Professional Paper

Accuracy Matters:

Reducing Measurement Error by Targeted HMH *Reading Inventory* Testing

Reading Inventory™

Overview

Traditional standardized grade-level reading tests are designed to measure grade-level standards. The scores' accuracy of these tests is not equal across all levels of ability. The further away from grade-level the student performs, the greater the degree of inaccuracy in the student's score. As a result, test scores of low- and high-ability students might not be accurate enough to be used instructionally or to monitor instructional growth from one year to the next.

The Lexile® score has the unique advantage of bridging assessment and instruction by reporting the complexity of instructional reading materials—such as textbooks, trade publications, and journal and magazine articles—and student reading ability on the same scale. A score from any standardized reading test can be associated with a given Lexile score.

In the case of a test linked to the Lexile Framework, such as HMH *Reading Inventory*, the accuracy of each score reported by the test affects the accuracy of the match between reader and text for the purpose of managing instruction. The accuracy of a test score can be compromised by the amount of Standard Error of Measurement (SEM) inherent in the test that produces the Lexile score. In the case of *Reading Inventory*, the accuracy of the test score can be substantially increased if the test is targeted by prior reading ability and grade level instead of grade level only.

Table of Contents

Introduction 1
Standard Error of Measurement 2
Grade-Level Tests 3
Computer-Adaptive Tests 4
Different Tests Produce Different Standard Errors of Measurement 5
FCAT's Standard Error of Measurement Graphs 6
Matching Students to Text for Managing Comprehension 10
Interpreting Individual Student Performance 12
Conclusion 14
References 15

Introduction

School districts and teachers are often frustrated when they try to use the results of a standardized grade-level test instructionally. The traditional scale scores produced by standardized grade-level tests have limited meaning for classroom instruction. Because these tests are designed to measure grade-level standards, they are valid measures of a standard, but the scores' accuracy is not equal across all levels of ability. The further away from grade level the student performs, the greater the degree of inaccuracy in the student's score. As a result, test scores of low- and high-ability students might not be accurate enough to be used instructionally or to monitor instructional growth from one year to the next.

The Lexile score, however, is a metric with powerful implications for classroom instruction in reading. No other measure bridges assessment and instruction by reporting the complexity of instructional reading materials—such as textbooks, trade publications, and journal and magazine articles—and student reading ability on the same scale. Theoretically, when students read text targeted to the same Lexile as their reading ability, they will demonstrate a comprehension rate of 75 percent or greater. Studies have shown that when students read text within the same difficulty range as their reading ability, they will comprehend what they read with 70- to 85-percent accuracy, and their reading comprehension ability can grow without frustration (Schnick & Knickelbine, 2000).

MetaMetrics, developers of the Lexile Framework for Reading, reports that "all standardized reading tests can report a Lexile score" (MetaMetrics, 2006). This means that a score from a standardized reading test can be associated with a given Lexile score. For example, a reading score from the Florida Comprehensive Assessment Test (FCAT) or SAT 10 can be tied to an equivalent Lexile score. Although every test can report a Lexile score, each type of test measures with greatest accuracy the constructs for which it was designed. The type of test and the constructs it measures affect the accuracy (in the form of the test's SEM) of each score reported by the test. In the case of a test linked to the Lexile Framework, such as *Reading Inventory*, the accuracy of the score affects the accuracy of the match between reader and text for the purpose of managing comprehension. As with other measurements, the accuracy of the match between text complexity and student reading ability reported as a Lexile score can be compromised by the amount of SEM inherent in the test that produces the Lexile score.

This paper illustrates how the SEM of a test linked to the Lexile Framework could influence a teacher's ability to manage student comprehension. It compares the SEM produced by three tests:

1) A standardized criterion-referenced grade-level state assessment
2) *Reading Inventory* when targeted by grade level only
3) *Reading Inventory* when targeted by prior reading ability and grade level

After establishing the SEM for each test this paper will discuss strategies for:

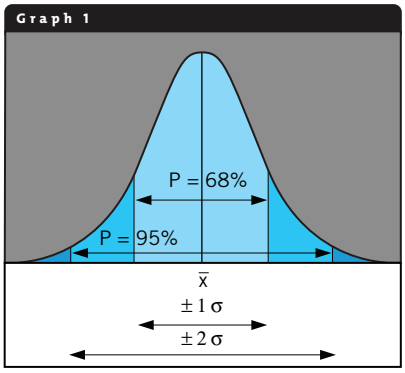
1) Interpreting fluctuation of Lexile scores generated from multiple administrations of *Reading Inventory* with little time between test administrations
2) Interpreting seemingly different scores from standardized reading comprehension tests linked to the Lexile Framework for Reading

Finally, the paper will examine how the SEM for each type of test score affects a teacher's ability to manage student comprehension by matching students to texts. The results of this analysis suggest that Lexile scores produced by computer-adaptive tests like *Reading Inventory* that are targeted by prior student performance are most accurate across all ability levels and most effective for detecting growth in student reading ability when administered three to four times a year.

