RESEARCH & VALIDATION

Scholastic Reading Measure

Reliability and Validity Study

March 2019

SCHOLASTIC

ABOUT AMERICAN INSTITUTES FOR RESEARCH

Established in 1946, the American Institutes for Research (AIR) is an independent, nonpartisan, not-for-profit organization that conducts behavioral and social science research on important social issues and delivers technical assistance, both domestically and internationally, in the areas of education, health, and workforce productivity.

CONTACT

For more information about this report, please contact Scholastic Research & Validation at ScholasticRV@scholastic.com.

SUGGESTED CITATION

Salinger, T., Dumani, S., & Vestal, D. (2019). *Scholastic Reading Measure reliability and validity study*. Washington, DC: American Institutes for Research.

Scholastic is not responsible for the content of third-party websites and does not endorse any site or imply that the information on the site is error-free, correct, accurate, or reliable.

Item #: 691659 TM ® & © Scholastic Inc. All rights reserved.

SCHOLASTIC and associated logos are trademarks and/or registered trademarks of Scholastic Inc.

LEXILE and LEXILE FRAMEWORK are registered trademarks of MetaMetrics, Inc.

AIR is a registered trademark of American Institutes for Research.

Other company names, brand names, and product names are the property and/or trademarks of their respective owners. Scholastic does not endorse any product or business entity mentioned herein.

RESEARCH & VALIDATION

AIR®
AMERICAN INSTITUTES FOR RESEARCH®

Scholastic Reading Measure Reliability and Validity Study

March 2019

Contents

Executive Summary 1

The Scholastic Reading Measure 2

The Reading Measure Study: Phase I & Phase II 3

Phase I: Review of CAT Algorithm 3

Phase II: Reliability and Validity Study of the Reading Measure 3

Study Administration 4

Technology Audit 4

Participants 5

Observations 5

Results 5

Conclusion 9

References 10

Tables

Table 1. Test-Retest Reliability Coefficients for Different Demographic Groups 7

Table 2. Convergent Validity Coefficients for the Reading Measure 8

Executive Summary

The *Scholastic Reading Measure* is a low-stakes, computer-adaptive test (CAT) that educators can use to select books, articles, and short reads at the right level for students' independent reading.

Lexile® measures make it easy for educators to:

- Personalize learning;
- Measure student growth; and
- Communicate with parents about their child's progress.[1]

Scholastic Inc. collaborated with MetaMetrics® to create the *Reading Measure*, and Scholastic Research & Validation partnered with American Institutes for Research (AIR), an independent research firm, to conduct a study to determine if the *Reading Measure* yielded valid and reliable data about students' reading levels. There were two primary phases of the work on the *Scholastic Reading Measure*: the first phase was an expert review of the CAT algorithm and the second phase was a field-based reliability and validity study, which yielded positive results.

Pearson correlation coefficients were calculated to determine the test-retest reliability of the *Reading Measure*. The overall reliability of the *Reading Measure* across two administrations was found to be high: r = .90. These data are strong evidence of the test-retest reliability of the *Reading Measure*.

In order to determine the convergent validity coefficient, Lexile measures collected from the reading portion of the *Scantron Performance Series*® fall assessment, provided by the study's participating school district, were correlated separately with Lexile measures obtained from two separate administrations of the *Reading Measure*. Both administrations of the *Reading Measure* correlated highly and significantly with the *Scantron Performance Series* assessment (r = .83 for Administration 1, r = .78 for Administration 2, and r = .84 when the average measures were used). These data are strong evidence of the convergent validity of the *Reading Measure*.

Analysis of the data collected in two administrations of the *Reading Measure* and data supplied by the district in which the study was conducted has confirmed the reliability and validity of the *Reading Measure*; that is, the resulting Lexile measures for two administrations of the *Reading Measure* were consistent and had high levels of convergent validity with the *Scantron Performance Series* assessment Lexile measure.

These results mean that teachers and students can use the Lexile measures determined by the *Reading Measure* to confidently identify books, articles, and short reads at the right level for students' independent reading. Results also support using the *Reading Measure* in conjunction with other information to track students' reading progress.

