

White Paper: Assessment Validity and Reliability

Purpose of this paper

Learning Today realizes that educators are increasingly being asked to evaluate and justify actions that they undertake in the process of educating students. This increase in accountability has added new demands on educators as they seek to evaluate and scrutinize outside goods and services being offered by many educational publishers or service providers.

The purpose of this paper is to clarify the many terms that surround this evaluation process and to shed additional light on where and how Learning Today's own products fit into the education process as a tool for schools and teachers. In this white paper, we will provide the following information:

- Description of the relationship between Learning Today and Let's Go Learn.
- Definition of the terms that surround educators' evaluation process.
- Description of how Let's Go Learn's web-based reading assessment meets the requirements of today's educational accountability requirements.

Relationship between Learning Today & Let's Go Learn

Learning Today (LT) is a PreK-5th Grade web-based curriculum developer, publisher and distributor specializing in Rapid Curriculum Development (RCDTM) & Deployment Methodologies since 2000. Let's Go Learn (LGL) is a K-12 web-based reading assessment developer and publisher since 2001. Both companies are pioneers in the development of web-based educational technology and are experts in the field.

LT originally released its web-based curriculum to three Miami-Dade County Public Schools on a pilot basis in September 2002. The original product was completely teacher-directed and was supported by state-of-the-art web-based applications. At that time, LT believed that reading assessment software was not a necessary component of our system due to the amount of assessment software, nationally normed commercial pen and paper assessments, and other text-based assessment tools already in use by the district. However, it was found that the integration of a diagnostic assessment tool with our curriculum would allow educators to target individualized needs and implement intervention strategies faster, resulting in improved academic gains. In addition to individual academic gains, this integration could reduce curriculum specialists' and teachers' workload by increasing their efficiency in the diagnostic/prescriptive phases of the initial and ongoing assessment periods.

Based on the need to target individualized needs and intervention strategies, increase administrative efficiency, and enhance teacher productivity, LT became an authorized distributor of Let's Go Learn's (LGL) Diagnostic Reading Assessment. LT chose LGL because of their superior native-web technology, and their in-depth research regarding the validity of their reading assessment product. Both development teams worked on integrating the two systems into one, seamless diagnostic/prescriptive reading solution, thereby expanding LT's products and services.

LT's new diagnostic/prescriptive reading solution was released in August 2003. Client schools will run a second reading assessment on their students in January 2004. Statistical performance results will be published at that time.

The successful union of LT and LGL has led both teams to once again partner in the development and integration of another diagnostic/prescriptive system. The LT diagnostic/prescriptive math solution will be released in March 2004.

LT's curriculum is 100% research-based as it is written according to and completely aligned with state standards in correlation with standardized tests. LT's pedagogy is aligned with the "Information Process Theory of Learning."¹ Therefore, the crux of this paper is dedicated to the validity of the LGL diagnostic assessment component of the LT diagnostic/prescriptive reading solution.

Terms

Validity – An assessment instrument is valid to the extent that it actually assesses the underlying skill or construct it is designed to assess. A properly calibrated bathroom scale is a valid way to assess weight. Assessing the component skills underlying a phenomenon as complex as reading, however, is a much more complex task than assessing weight with a scale. The difference is that weight is a directly observable feature of physical reality; whereas, reading skills are latent (not directly observable) traits. The validity of an instrument designed to assess such latent traits is established in two ways:

- (1) Construct Validity: The theoretical connection between the instrument and the skill to be assessed—provided by the experts in the field who create the instrument.
- (2) Criterion Validity: The empirical connection between performance on the instrument and other outcomes recognized as correlates of the skill to be assessed—such as correlation with other assessment instruments or relevant behaviors.

Reliability – An assessment is reliable to the extent that its results are consistent over repeated administrations. Reliability is a necessary condition for an instrument to be valid. A perfectly valid and reliable instrument will give the same score over and over when assessing the same person in the same skill state. In reality, however, repeatedly giving the same assessment to a single individual would not give the same score, because the person's score would vary due to practice effects and other variables. So the reliability of an instrument must be established by other means, such as comparing one part of the instrument to another part (split-half reliability) or by comparing each part to the whole (computation of the "alpha" reliability coefficient).

Nationally Normed – This term is often misleading and carries more weight than it should. When a test is nationally normed, it means that a particular test has been given to an extremely large pool of test-takers across the nation and therefore scores can be compared to a national average. Unfortunately, nationally norming by itself says nothing about the accuracy or validity of a test. A test can be nationally normed and still be an invalid test. Therefore, it is important to note the test's validity first and foremost. Nationally norming then becomes relevant if a percentage comparison is needed against the national average. National norming is also necessary for accountability tests that rank individual programs to national averages.