Standard Error of Measurement

A test's accuracy is estimated by a number called the standard error of measurement. The SEM indicates to what degree a student's test score reflects his or her "true ability." Every test contains a unique SEM. Understanding a student's test score in light of a test's SEM tells us how much confidence we can place in the test's ability to measure the student's "true ability" to comprehend what he or she reads.

Evidence of SEM may be observed in the fluctuations of a student's test score. Test scores fluctuate for a variety of reasons. Although we will not discuss all the reasons that scores fluctuate in this paper, a full account of them can be found in *Why Scores Change* (Williamson, 2006). To illustrate SEM, assume that a student was administered *Reading Inventory* on Monday and scored 600L and took the test again on Tuesday and scored 557L (note: all Lexile scores are reported with an "L" suffix). If the student experienced exactly the same testing conditions and exhibited the same testing behaviors—content sampling, incorrect answers, misinterpreted instructions, guesses, and personal state of mind (for example, fatigue, nervousness, hunger, discomfort, etc.)—how can we explain the apparent difference in the student's reading ability from one day to the next? Under normal conditions, one day is not enough time between administrations for the student to gain or lose reading ability. To make sense of the change in student performance, the SEM of the test needs to be considered. The SEM of *Reading Inventory* is 93L (when targeted by grade level alone) and 56L (when targeted by grade level and prior ability). The difference in the student's Monday and Tuesday test scores (43L) is within one SEM (93L or 56L) for either way of targeting *Reading Inventory*. Therefore, the change in the student's test score could be attributed to the test's SEM. In this paper, we will operationally define the SEM as the number that tells us to what degree fluctuations in a test score reflect randomness (instead of a change in a student's "true ability").



P represents probability

The SEM helps us understand how accurately a test is able to measure a student's ability. For example, as shown in Graph 1, if a student took a test 100 times, 68 times out of 100 the student's score would fluctuate +/– one standard error of measurement due to random chance alone. In this case the SEM is the same as the standard deviation for the distribution of the student's scores. Changes in a student's score that fall within this range cannot be attributed to changes in the student's "true ability." Ninety-five times out of 100, the student's score will fluctuate +/– two standard errors of measurement due to chance alone. To say that a student has experienced a change in ability that is not attributable to chance, the student will need to exceed his original score by the number of points that represent two standard errors of measurement. The smaller a test's SEM, the more accurate the student's score will be, making it easier to detect real growth in ability, not merely change due to random chance.

The size of an assessment's SEM is affected by the test's purpose. As noted above, each type of test measures with greatest accuracy the purpose for which it was designed. The purposes of a high-stakes grade-level test and *Reading Inventory*, a computer-adaptive test, are very different. The amount of SEM in each type of score across the range of possible scores reflects the purposes for which each test was designed to be used.

Grade-Level Tests

Traditionally, grade-level tests are delivered in paper-and-pencil format. They are designed to describe student performance in relation to either:

1) The performance of a national sample of grade-level peers (i.e., norm-referenced), or
2) Grade-level standards (i.e., criterion-referenced)

Norm-referenced and criterion-referenced tests traditionally deliver the same test items to every student, regardless of the student's reading ability. Grade-level test items include a majority of items appropriate for the grade level being tested and some items appropriate for one grade level above and one grade level below the grade level being tested.

On norm-referenced tests, student performance is distributed on a normal curve around an average score. Student scores are then rank-ordered and a national percentile rank is assigned to each score. Students who complete subsequent administrations of the norm-referenced test get a scale score and a percentile score that describe their performance in relation to the norm group—the initial group of students whose demographic characteristics reflect those of students nationwide at the same grade level.

Unlike norm-referenced tests, criterion-referenced state assessments are designed to measure a grade-level standard that is applied to all students who attend school in a particular state. A set of test items is selected to sample the curriculum benchmarks of a given grade level. Some criterion-referenced tests are designed to test minimum competencies. Most students who take this type of test are able to pass; therefore, the scores produced by these tests are not evenly distributed around an average score. Other criterion-referenced tests, such as FCAT, measure high standards. Unlike a minimum-competencies test, student scores are distributed across all possible scores.

State standards vary nationwide. State proficiency standards are established through a process that includes various stakeholders such as parents, teachers, and various community members. Even so, all state criterion-referenced tests share the purpose of measuring the degree to which students demonstrate benchmark skills using test passages that reflect content that students are expected to be exposed to during a school year.

