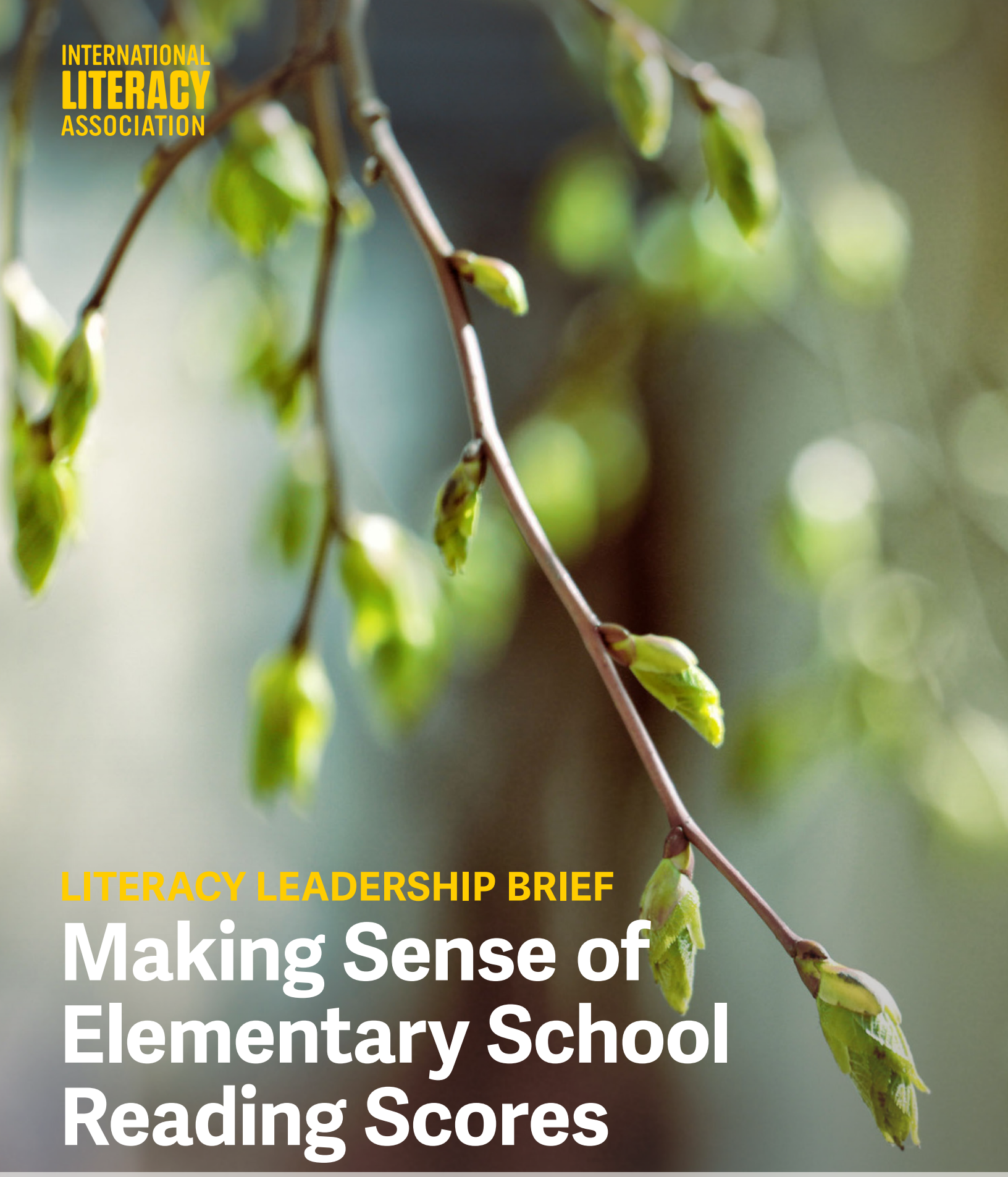

INTERNATIONAL LITERACY ASSOCIATION

LITERACY LEADERSHIP BRIEF

Making Sense of Elementary School Reading Scores

International Literacy Association | 2020

Mr. Garcia teaches second grade in a rural school district in southeastern United States. Last March, he opened his reading test spreadsheet file on his computer. The spreadsheet listed each of the 28 children in his class along with columns showing each student's scores for the 13 reading tests administered during the school year. There were test scores for Mr. Garcia's personally constructed assessments for students' sight word and phonics development and for overall reading level. Then there were state-required assessments to gauge students' word reading fluency and comprehension—repeated three times a year. And, finally, there was a column for the state-required end-of-grade standardized test to be given in May. On that March day, Mr. Garcia sat back, scanned the spreadsheet, took a deep breath, and said to himself, "Is all this necessary? What am I learning from all these scores? What's happening to my students as a result of all this assessment?"

