Houghton Mifflin Harcourt

Interpreting Assessment Results

PROFESSIONAL PAPER

TABLE OF CONTENTS

INTRODUCTION

Every school's assessment program is designed to meet a variety of needs, including screening, placement, progress and growth monitoring, and accountability. Strong assessment programs use multiple sources of data to inform instructional planning. Strong programs also focus on ensuring that the data collected are used for their intended purposes.

Many schools use HMH *Reading Inventory* and HMH *Math Inventory* to make inferences about learning and instruction. These programs provide metrics to determine the extent to which students are performing to grade-level expectations. Educators can also use these measures to determine if expected growth of all students is met.

This paper provides guidance to educators as they review student results from the HMH *Reading Inventory* and HMH *Math Inventory* as part of their assessment program. There are times when results do not align to expectations for individual students. This paper interprets these situations.

EDUCATIONAL MEASUREMENTS

In educational assessment, we strive to measure attributes that are not directly observable: reading comprehension and mathematical understandings. We cannot actually see the act of reading or the process of mathematical thinking, so instead we measure a proxy of what we can observe. In *Reading Inventory*, we measure students' responses to questions about a passage of text. In *Math Inventory*, we measure students' responses as they solve math problems.

Measuring reading ability and math understanding is complex because the measures are indirect and involve human behavior. Our measurement instruments bring precision and reliability to the task, which by its nature can only get "so close" to the true measure of students' ability and growth.

General Expectations in Educational Measures

When we measure students, our general expectation is that their scores will improve over time. This is a reasonable assumption because students *do* grow over time in response to instruction, and our measurements of them are designed to *reflect* their progress.

In our everyday lives, simply by looking at our grade books or observing students in our classroom, we know that their performances vary. Score fluctuation is a standard part of assessment and should be expected.

From a statistical perspective, score fluctuations are called *errors in measurement*, which is defined as "the difference between a measured value of quantity and its true value."

Measurement error is not a "mistake." Measurement error does not necessarily need to be corrected. Variability is inherent in the measurement process. Every test yields an error of measurement.

UNDERSTANDING MEASUREMENT ERROR

The cause of measurement error is attributed to two sources:

- Systemic error (repeatable factors inherent in the measuring instrument)
- Random error (unintended factors for which we cannot repeatedly control)

Reducing Systemic Error

Systemic error refers to the limits to the testing instrument itself and is easier to control for than random error. Systemic error tends to be reproducible: it is a function of the test instrument and recurs consistently. Because systemic error can be reproduced, it can be controlled. Systemic error does not contribute to score fluctuations as much as random error does.

Most commercial assessments and state exams are subject to research studies to determine their reliability—the reproducible consistency of their measure. Tests with a high level of reliability have a lower level of systemic error and are more desirable instruments.

Third-party reviews of *Reading Inventory* and *Math Inventory* from the National Center on Response to Invention and the National Center on Intensive Intervention resulted in the highest ratings for reliability. These assessments demonstrate low systemic error, can identify the sources of error, and produce consistent measures.

Because systemic error can be studied and documented, it can be mitigated by knowledge of the instrument and program features. *Reading Inventory* and *Math Inventory* endeavor to reduce systemic error with these program features:

- **Targeting**: In *Reading Inventory* and *Math Inventory*, before the first test, teachers are asked to identify the general level of each student's proficiency. This practice, called targeting, identifies a starting point for the first question. A first question delivered closer to the students' ability will result in greater accuracy of the first test.
- **Save Test**: In *Reading Inventory* and *Math Inventory*, a test can be saved at any time. This allows teachers to increment testing over a number of days to compensate for test fatigue.
- **Locator Test**: In *Reading Inventory,* for students in Grade 7 or above who are below grade level, two or three more items at the beginning of the test are included to locate their true start point.
- **Skip Items:** In *Reading Inventory* and *Math Inventory*, students can skip up to three items if the context of the passage/question is unclear to them.