Computer-Adaptive Tests

Computer-adaptive tests are delivered by a computer software program. Each time a student takes a test, the student is presented with a unique set of items whose difficulty level is targeted to match the student's ability level. With proper targeting, a student has a 50-percent chance of answering each question correctly. As a student answers a question correctly or incorrectly, the software adjusts the difficulty of the next test item to maintain the student's chances of selecting a correct answer.

Reading Inventory is a computer-adaptive test. The purpose of *Reading Inventory* is to describe what level of text complexity a student can read and comprehend with 75-percent accuracy, regardless of the student's initial ability level. *Reading Inventory* contains about 6,000 items that can detect and measure a student's reading comprehension between 100 and 1500+ Lexiles. The resulting Lexile score is accurate across all ranges of abilities, not just the range of abilities common to students at a particular grade level. The score is designed to be used instructionally, to guide students to text that they can read and comprehend with an accuracy rate of 75 percent or greater. The range of Lexile scores reported is operational, meaning the scores produced by *Reading Inventory* match the full range of reading materials common to students in first grade through graduate school.

Grade-level tests that report Lexile scores are linked to the Lexile Framework through formal linking studies. The Lexile score resulting from linking studies contains the same degree of random error across the range of possible scores as the parent test on which the Lexile score is based. Because *Reading Inventory* administers a unique test that matches each test taker's ability, each test has a unique SEM. The next section of this paper will examine the SEM of three tests: the Florida Comprehensive Assessment Test—Sunshine State Standards Reading for Grades 3–10, *Reading Inventory* targeted by grade level only, and *Reading Inventory* targeted by grade level and prior reading ability.

Different Tests Produce Different Standard Errors of Measurement

Each test and each grade-level test within a particular instrument yields its own SEM. To discern "true ability," an understanding of tests' SEM is essential.

To illustrate how a test's purpose affects its SEM, graphs 2–11 show how the SEM for the following three types of test varies:

1) FCAT by grade level
2) *Reading Inventory* when targeted by grade level only
3) *Reading Inventory* when targeted by grade level and prior ability (i.e., far above, above, at, below, or far below grade level)

Graphs 2–9 showing the FCAT's SEM by grade level and across scores illustrate the amount of inaccuracy in a grade-level test whose purpose is to measure a standard. The FCAT is not linked to the Lexile Framework by a formal linking study. The FCAT's SEM, if it were linked formally to the Lexile Framework through the concurrent administration of a Lexile Theory test, would be inherent in each Lexile score to the same degree across all possible scores. That is, the resulting Lexile score's SEM would be greatest for the lowest and highest scores and least for scores that reflect each grade-level standard, so the scores for students on grade level would be much more accurate. The standard error of multiple tests is graphed to illustrate:

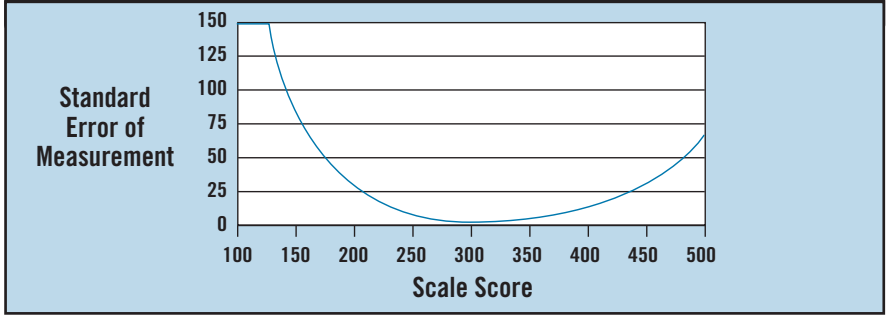
1) The difficulty of attributing fluctuations in student scores to growth in ability, even when they are reported on the same Lexile scale, without considering the SEM
2) The Lexile scores reported by a grade-level test may not be accurate enough to manage student comprehension with 75-percent accuracy when matching students to texts

The graphs show the SEMs of the FCAT SSS Reading Test as reported in the FCAT Technical Manual (Florida DOE, 2002). The graphs demonstrate that:

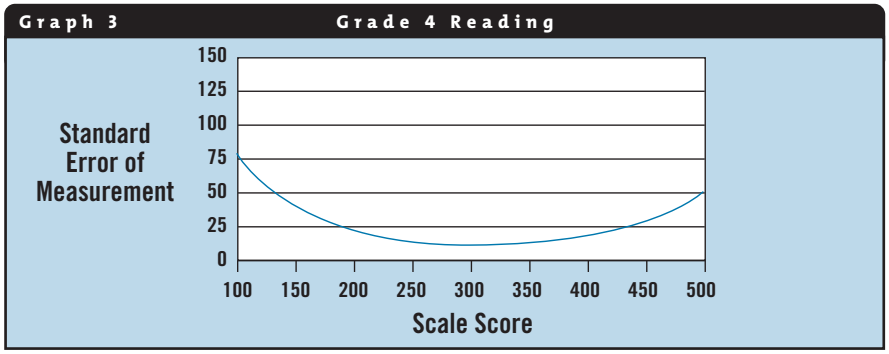
1) The SEM is not the same across all grade-level tests
2) The minimum standard error on each grade-level test falls within a specific range of scores