Mr. Garcia is not data driven, he's data *overwhelmed*. He values some of the reading assessments because they help him to make instructional decisions. He likes his teacher-created checklists for students' sight word and phonics development. They are useful, especially at the beginning of the year, as they give him clues for instruction.

Other scores are less useful to him. The district-required progress monitoring test scores are meant to help Mr. Garcia to know his students' progress toward the end-of-year goals. But they don't help him much because they tend to confirm what he already knows from his day-to-day instruction about each child's reading progress. In aggregate, those reading scores might help the principal or the school district officials to get a sense of how things are going for the entire population of students, but their value in daily instruction may be limited.

What matters most for elementary-grade teachers when thinking about all those reading scores, and what could policymakers do to help teachers? Three positions are worth pursuing in this regard:

1. Every reading assessment should have a clear purpose.
2. Formal reading test score use must be supported by validity and reliability evidence.
3. Testing should be done only when necessary.

Testing should be done only when necessary.

Clear Purpose

Reading scores exist for a continuum of purposes, from informal assessment to formal standardized tests.

Reading scores exist for a continuum of purposes, from informal assessment to formal standardized tests. Informal teacher-made checklists used on a day-to-day basis can indicate the degree to which some very specific reading knowledge has been mastered, such as knowledge of particular phonics patterns or sight words. Or a running record or informal reading inventory may provide an indicator of a student's overall reading level. In those cases, the purpose of the score is to directly inform instructional planning and implementation.

Reading scores obtained from benchmark assessments may be useful to teachers, administrators, and policymakers alike, or useful to just one of those groups. Benchmark assessments are administered multiple times in a school year, marking a student's progress for a particular reading ability. For instance, they may be obtained for a student's oral reading fluency. Some benchmark scores provide information that is useful to teachers. For example, knowing the current status of students' phonics knowledge compared with a benchmark goal may lead a teacher to intensify phonics instruction. Those benchmark scores—when aggregated for the whole school—can also provide administrators with data for determining the best allocation of instructional resources, such as which intervention supports to make available to second-grade students.

Reading scores from formal standardized tests and end-of-grade assessments tend to be most advantageous to administrators and policymakers for accountability purposes. Families may also appreciate those scores as they allow them to monitor their children's progress compared with other children in the district or state.

Validity and Reliability

Validity is not a property of a reading test. Rather, it references the degree to which evidence supports the use of test data to make particular decisions or take specific actions. For instance, when teachers use reading scores from standardized tests for a given purpose, they should look for evidence that supports using those scores *for that particular purpose*. If, for instance, a reading comprehension test is overly affected by other factors, such as students' topic knowledge, then the test results should

not be used to make decisions about reading comprehension instruction.

Various sources of validity evidence exist. One source is convergent validity. Different tests of the same skills should obtain similar results. One way of determining validity is to correlate the results of one reading test with the results of another. As an example, one might expect students' scores from one reading vocabulary assessment, such as a newly published test, to be highly correlated with scores on another reading vocabulary assessment. A strong positive correlation provides convergent validity evidence to support the use of the new reading vocabulary test for the kinds of decisions that were previously made with the older assessment.

Reliability is the degree to which items on a test consistently measure the same thing (internal reliability) as well as how stable test scores are across multiple administrations (test–retest reliability). We would want all of the items on a reading test to assess students' reading consistently. Likewise, if a student were to take a reading test again on a different day, we would expect the results to be similar.

Reliability estimates are usually reported as coefficients that range from 0.00 to 1.00. There is no definite cut-point for considering whether a reliability estimate is "good." Evaluation of what is sufficient depends on the type of reliability reported and how the test scores are used. But general guidelines suggest that an internal reliability coefficient (e.g., Coefficient Alpha or KR20) of .90 and above provides excellent indication of reliability, below .70 is inadequate and the resulting scores may have limited applicability, and scores between .70 and .90 are sufficient for most purposes.