Criterion-Referenced – Often tests fall into two categories, accountability testing and diagnostic testing. Diagnostic tests generally use criterion referencing. In other words, these tests compare specific abilities to detailed measures or standards. For instance, reading specialists may state that by the low 2nd grade students should have mastered certain phonological rules. If a student has not mastered those particular "criteria" then he or she is considered sub-2nd grade in that skill. The way that criteria are defined can vary by the experts who defined them. However, for diagnostic purposes, what is more important is that the same measurement is used to plot progress. For instance, a child may grow in height over the years. Whether one measures the child in inches or centimeters does not matter. What is important is that the measurer uses the same system so that when comparing measurements, growth or lack of growth can be recognized.

Researched-Based – This term is often used very loosely. Something is research-based if it was developed by recognized experts in a particular field. If these experts have statistical data or studies to support their finding the claim is stronger. Often ties to universities or other research organizations help clarify claims of being "research-based".

Foreword by Dr. Richard McCallum, Co-Founder of Let's Go Learn (Reading Assessment)

Let's Go Learn was founded on the belief that timely and accurate assessment data is a key component of successful learning. This fact is especially true in reading – parents and teachers need both diagnostic and on-going assessment data to make effective instructional decisions for students. The goal of the LGL Reading Assessment is to bring “best practices” in literacy assessment into an intelligent online application. To achieve this goal, we began to assemble a set of reading instruments, delivered online, that will provide:

- (1) Individualized assessment data in reading.
- (2) A management system for the reporting and analysis of student's scores.

The most recent version of the LGL Reading Assessment provides an online tool for collecting information that might normally be collected by a teacher or specialist using informal reading inventories, word lists, reading passages, and other classroom based diagnostic measures of reading ability. Our goal is to utilize the strengths of online technology to get individualized diagnostic assessment data in the hands of educators. Our first attempts at developing such tools have been warmly greeted by parents, educators and administrators in schools.

There are several distinct advantages for teachers using the LGL Reading Assessment. First, teachers save time using this tool. Collecting individualized assessment data is time consuming, and teachers will tell you that the time commitment alone is enough to mitigate against collecting such data. Second, when students are assessed in an online environment, no data is lost. That is, when a teacher or specialist assesses a child, subtle patterns in their behavior may be lost if the assessor is not highly trained and aware of the many nuances involved. In the LGL Reading Assessment though, the thoughtful design of the test items and the database structure, allow for all test data to be captured. For example, in the word analysis sub tests in the LGL Reading Assessment distracters were chosen with several key variables in mind: the nature of the sound pattern and it's position in the word. Over the course of a subtest this information can be used to identify subtle patterns in the student's response within the measure.

Educational Expertise on Dr. Richard McCallum, Co-Founder of Let's Go Learn

Let's Go Learn was co-founded by Richard McCallum, PhD. For the past eight years, Richard McCallum has been the Academic Coordinator for the Advanced Reading and Language Program in the Graduate School of Education, at the University of California, Berkeley. In Dr. McCallum's program, graduate students earn a Masters degree in Reading Education and a California teaching credential as a reading specialist. In addition, to the course work required for the degree, Dr. McCallum's graduate students also receive extensive field training through Cal Reads, a nationally recognized school-site intervention reading program.

Cal Reads provides individualized one-to-one tutoring for low achieving intermediate and middle school students and, as is the case with all effective interventions programs, Cal Reads administers individualized diagnostic reading assessments for all children served by the program. Based on these measures, an individual literacy profile is developed for every child. This profile provides the instructional roadmap

Cal Reads succeeds, in part, because the program collects both diagnostic and on-going assessment data on students. This detailed information is essential if we are to bring students' reading abilities back up to grade. Unfortunately, parents and classroom teachers are not in a position to collect the type of assessment data a reading specialists or intervention program might utilize. For this reason, Richard McCallum and a small group of other experts in education and web-based business technology founded Let's Go Learn.

LGL Reading Assessment: Construct Validity

Dr. Richard McCallum is a recognized expert in reading. His credentials at U.C. Berkeley as well as his published papers have clearly established his status. As the co-founder of Let's Go Learn, Dr. McCallum brought with him his rich experiences in reading assessment and instruction. With Dr. McCallum as the Chief Educational Architect, Let's Go Learn has gone far beyond just being an electronic version of a paper and pencil diagnostic reading assessment. It has combined the skills of reading specialists and best practices of literacy assessment into an intelligent online application. It is important to remember that even the best paper and pencil assessment is only as good as the certified reading specialists administering the test. Let's Go Learn has greatly reduced the need of a reading specialist in the first stage of reading remediation, which is diagnostic assessment. This will allow for the logical expansion of diagnostic assessment for all early readers, which today is impossible given the limited number of reading specialists that exist across the nation.