FCAT's Standard Error of Measurement Graphs

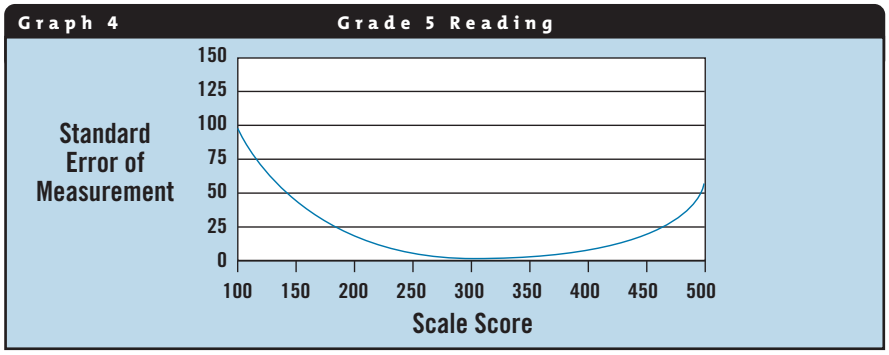
Graph 2 Grade 3 Reading



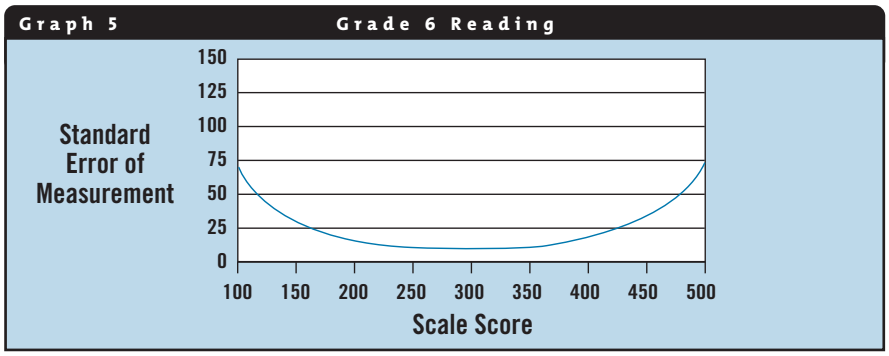
Graph 3 Grade 4 Reading



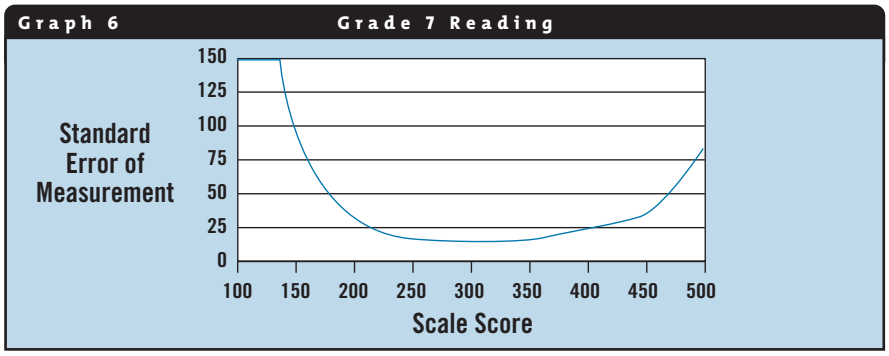
Graph 4 Grade 5 Reading



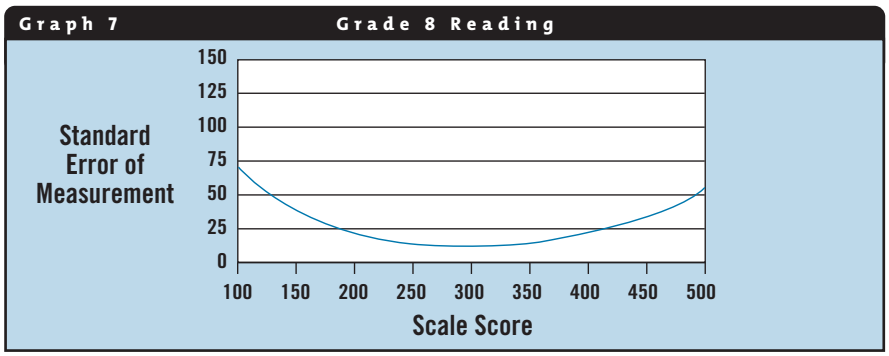
Graph 5 Grade 6 Reading



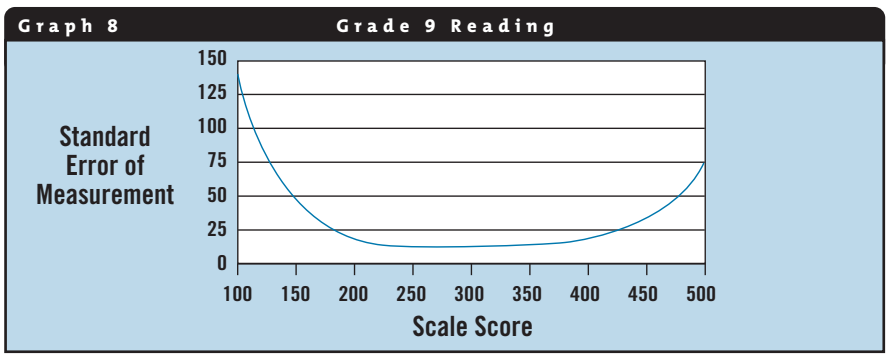
Graph 6 Grade 7 Reading



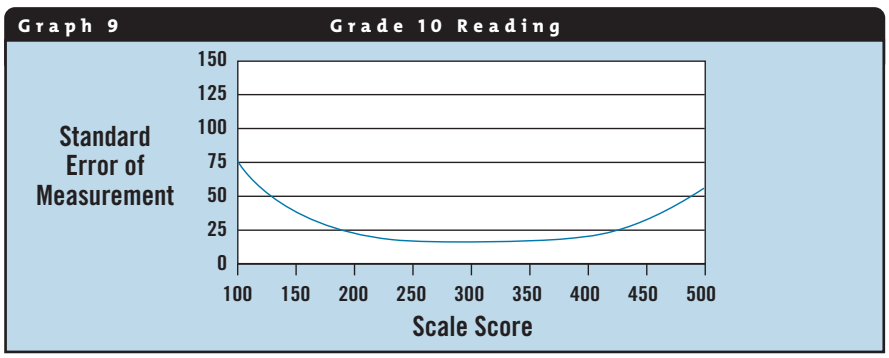
Graph 7 Grade 8 Reading



Graph 8 Grade 9 Reading



Graph 9 Grade 10 Reading



The importance of targeting by grade level and by known prior reading ability in determining students' "true ability" is key when accuracy of the score is the desired outcome. Table 1 shows the SEM of two different administrations of *Reading Inventory*. The second column shows the SEM of *Reading Inventory* scores reported when the test is targeted by