[1] Retrieved from: https://lexile.com/educators/understanding-lexile-measures/

The Scholastic Reading Measure

The *Scholastic Reading Measure* is a computer-adaptive test (CAT) developed by Scholastic Inc. and MetaMetrics® that can be used by teachers and students to determine students' reading levels, as indicated by resulting Lexile® measures. Lexile measures are also routinely used to indicate the difficulty of a book or other text in a unique way that is not tied directly to a student's grade level. The two, when used together, can be informative in instructional settings because Lexile measures guide teachers to identify and recommend books, articles, and short reads that will be at the right level for their students' independent reading. Independent reading is a strong contributor to students' reading achievement and helps them become proficient and advanced readers (Allington, 2014; Allington & Gabriel, 2012). One measure of students' progress as readers is that they are able to read more and more complex texts—and do so on their own.

A student logging into the *Reading Measure* begins with a brief tutorial and one practice question. Next, the student begins the measure with a few questions at grade level.[2] The student then progresses through approximately 33 short reading passages, followed by a series of corresponding multiple-choice questions. A student's performance on each question determines whether the difficulty level of the next passage is lowered or becomes increasingly challenging. The *Reading Measure* continues to present passages and corresponding questions until a level of certainty is reached about a student's Lexile measure or the maximum number of questions per administration (33) have been answered.

[2] Prior to students taking the *Reading Measure*, educators can set different parameters, such as a benchmark (choosing beginning, middle, or end of year) and teacher appraisal (choosing below, on, or above grade level). Depending on the parameters that educators set for each student, the Lexile measure of the first question a student receives will vary.

The Reading Measure Study: Phase I & II

There were two primary phases of the *Scholastic Reading Measure* study: the first phase was an expert review of the CAT algorithm and the second phase was a field-based reliability and validity study of the *Reading Measure* with American Institutes for Research (AIR), an independent research firm.

Phase I: Review of CAT Algorithm

A senior psychometrician at the independent research firm (AIR), who is an expert on CATs, reviewed the *Reading Measure* CAT algorithm.

Phase II: Reliability and Validity Study of the Reading Measure

Prior to initiating the reliability and validity study, the *Reading Measure* was subjected to usability and other product testing to evaluate how real students would interact with the measure and to address any technical issues that might affect the performance of the *Reading Measure*.

A field-based reliability and validity study was conducted in order to ensure:

- confidence in replication of consistent results (reliability)
- confidence in the accuracy of the measure (validity)

Approximately 10,000 students in Grades 1–6 were targeted for participation in the study to ensure a robust sample for analyses, with approximately 500 students at each grade level providing complete data (i.e., two complete administrations of the *Reading Measure* and either progress monitoring/formative assessment data or state reading assessment scores). Additionally, a minimum of 400 students with complete data were targeted for subgroup analyses for English learners, students with disabilities or those eligible for special education services, students receiving free and reduced-price lunch, and other groups of interest.

As part of the reliability and validity study, the independent research firm also intended to gather contextual data through classroom observations.

In order to conduct the reliability study using the test-retest model, students would take the *Reading Measure* two times in a short time frame: *Reading Measure* during Week 1 (Administration 1), no research activities in Week 2, and *Reading Measure* during Week 3 (Administration 2). Additionally, in order to conduct the validity study, data from the reliability study would be analyzed in conjunction with achievement data from the district for students who participated in the reliability study.

Study Administration

Once a participating district was identified, Scholastic, AIR, and the district entered into a memorandum of understanding for the study. All parties were in compliance with the standard requirements for data sharing and secured Institutional Review Board approval for procedures and all materials used. Information regarding the study was provided to all school staff and students' families and all participants had the opportunity to refuse participation without repercussions.

A school district in the Southeast with more than 30,000 enrolled students participated in the *Scholastic Reading Measure* reliability and validity study. Starting in kindergarten, all students in this school district receive a digital device from the district. The student-teacher ratio averages 13:1, and the high school graduation rate is 80%. According to the U.S. Census, the racial/ethnic composition of the district is 87.1% White, 9.6% Black or African American, and 4.6% Hispanic or Latino.