LGL Reading Assessment: Criterion Validity

The nationally recognized Cal Reads program started by Dr. Richard McCallum established diagnostic reading assessment as an essential part of reading remediation. It has proved that diagnostic assessment guiding targeted reading instruction can result in dramatic gains across schools, age, ethnicity and gender. The latest year's results showed a 3 and 1.8 year gain for two schools' respective students-served groups. The control groups in both only gained 0.5 years gain.

Let's Go Learn designed its assessment to measure the same reading-measures that Cal Reads uses for its assessment. Given that Dr. McCallum founded both Cal Reads and co-founded Let's Go Learn, the goal was to create an online assessment that could rival the human one-to-one assessments carried out by the reading experts of Cal Reads. Let's Go Learn succeeded in doing this by developing an intelligent and innovative custom web application that could mimic the decisions that a reading specialist would make when face to face with a test taker.

During the pilot performed in early 2002 funded in part by the U.S. Department of Education, Let's Go Learn demonstrated that its online assessment was highly correlated to the human reading specialists of Cal Reads. A portion of the final pilot report submitted to the U.S. Department of Education can be found in this document on page 6 under the subtitle, "Let's Go Learn Correlates Significantly to the Nationally Recognized Cal Reads Program."

Standards-Alignment and Criterion-Referencing

The Let's Go Learn Reading Assessment based its initial criteria on the California Standards for grade-level achievement in reading. In addition, it draws from the best practices of current literacy assessment. As a result, this format compares students' individual performance in six reading measures against a criterion of mastery for each sub-skill, resulting in an individual reading profile for every student.

Correlation to State Language Arts Standards

By virtue of being a diagnostic reading assessment that collects individual assessment data across six measures in reading, Let's Go Learn was able to correlate its assessment to many of the most rigorous State Language Art Content Standards. Immediately after a student completes an assessment, teachers or parents are able to view each student's performance against the State standards of their choice. Currently, 20 State standards are included. The remaining states are scheduled to be added by 2004.

Accuracy and Refinement

Assessment, by its very nature, attempts to measure a skill through a sub-sampling of test items. Common sense dictates that the shorter the test, the less accurate the test becomes. For this reason, the Let's Go Learn Reading Assessment sees itself in a class separate from the 10-15 minute reading assessments offered by multiple other educational publishers. Our assessment takes approximately 60 minutes and adapts to each test-taker as he or she undertakes an assessment. The detailed report that we produce for each student is far more accurate and diagnostic than what a "short" test can offer. Of course, assessment improvement is an ongoing process. Through detailed item analysis and sub-test refinement we are continuing to improve our assessment. By improving our assessments adaptation to the test-taker, we expect to be able to reduce the total test time by 25% by 2004. By performing ongoing item analysis with specific pools of students, we are improving our accuracy even more.

Our first major item analysis study took a pool of 1,000 students across the nation and examined their responses to each question across all 6 sub-tests. The very nature of being a web-based application means that no data is lost. As a result of this study, we used this data to improve our assessment and will continue to do so in the future.

Test-Retest Study: Q1 2003

Test-Retest is the ability of a test to be taken once and then immediately again and have similar results. Let's Go Learn has undertaken an initial study to gather data on the repeatability of the current Let's Go Learn assessment. Because the number of students involved in this initial analysis was smaller, the margin of error is higher than we could achieve with a larger study involving more test subjects. Nonetheless, the results were excellent. Variability was low meaning that the test can be re-administered with low bias.

Standard deviations (sd) measured in grades. Margin of error of this analysis: 22% Grade level variance is in criterion-referenced grade levels. SE is standard error for mean delta in grades.

• Sight-Word familiarity	sd=0.3	Grade level variance: 0.46	SE=0.15
• Word Recognition	sd=0.3	Grade level variance: 0.74	SE=0.20
• Word Analysis	sd=0.3	Grade level variance: 0.19	SE=0.10
• Word Meaning	sd=1.3	Grade level variance: 1.41	SE=0.28
• Spelling	sd=0.5	Grade level variance: 0.27	SE=0.12
• Silent Reading	sd=0.3	Grade level variance: 0.90	SE=0.21

Sub-test Specifications

- High-Frequency Words: 72 criterion referenced words. 24 words per grade from 1st to 3rd grade.
- Word Recognition: 120 criterion referenced words. 10 words per grade from 1st to 12th grade.
- Word Analysis: 80 criterion referenced words. 20 words per grade from 1st to 4th grade.
- Word Meaning: 60 criterion referenced words. 5 words per grade from 1st to 12th grade.
- Spelling: 60 criterion referenced words. 5 words per grade from 1st to 12th grade.
- Silent Reading: 12 Flesch-Kincaid leveled passages with 6 questions per passage. 1 passage per grade.