grade level only. The third column shows the SEM of *Reading Inventory* scores reported when student grade level and prior reading ability are used to target the test for each test taker. In every case, the SEM decreases substantially going from the second column to the third column. This means that when prior reading ability is used in addition to grade-level in targeting the test, the accuracy of the scores increases because the SEM is much smaller.

Table 1
Average SEM on *Reading Inventory* by Extent of Prior Information

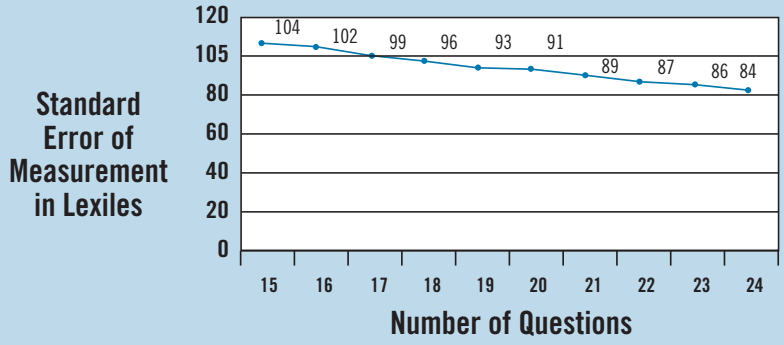
Number of Questions	SEM: Grade Level Known	SEM: Grade and Reading Level Known
15	104L	58L
16	102L	57L
17	99L	57L
18	96L	57L
19	93L	57L
20	91L	56L
21	89L	56L
22	87L	55L
23	86L	54L
24	84L	54L

The same information on Table 1 is presented in graphs 10 and 11. These graphs show the SEM of:

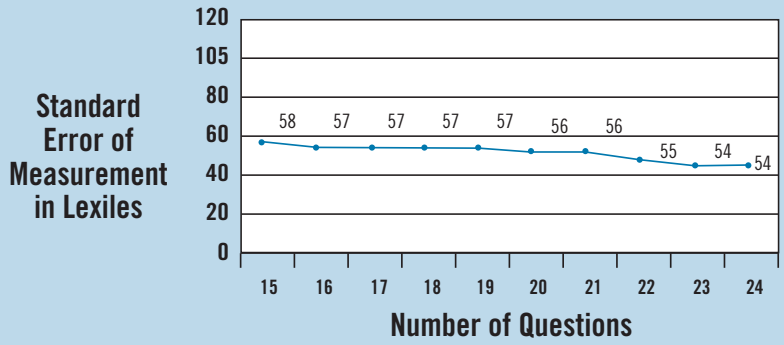
1) *Reading Inventory* when targeted by grade level only
2) *Reading Inventory* when targeted by grade level and prior ability as it relates to the student's grade level (far above, above, at, below, or far below grade level)

The graphs demonstrate that the SEM is consistent across all tests by the number of items, even though each test is unique. More importantly, note that the SEM of the Lexile score produced by *Reading Inventory* is cut in half simply by targeting the student's prior ability. If a student's prior reading ability is not specified before the first administration of *Reading Inventory*, the student will need to answer approximately 40 items before the test is targeted to the student's ability. In other words, three *Reading Inventory* tests will be needed to produce a Lexile score accurate enough to match a student to the text he or she comprehends with 75-percent accuracy.

Graph 10 *Reading Inventory* Targeted by Grade Level Only



Graph 11 *Reading Inventory* by Grade Level and Reading Level



Because *Reading Inventory* is an adaptive test, the SEM of the scores at the highest and lowest points of the range will be more accurate where grade-level standardized tests are less accurate.

Matching Students to Text for Managing Comprehension

Reading Inventory Educator's Guide discusses what students experience when they read targeted text. Targeted text has a difficulty level within a range of Lexile scores. *Reading Inventory Educator's Guide* recommends the range of reading materials span 100 Lexiles below to 50 Lexiles above the student's score. The comprehension range produced by reading targeted text spans 70-to 85-percent accuracy. After accounting for variables—such as student, content, and text—students who read within their targeted range are able to experience engagement with the text. It isneither too easy nor too difficult.

A test's SEM can affect whether a score can be used to manage a student's comprehension of text within the targeted range. For example, let's compare the accuracy of comprehension a student will experience when matched to text from an *Reading Inventory* test that has been targeted by prior reading ability with an *Reading Inventory* test that has been targeted by grade level only. The average SEM of *Reading Inventory* when targeted to match the student's prior reading ability is 56L. When *Reading Inventory* is targeted by grade only, the average SEM is 93L—almost double. Notice in Table 2 that the probability is 82 percent that a student's "true ability" is included within the range of text that matches his or her comprehension ability when *Reading Inventory* is targeted by prior ability. This means that a student's "true ability" overlaps the random error of the test by two standard errors of measurement. The test score produced by *Reading Inventory* when targeted by a student's prior ability is accurate enough to be used to manage the student's comprehension of text. The overlap between the SEM and the reading level of the text proves that student comprehension is being managed.

Table 2
Targeting *Reading Inventory* by Grade Level and Reading Ability

95% probability that student's "true ability" is included in the range of SRI scores that spans two SEM	288 2 SEM (-112)	344 1 SEM (-56)	**400** **SRI Score**	456 1 SEM (+56)	512 2 SEM (+112)
Range of targeted text reading levels	300	350	**400 SRI Score**	450	500
Accuracy of comprehension	85%	80%	**75%**	70%	65%
Probability that student's true score is within recommended range of text	14%	34%		34%	13%
Cumulative probability that student's true score is within recommended range of text	14%	48%		82%	95%

Notice in Table 3 that when students read text within the recommended range after completing *Reading Inventory* targeted to their grade level only, the text will match the student's "true ability" only 66 percent of the time. The standard error of a test targeted by grade level only is great enough that a student's comprehension will be managed little more than half the time a student reads text within his or her recommended range.

Table 3

Targeting *Reading Inventory* by Grade Level Only

95% probability that student's "true ability" is included in the range of SRI scores that spans two SEM	214 2 SEM (-186)	307 1 SEM (-93)	.5 SEM	**400 SRI Score**	447 .5 SEM	493 1 SEM (+93)	586 2 SEM (+186)
Range of targeted text reading levels		300		**400 SRI Score**	450		
Accuracy of comprehension		85%		**75%**	65%		
Probability that student's true score is within recommended range of text	14%	16%	18%		18%	16%	14%
Cumulative probability that student's true score is within recommended range of text	14%	30%	48%		**66%**		

At the National Lexile Conference, Knutson (2006) presented preliminary research findings that suggest the accuracy of the match between student scores and the range of recommended text correlates to gains in reading comprehension. Considering the SEM of grade-level tests, the mismatching of students to text illustrated in the table above will only be amplified for the lowest- and highest-scoring students.