The district administers the *Scantron Performance Series*® assessment to students in Grades 1–10 for progress monitoring in the fall, winter, and spring, and currently uses it as the state summative assessment. This is a CAT that can be used to measure proficiency as well as determine program placement.[3] As part of the output, the *Scantron Performance Series* assessment provides a Lexile measure for students in all grades. This Lexile measure was provided as part of the study and used to conduct validity analyses for students in Grades 1–6 who participated in the *Reading Measure* study.

Technology Audit

Prior to the *Reading Measure* reliability and validity study, a Scholastic team visited two schools in the district to use the *Reading Measure* on the district's technology and network to ensure the best possible experience for students during the study. Student accounts were created specifically for the audit and used on district issued laptops connected to school wireless Internet. The team mimicked research conditions at two schools on multiple accounts. At the end of the audit, both Scholastic and the district were satisfied with the outcome and confident that students would have a seamless experience using the *Reading Measure* during the reliability and validity study.

[3] *The Scantron Performance Series* assessment is a Web-based CAT that provides scaled scores to measure proficiency and national norming information. The assessment can be used to personalize learning, measure growth over time, serve as a universal screener, and/or determine program placement. For more information on the assessment, please visit www.performanceseries.com.

Participants

Twenty-four elementary/intermediate schools participated in the study, inclusive of students in Grades 1–6. Consented students completed the *Reading Measure* in two separate administrations (referred to as Administration 1 and Administration 2) during one month in the fall. Classroom teachers or other school staff (e.g. Reading Specialists, Library-Media Specialists, etc.) read students the informed assent information, provided instructions for the *Reading Measure*, and then monitored administration sessions. Students took the *Reading Measure* on their own school-assigned devices.

Participating schools, teachers, and students were compensated for their participation.

Observations

Observations conducted during administration of the *Reading Measure* indicated that students, even the youngest ones, appeared familiar with technology and easily located the *Reading Measure* application in the Web portal on their devices. Students worked their own way through the *Reading Measure* and no significant issues or difficulties were observed.

Results

The reliability and validity studies' results are predicated on the assumption that the *Reading Measure* will be used only for low-stakes decision-making in the service of instruction, either in direct conjunction with *Scholastic Literacy Pro*®[4] or independently. As such, the *Reading Measure* cannot be used to make high-stakes decisions; for example, as part of the criteria to determine students' grade retention or as the primary measure on which assignment to a Tier 2 or 3 intervention is based. Additionally, the *Reading Measure* has not been validated for use in evaluating teachers' performance.

[4] *Scholastic Literacy Pro* is a blended independent reading program designed for K–6 students that recommends personalized collections of books for students that are aligned with their reading level and self-identified interests. Students access *Literacy Pro* at school or at home from a computer or tablet with an Internet connection; the program allows them to set and monitor their personal independent reading goals and complete *Think More* reading comprehension check-in activities. Teachers can use *Literacy Pro* as a digital classroom management tool to monitor students' independent reading both at school and home. The program also produces reports that when used in conjunction with other information can help school administrators track students' reading activities and progress.

Findings from the Reliability Study

All students whose data were included in the analyses completed the *Reading Measure* in two separate administrations. The Lexile measures obtained from these two separate administrations were used to calculate the test-retest reliability of the *Reading Measure*. Due to the nature of a CAT, the same student may not receive the same set of questions across two administrations of the *Reading Measure*; thus the test-retest reliability of the current *Reading Measure* relies on the assumption that the different items administered in both administrations are still measuring the same underlying reading ability.

The initial data sets included 11,329 entries for Administration 1 and 10,506 entries for Administration 2 of the *Reading Measure*. After cleaning the data set, the final matched sample included 7,425 students who completed the *Reading Measure* for both administrations and had valid final Lexile measures. This data was matched with a data set that included demographic information provided by the district. The demographic information used for subsequent analyses included grade, gender, race, free and reduced-price lunch eligibility, English learner, special education, and gifted student status information.

The independent research firm calculated Pearson correlation coefficients to determine the test-retest reliability of the *Reading Measure*. The overall reliability for the *Reading Measure* across the two administrations was found to be high: $r = .90$.

The breakdown of the test-retest reliability, based on different demographic groups, is presented in Table 1. In every instance, the test-retest reliability of the *Reading Measure* remained significant and desirable (ranging from $r = .76$ to $r = .94$). These results are strong evidence of the test-retest reliability of the *Reading Measure*.