Reading Level Calculation Adherence

In the silent reading subtest of the Let's Go Learn Reading Assessment, the following method for reading level calculation was chosen. The 6th subtest, silent reading, is made up of 12 reading passages for grades 1st through 12th. These passages were systematically constructed to adhere to the Flesch-Kincaid Reading Grade Level Index. This index calculates the number of words, syllables, and sentences in a given passage. It then utilizes the average syllables per word and words per sentence to articulate a readability formula. After researching the methods used by text book publishers and children's book publishers our research found this to be the most reliable and widely accepted system for leveling reading material.

Let's Go Learn Correlates Significantly to the Nationally Recognized Cal Reads Program

In it's FIRST comparison to one-on-one paper and pencil assessments by Cal Read reading specialists, Let's Go Learn achieved the following correlations with statistical significance beyond the $\alpha=.01$ level:

- Sight-word familiarity $r=.89$ (n=17)
- Word recognition $r=.81$ (n=20)
- Word meaning $r=.60^*$ (n=20)
- Spelling $r=.78$ (n=20)
- Silent reading $r=.89$ (n=19)

This study was conducted in Tahoe/Truckee Unified School District in California in 2/02. Students were both tested by Cal Reads reading specialists and online using Let's Go Learn within a 3 week time period. The Word Analysis subtests were not compared because of incompatible methods in which Cal Reads and Let's Go Learn report their final results.

** Since this initial study, the Word Meaning component of Let's Go Learn has been significantly modified to improve its correlation.*

Item Analysis Major Revision: Q1, 2003

Item analysis can be performed across all valid assessments completed on the LGL Reading Assessment system because all data is recorded. A pool of 1000 students was selected from a total of 3100 completed assessments. This final pool of students were selected after eliminating evaluation accounts, suspect school pilots, internal testing accounts, or any other account that we had questions concerning its reliability.

Items across all six subtests with errors either above the 75% level or below the 25% level were flagged. These extreme values represent errors that are outside the mean error rate that we expect from a normally functioning item.

Example1: The word "and" received an usually high error rate. The word "an" was a distracter that was selected with a high percentage. Conclusion: The audio of "and" and "an" are too similar. Students might not be hearing the /d/ sound and thus think the target word is "an".

Example2: Target vocabulary word: "Caravan". Too often students choose a picture of 1 car and 1 van. Overall the error rate was too high for this particular word. Conclusion: Many students define "caravan" as the Dodge Caravan vehicle and not a line of camels walking through the dessert.

LGL Reading Assessment Comparison to Nationally Normed Paper and Pencil Assessments:

April 2003

Tested students (Grade range: 2-6) within one weeks time on the LGL Reading Assessment and the following paper and pencil tests.

LGL HFW subtest and the Slosson Oral Reading Test

LGL WR subtest and the Woodcock Word Identification Test

LGL WA subtest and the Woodcock Word Attack

- Correlation (HFW & SORT): 0.95 SE=0.073 (n=21)
- Correlation (WR & WI-W): 0.92 SE=0.088 (n=21)
- Correlation (WA & W-WA): 0.91 SE=0.097 (n=21)

High correlation demonstrates concurrent validity of the LGL Reading Assessment.

Recalculation of April SP and SR results + New Tests October 2003

LGL SP subtest and the WRAT

LGL SR subtest and the Gray Oral Reading Test

- Correlation (SP & WRAT): 0.85 SE=0.210 (n=21)
- Correlation (SP & GORT):* 0.65 SE=0.250 (n=21)

Medium to high correlations demonstrate concurrent validity of the LGL Reading Assessment.

** Lower correlations to the GORT is attributed to a high variability observed in the GORT results. Students inconsistently tested well above their grade levels on the GORT. Subsequent SR comparisons with a more consistent paper and pencil assessments is recommended.*

Awards

In September 2001, the U.S. Department of Education awarded Let's Go Learn a prestigious grant reserved for private companies. This SBIR grant seeks to support innovative new technologies that bring with them the ability to realize scalable solutions in education. The U.S. Department of Education recognized Let's Go Learn's unique team of experts and product plan and gave it extremely high marks across the entire spectrum of review criteria.

Next Steps for Learning Today and Let's Go Learn

- The new Learning Today diagnostic/prescriptive math solution will be released in March 2004.
- The entire Learning Today system will be aligned to national and all state standards by January 2004.
- Custom reports (i.e., AIPs, IEPs) will be available by January 2004.

1 Information Process Theory of Learning: Atkinson & Shiffrin, Kintsch; Klatsky, Loftus & Loftus; George Miller (1956); Newell, Shaw and Simon (1955-60);

Gagne' and Dick Anderson (1984); Rothkopf (1970); <http://tiger.coe.missouri.edu/~t377/IPTheorists.html>

a). Students are actively processing, storing and retrieving information

b). Teaching involves helping learners to develop information processing skills and apply them systematically to mastering the curriculum.

Cognitive structures relate to structure of the subject matter. Information processing emphasizes cognitive structures built by the learner.