Reducing Random Error

Random error refers to error produced by normal human behavior. Random error can arise from:

- **Test Administration Practices**—timing, interruptions, conditions in test room, clarity of the test directions, attitude of the test administrator, and the perceived consequence of the scores

- **Student Motivation**—state of mind; alertness; feelings of fatigue, hunger, or illness; lack of interest or attentiveness; guessing; speed; and/or carelessness;
- **Other Factors**—misreading items or answers, misunderstanding the instructions, clerical errors, test "prepping," etc.

Error related to human behavior can be reduced with knowledge, awareness, management, and communication.

Test Administration Practices

To reduce error inadvertently introduced during test administration, school leaders can take a number of steps to prepare the environment by:

- Helping staff understand the purpose of the assessment
- Valuing the accuracy of scores rather than the achievement of "high" scores
- Creating flexible testing schedules to support retesting
- Ensuring that the testing rooms are clean, well heated or ventilated, well lit, and in quiet locations that are respectful of the students
- Ensuring that technology is properly functioning with adequate server capacity

The most important factor in receiving highly accurate results with *Reading Inventory* and *Math Inventory* is to empower teachers and proctors to manage the test environment actively.

For example, on any day there will be students who are unable to participate fully in the demands of school. Teachers should have the ability to excuse a student who is not well or is unable to make a genuine effort on testing day.

Similarly during testing, a teacher should be able to make the decision to excuse a student for attention or behavior issues. Since the goal of the assessment is to receive accurate information on students' abilities, and since *Reading Inventory* and *Math Inventory* scores are comparable over a number of weeks, a process should be in place to reschedule testing for the excused student.

Student Motivation

Student motivation is enhanced if the assessment is connected to their goals. Before any assessment, students should be familiarized with the purpose of the test and the instrument itself. Students should understand how their scores will help them reach their goals and how to track their own progress.

Detailed guidelines to enhancing student motivation are presented in each program's **Educator's Guide**. These resources can be accessed from the HMH Product Support site.

INTERPRETING RESULTS

For any test, there may be a set of scores that do not align with the teacher's expectations or with other evidence about students' performance. When results do not align with expectations, teachers can enact the following steps.

Reviewing Tests

1. **Speak with the student:** When interviews are conducted in a positive spirit, most students will admit that they were confused, demotivated, inattentive, or distracted during the assessment.

2. **Review the student's test:** A teacher should review the student's test looking for two indicators that suggest lack of attention and motivation. These are time-on-task and the student's response patterns. The average student takes 20 minutes to complete *Reading Inventory* and 30–40 minutes to complete *Math Inventory*. A time stamp is provided on each test instance; any test completed in a highly abbreviated amount of time should be reviewed. The teacher can also look for response patterns (all As or all Bs, etc.) as a proxy for inattentive responding.

3. **Ask about the test environment:** Ask the proctor if there were interruptions or distractions during testing that could have affected a student's results.

Responding to Student Behavior

When a teacher concludes that a student's behavior and/or the testing environment have affected the accuracy of the result, there are three possible next steps:

- **Let the test stand:** a good choice if retesting is impractical or if the student, with reasonable accommodation, has demonstrated the inability or inattention to take the assessment as intended
- **Retest the student after four-to-six weeks:** a good choice if results are in the general range of expected performance for the student at his or her grade level
- **Remove the test from the program's management system and retest the student:** a good choice if motivation, inattention, evidence of cheating, misadministration, or poor testing conditions were major factors affecting the test result.

Establishing Test Removal Procedures

For the protection of the teacher and student, HMH strongly suggests that school administrators create an approval process for test removal and retesting. Tests should not be removed because of:

- Variations in scores
- Gut feelings or intuition
- Trying for a better score

INTERPRETING RESULTS CONTINUED

If a motivated student is retested in *Reading Inventory* and *Math Inventory* and receives a similar result, there should be no further retesting. Continued retesting of students who are not producing desired scores should be discouraged. This practice contributes to test fatigue and over familiarity with the test instrument, and it displaces the student's time in instruction.