Table 1. Test-Retest Reliability Coefficients for Different Demographic Groups

Demographic Group	Reliability	
	r	*N*
Male	.90	3,815
Female	.90	3,605
White	.90	5,080
Black/African American	.86	828
Hispanic	.90	843
Asian	.93	71
American Indian/Alaskan Native[a]	.94	18
Native Hawaiian or Other Pacific Islander[a]	.78	5
Multiple Races	.90	562
Grade 1	.83	1,238
Grade 2	.87	1,280
Grade 3	.83	1,278
Grade 4	.84	1,401
Grade 5	.82	1,267
Grade 6	.82	956
Free/Reduced-Price Lunch—Yes	.88	3,466
Free/Reduced-Price Lunch—No	.90	3,954
English Learner—Yes	.86	455
English Learner—No	.90	6,965
Special Education—Yes	.86	993
Special Education—No	.90	6,427
Gifted—Yes	.76	718
Gifted—No	.89	6,702

Note: *r* indicates Pearson correlation coefficient and *N* indicates the sample size.

[a] indicates that the reliability coefficient should be interpreted with caution due to insufficient sample size.

Findings from the Validity Study

The independent research firm also calculated the validity coefficient for the *Reading Measure*. The Lexile measures from the reading portion of the *Scantron Performance Series* fall assessment were correlated separately with Administration 1 and Administration 2 Lexile measures from the *Reading Measure* to obtain the convergent validity coefficient. The *Scantron Performance Series* assessment Lexile measures can be interpreted the same way as the Lexile measures on the *Reading Measure*.

Convergent validity is one source of construct validity, and it shows the degree to which two different measures of constructs that are theoretically similar are related to each other (Campbell & Fiske, 1959). One would conclude that a measure has convergent validity if it is highly correlated to another measure that assesses the same underlying construct. Therefore, the current validity study was conducted under the assumption that both the Lexile measure from the *Reading Measure* and the Lexile measure from the reading portion of the *Scantron Performance Series* assessment are comparable assessments of reading ability with similar theoretical underpinnings (Stenner, Smith, & Burdick, 1983).

A final sample of 6,148 students who had complete data on both the *Reading Measure* and the *Scantron Performance Series* assessment Lexile measure were used for the validity analyses. Both administrations of the *Reading Measure* correlated highly and significantly with the *Scantron Performance Series* assessment (r = .83 for Administration 1, r = .78 for Administration 2, and r = .84 when the mean scores for Administrations 1 and 2 were used). Table 2 presents these coefficients. Overall, these data are strong evidence of the convergent validity of the *Reading Measure*.

Table 2. Convergent Validity Coefficients for the Reading Measure

Administration	Validity	
	r	*N*
Administration 1	.83	6,148
Administration 2	.78	6,148
Overall	.84	6,148

Note: *r* indicates Pearson correlation coefficient and *N* indicates the sample size.

Conclusions

Analysis of the data collected in two administrations of the *Scholastic Reading Measure* has confirmed its reliability and validity; that is, the resulting Lexile measures for the two administrations were consistent and had high levels of convergent validity with the *Scantron Performance Series* assessment Lexile measure. Teachers and students can use the *Reading Measure*'s resulting Lexile measures confidently as they identify books, articles, and short reads that are at the right level for students' independent reading. Results also support using the *Reading Measure* Lexile measures to inform instruction and in conjunction with other information to track students' reading progress.

References

Allington, R. L. (2014). How reading volume affects both reading fluency and reading achievement. *International Electronic Journal of Elementary Education*, *7*(1), 13–26.

Allington, R. L. & Gabriel, R. (2012). Every child, every day. *Educational Leadership*, *69*(6), 10–15.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multi-trait method matrix. *Psychological Bulletin*, *56*(2), 81–105.

Stenner, A. J., Smith, M., III, & Burdick, D. S. (1983). Toward a theory of construct definition. *Journal of Educational Measurement*, *20*(4), 305–316. Retrieved from https://doi.org/10.1111/j.1745-3984.1983.tb00209.x

Notes

SCHOLASTIC

Item #691659