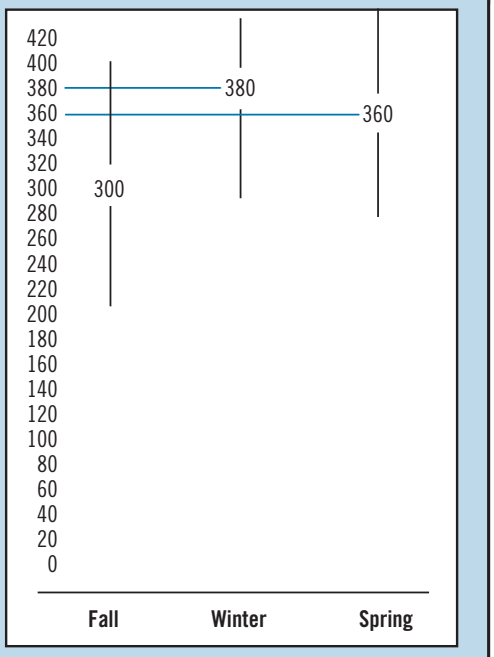
Individual Student Performance

Given the varying SEM of *Reading Inventory* when targeted and not targeted by prior reading ability, many educators ask how to interpret the fluctuations they see in student test scores. Because a student's test score will fluctuate by two SEM in 95 percent of 100 administrations due to random chance, student scores should be considered in relation to the test's SEM.

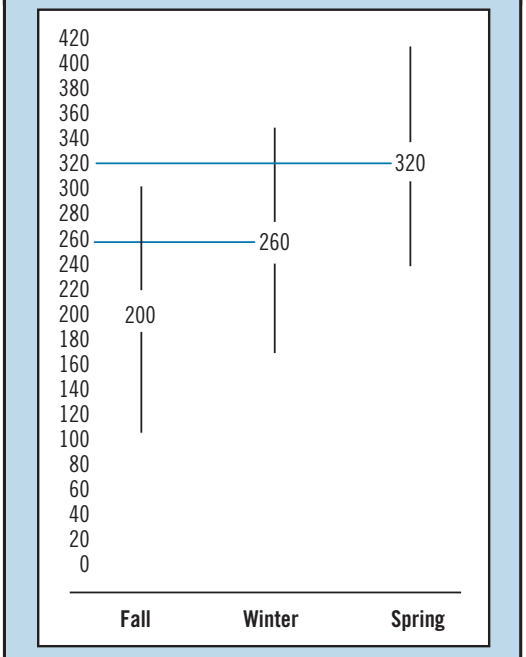
For example, suppose a student takes *Reading Inventory* targeted by his prior reading ability at the first administration in the fall and then completes two more administrations, one in the winter and one in the spring. The student's results are shown in Graph 12. In this case, the student's true reading ability has not changed, either positively or negatively. The vertical lines extending above and below the student's scores represent +/– two standard errors of measure. The horizontal lines extending from the student's winter and spring scores intersect the vertical line above the fall score; i.e., the winter and spring scores are still within the SEM of the fall score. The student's winter and spring scores are not significantly different from the fall score.

Graph 13 shows a student whose score has changed significantly from fall to spring. Notice the student's spring score of 320 exceeds the fall score of 200 by two standard errors of measure.

Graph 12



Graph 13

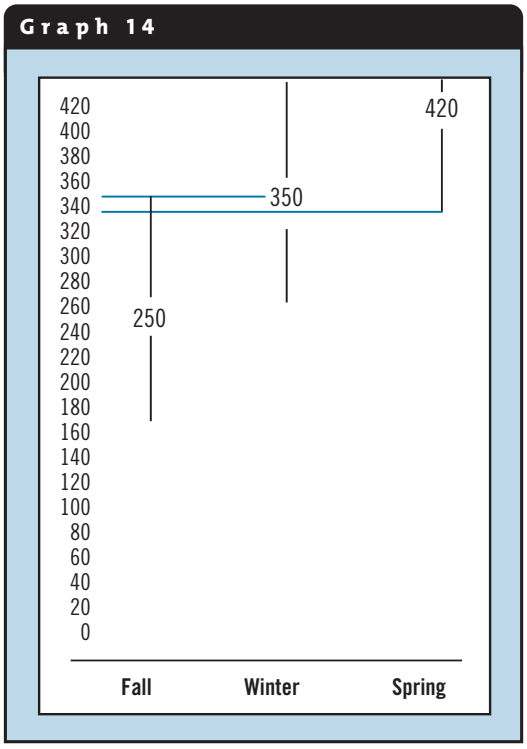


Graph 14 shows a range of scores that suggests the student's "true ability" has changed positively. The winter score of 350 exceeds the fall score of 250 by two standard errors of measure; however, the range of winter scores within −2 SEM still overlaps the range of fall scores. But the range of spring scores +/−2 SEM exceeds the range of fall scores +/−2 SEM. There is no overlap: the "true ability" of the student has changed between fall and spring.

