scale), the difference between a reader's Lexile measure and a text's Lexile measure can be used to forecast the comprehension that the reader will enjoy with the text. One of the realities in U.S. K-12 education that tends to be neglected is student mobility. Every year, a great number of students move from one state to another. To the degree that all states are using tests that have been linked to the Lexile Framework, the students' test scores can travel with them.

With this continuity of the measurement system, schools will not lose important test data. Using the Lexile Framework, each state can establish its own proficiency level benchmarks, but by using the same scale, improvement can be viewed in very concrete terms. States could define their AYP for reading in terms of the Lexile scale and use the Lexile measures from the test results to document the growth.

Currently, every major test publishing company has linked their norm-referenced reading tests to the Lexile scale. Some examples include:

Harcourt Assessment

- Stanford Achievement Tests, Ninth and Tenth Editions (SAT-9 and SAT-10)
- Metropolitan Achievement Test, Eighth Edition (MAT-8)
- Stanford Diagnostic Reading Test, Fourth Edition (SDRT-4)

CTB/McGraw-Hill

- TerraNova Assessment Series (CTBS/5 and CAT/6)

The Riverside Publishing Company

- Gates MacGinitie Reading Tests, Fourth Edition (GMRT-4)
- The Iowa Tests (ITBS and ITED)

Northwest Evaluation Association (NWEA)

- NWEA Achievement Level Tests (print and electronic versions)
- Measures of Academic Progress (MAP)

In addition, Scholastic Reading Inventory, or SRI (Scholastic Inc., 1999), is a standardized assessment designed to measure how well students read literature and expository texts of varying difficulties. SRI began as a targeted-level pencil-and-paper test, but is now available in a computer-adaptive test format.

All of these instruments are able to report out Lexile measures for every student. Because the Lexile scale is a common, supplemental scale that has been linked to

the underlying scale of each instrument, we have moved closer to the concept of objectivity in measurement, and hence a unification of a construct. Just as when we measure temperature using a thermometer, we assume that the measure we obtain is not dependent on which thermometer we used. Likewise, in the measurement of reading ability, we assume that the measurement is not dependent on which assessment we used (e.g., the SAT-10, TerraNova, GMRT, etc.). This attribute, termed "general objectivity," is what has historically distinguished measurement in the physical sciences from that in the behavioral sciences.

Many states use tests that are already linked to the Lexile Framework. Any time a student takes one of these tests, he or she can receive a Lexile measure.

Having state assessment results also reported on the Lexile scale also enables parents, teachers and students to act on the information. With a Lexile measure, parents can actively support and encourage reading by helping their children select appropriately targeted books (tens of thousands of titles are available at www.Lexile.com). Lexile measures help teachers to differentiate instruction and select textbooks, classroom materials and periodicals that have been measured on the Lexile scale. Since tens of thousands of trade books, thousands of textbooks, and tens of millions of articles have been measured, the annual state test data can now be linked to the classroom text resources.

Students can also benefit from knowing their Lexile measure. Depending on the age of the student, the Lexile Framework can help older students select appropriately targeted research materials for projects. For younger children, a Lexile measure helps to ensure a positive reading experience. Targeted readers report confidence, control of the text, and comprehension and enjoyment of the reading material.

Another benefit is in describing "proficiency levels" with real-life text. What does proficiency at the fourth-grade level mean compared to eighth grade? The Lexile Framework provides a way for teachers, school districts and states to describe proficiency levels in terms of actual text that can be read and comprehended. Using the same label ("proficient") across all grades fails to communicate efficiently with parents. For example, fourth-grade proficiency could be described in terms of the types of text that a reader can comprehend. Concretely, this information could be presented with well-known titles at the different grade levels.

A final benefit is that the Lexile Framework permits school administrators to build longitudinal growth profiles on each student. Since the Lexile scale is a common supplemental metric that cuts across multiple instruments, these growth profiles can be built from multiple data points over many years. For example, if a district is using the SAT-9 as their norm-referenced test, a statewide test that has been linked to the Lexile Framework (e.g., North Carolina) and any interim assessments that reports Lexile scores (e.g., SRI), then there could be three data points for every year on the student's profile.

Conclusion

As our various sanctioning and professional bodies in the behavioral sciences have rightfully pointed out, we should not make high-stakes decisions from a single administration of a test. This standard has resulted in the necessity of multiple measures. Unless, however, there is general objectivity of measurement of the underlying constructs (i.e., reading, mathematics and science), we are still left with subjectivity and sliding state standards where "proficiency" in one state means something entirely different in another state.

About the Author

Malbert S. Smith III, Ph.D., is the president of MetaMetrics, Inc. and co-developer of The Lexile Framework for Reading. In 1984, Smith and A. Jackson Stenner, Ph.D., co-founded MetaMetrics. Together, they continue to lead the company as Lexile measures become the global standard for matching reader ability to text. Smith earned his doctorate of philosophy in educational psychology from The University of North Carolina at Chapel Hill, and his undergraduate degree in psychology from Duke University. He has published and presented numerous papers in the field of educational assessment and measurement, and teaches graduate seminars at Duke University and 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

Baker, Eva L., "Multiple Measures: Toward Tiered Systems," Educational Measurement: Issues and Practice, Vol. 22, No. 2, 2003.

Baker, E. L., R. L. Linn, J. L. Herman, D. Koretz, "From the Directors: Standards for Educational Accountability Systems," The CRESST LINE, Winter 2002.

Blaise, Clark "Time Lord: Sir Sanford Fleming and the Creation of Standard Time," Pantheon Books: New York, Summer 2000.

Kingsbury, G. Gage, Allan Olson, John Cronin, Carl Hauser, Ron Houser, "The State of State Standards: Research Investigating Proficiency Levels in Fourteen States," Northwest Evaluation Association, 2003.

Koretz, Daniel, "Using Multiple Measures to Address Perverse Incentives and Score Inflation," Educational Measurement: Issues and Practice, Vol. 22, No. 2, Summer 2003.

Smith, Malbert III, Ph.D., "Multiple Measures within NCLB: The Need for Objective Measurement," MetaMetrics, Inc., July 2003.

"2001 'No Child Left Behind' Act," The Lexile Times, Vol. 2, Issue 1, August 2002.

Lexile: Matching readers to text

LEXILE®

www.Lexile.com
1.888.LEXILES