WHY SCORES VARY

On any instrument, individual scores can fluctuate from test to test. When we expect students to grow at a certain trajectory, what is the context to interpret scores that fluctuate for individuals from test to test?

Decreases From Test to Test

Both *Reading Inventory* and *Math Inventory* are formative assessments and, in aggregate, students' scores trend upward over the course of an academic year. For some students, however, scores decline from one test administration to the next.

A component of systemic error called Standard Error of Measurement (SEM) can help us interpret the variability of individual results. Because no single test administration can ever fully capture a student's ability growth, SEM is a metric developed to indicate an estimate of how repeated test administrations given at a close interval of time tend to be distributed around a student's "true score." SEM is a theoretical construct that creates an accuracy range.

The SEM in both assessments (*Reading Inventory* and *Math Inventory*) decreases with each test administration. If the grade range and approximate reading level are entered into the Student Achievement Manager, the SEM for the first administration of *Reading Inventory* is approximately 56L. To interpret a student's score of 700L, we should think that the range of the student's true ability will fall between 644L and 756L. Subsequent tests should be considered more accurate than previous tests, as the SEM falls with an increased number of items. In both assessments 40 items are generally needed to achieve a low stable SEM. For *Math Inventory*, the SEM is 63Q.

For more information about standard error of measurement, please refer to the Scholastic Professional Paper *Accuracy Matters*, by Kimberly Knutson, Ed.D., the *Reading Inventory Technical Guide* and the *Math Inventory Technical Guide*, both available at the HMH Product Support site.

Increases From Test to Test

Standard Error of Measurement applies to both sides of the score. In general terms, if a fourth grader receives a measure of 700L in the fall, and receives a measure of 715L in December, his teacher could infer that his testing demonstrated no significant growth.

However, if the second test shows a significant gain, the teacher should investigate to determine if the change is due to real growth in performance or random error. If the student is receiving reading intervention services designed to accelerate growth, the change in test performance, with other evidence, can be attributed to the positive effect of targeted instruction.

WHY SCORES VARY CONTINUED

Educators also need to interpret growth in a developmental context: younger and less proficient readers grow more than older and more proficient readers.

The *Growth Expectations Guide* is available at hmhco.com/readinginventory. This report discusses an empirical study of individual student growth by grade level and Lexile® level focusing on what amount of growth should be expected for students of varying ability. A similar study is underway for HMH *Math Inventory*.

Finally, as formative assessments, *Reading Inventory* and *Math Inventory* measure growth. It is important that students are not overtested with instruments. For the average student, growth measures will not show change if they are retested in less than six weeks. While 30 days lapse can be used in intensive interventions, nine weeks between testing is the most optimal range.

CONCLUSION

HMH assessments offer educators the opportunity to screen students, monitor progress, and measure student growth in math and reading. When interpreting the measures from these assessments, educators should understand that variations in scores may be due to multiple factors including test administration practices, messaging, student motivation, and the test's standard error of measurement. For this reason, it is strongly recommended that an approval process for test removal and retesting is followed in order to ensure that the methods for review of questionable scores follow a coherent process throughout the district.

REFERENCES

Dodge, Y. (2003). *The oxford dictionary of statistical terms*. Oxford University Press.

Knutson, K. A. (2006). *Accuracy matters*. Scholastic Inc.

Knutson, K. A., Scholastic, and MetaMetrics (2011). *Growth expectations guide*. Scholastic Inc.

Scholastic. (2010). *Scholastic math inventory educator's guide*. Scholastic Inc.

Scholastic. (2010). *Scholastic reading inventory educator's guide*. Scholastic Inc.

Scholastic. (2007). *Scholastic reading inventory technical guide*. Scholastic Inc.

Scholastic. (2012). *Scholastic reading inventory technical guide*. Scholastic Inc.

Williamson, G. (2004). *Why do scores change?* Metametrics, Inc.

PROFESSIONAL PAPER

Connect with us:

Houghton
Mifflin
Harcourt