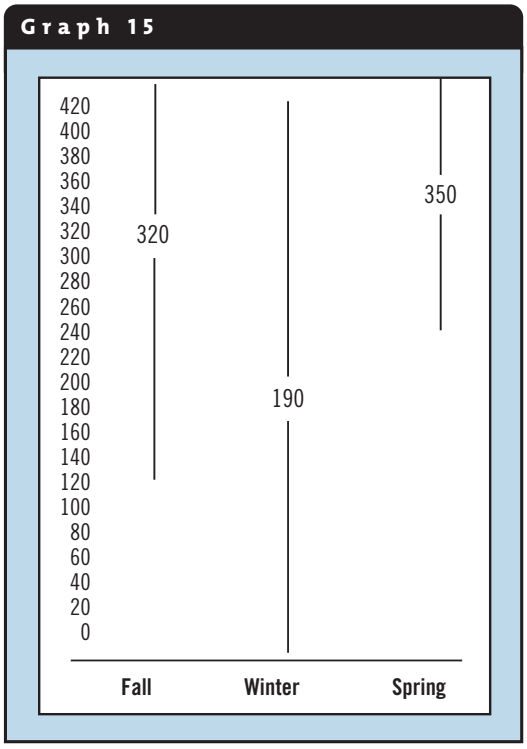
Graphs 12, 13, and 14 show scores in relation to their SEM when the test is targeted by the student's prior ability. Graph 15 shows how to interpret fluctuation in scores when the first test is not targeted by the student's prior ability. Notice that the SEM bands of the fall and winter administrations are double those of the spring administration. Initial scores that result from tests not targeted by student ability will become targeted after 40 items are completed, i.e., usually by the third test administration.

It is important not to test using *Reading Inventory* more than three to five times a year, as recommended by *Reading Inventory Educator's Guide*. Because there are about 6,000 test items available and the range of Lexiles measured extends from 100 to 1,500, with frequent testing—say, on a monthly basis—a student may see items duplicated across administrations. *Reading Inventory* is an assessment, not an intervention to improve reading. *Reading Inventory* should not be used as a means to practice test-taking, especially if the school district uses the results as one of several measures of a district or state grade-level standard.

Graph 14



Graph 15



Conclusion

Reading Inventory offers educators the opportunity to use reading comprehension scores in meaningful ways. When Lexile scores are aligned to state standards, teachers can monitor students' progress toward grade-level standards in terms of the complexity of text the students can comprehend. Teachers can support reading comprehension growth by matching students to text. They can use Lexile scores instructionally by creating lessons that use differentiated materials.

Since the students who are most likely to need monitoring are those reading below grade-level—the same students whose performance places them in the range of scores where grade-level standardized tests contain the greatest amount of inaccuracy—it is crucial to use a test that reports a score with the least amount of random measurement error, as is the case with *Reading Inventory*. When administering *Reading Inventory*, teachers need to be aware that the accuracy of the score can be substantially increased if the test is targeted by prior reading ability and grade level instead of grade level only.

Note: Prior to 2015, the HMH Reading Inventory was known as the Scholastic Reading Inventory (SRI).

References

Florida Department of Education. (2002). *Florida Comprehensive Assessment Test (FCAT) for Reading and Mathematics: Technical report for test administrations of FCAT 2002*. Tallahassee, FL: Florida Department of Education.

Knutson, K. (2006). *Because you can't wait until spring: Using the Reading Inventory to improve reading performance*. New York: Scholastic Inc.

MetaMetrics. *Frequently asked questions*. Retrieved September 12, 2006, from MetaMetrics Web site: http://www.lexile.com

Schnick, T., & Knickelbine, M. (2000). *The Lexile framework: An introduction for educators*. Durham, NC: MetaMetrics.

Scholastic. (2006). *Scholastic reading inventory interactive educator's guide*. Scholastic Inc.

Scholastic. (2001). *Scholastic reading inventory interactive educator's guide*. Scholastic Inc.

Williamson, G. (2006). *Why scores change.* Retrieved September 12, 2006, from MetaMetrics Web site: http://www.lexile.com

Professional Paper

Reading Inventory™

Connect with us:

Houghton Mifflin Harcourt™