When Validity Evidence and Reliability Estimates Matter Most

Particularly strong validity evidence and reliability estimates are essential if reading scores are used for a consequential purpose.

Particularly strong validity evidence and reliability estimates are essential if reading scores are used for a consequential purpose, such as determining whether an individual student will be promoted to fourth grade. In other situations, less validity evidence and lower reliability are acceptable. Generally, informal, day-to-day scores from teacher-made assessments require little formal technical support. The main interest in such a case is face validity—do the test items appear to measure the construct they are supposed to measure? And matching a student's

reading level to a book level for next week's independent reading is a low-stakes decision that can easily be adjusted if the book proves to be too hard or too easy.

Assessment Procedures Matter

Importantly, to ensure valid interpretations of reading scores and reliable scores, test makers' assessment procedures must be followed precisely. Unfortunately, shortcuts are taken sometimes, which may render the resulting reading scores unreliable and the interpretations of the scores invalid. For example, one popular benchmark test requires students to read two text passages aloud for one minute each. If, in the interest of saving time, teachers use only a single passage, the score validity and reliability is seriously impacted, and the resulting scores become less trustworthy.

Reading scores tend to be more reliable and score interpretations more valid when the test materials are closest to the student's performance levels versus when materials are too hard or too easy for the student. As an example, third-grade readers who are two years behind in reading may be evaluated more profitably with first-grade texts rather than third-grade texts.

Test Only When Necessary

There has been a growing concern about the amount of testing that students are now being subjected to in our schools.

In the past few years, there has been a growing concern about the amount of testing that students are now being subjected to in our schools. Indeed, in one study of 66 city school districts, the average number of mandated tests (mainly reading and mathematics tests) from prekindergarten through fifth grade was 48.5! Anecdotal reports in the public press suggest that some students feel so much test anxiety that they don't want to go to school, and that, because tests tend to drive instruction, instructional time is spent on preparing for and taking tests rather than for learning to read. As a result of the tumult, the United States Department of Education has spelled out ways that states might reduce redundant and low-quality tests, and some state school chiefs have been reviewing the panoply of required tests from kindergarten through 12th grade in an effort to reduce testing.

What can be done regarding over-testing of reading? One solution is to eliminate short-term retesting. Sometimes educators and policymakers require repeated reading testing primarily for the appearance of rigor and vigilance. But repeated

Repeated testing in a relatively short time span yields little or no new information.

testing in a relatively short time span yields little or no new information and, as students become fatigued at the repetition, may even be untrustworthy. During repeated short-term testing, a student's reading scores will tend to fluctuate. One reason for score fluctuation is that all reading scores have some degree of *un*reliability, and thus changes in scores, especially when repeated in a short time span, do not necessarily signal any actual change in a student's reading ability.

Educators and policymakers could consider two factors to determine a repeated test schedule that would yield substantive and usable information about students' reading abilities. They could consider the size of the standard error of measurement for the reading assessment in conjunction with how quickly students typically progress in the reading skill or ability under examination. Standard error of measurement is an indicator of how much a student's repeated scores will vary around the student's "true" reading score.

Here is an example of how the two factors can be used to determine reasonable spacing between repeated tests. Let's say district officials want to closely monitor their second-grade students' oral reading fluency measured as words read correctly per minute. First, imagine that the oral reading test's standard error of measurement is 10 words correct per minute, that is, with repeated testing, students' scores on this test are likely to fluctuate by about 10 words per minute. That would mean if a student were tested and received a score of 60 words per minute, upon retesting it would be common to see this score vary between 50 and 70 words per minute.

Second, let's say that district officials know from national fluency norms or from previous district data that second-grade students tend to improve oral reading fluency by about one word per week on average. Taken together, the standard error of the test and the typical rate of learning would suggest readministering this test about every 2.5 months, rather than weekly as some testing mandates require.

Likewise, teachers sometimes administer unnecessary tests. For instance, if they have been teaching reading from a core textbook program, the publisher may provide weekly or unit tests to assess students' reading abilities for the content covered. Many teachers already have a very clear understanding of their students' performance level from daily observation. In

those cases, time spent on instruction would be more beneficial than time spent on testing.

What Can Policymakers Do?

Policymakers should be cautious when mandating literacy assessments. First, they should make sure that the testing required to accomplish particular accountability or reporting mandates is truly necessary. When requiring particular regimens of ongoing monitoring assessments, they should consider teachers' opinions and experiences. Requiring the analysis of teacher insights into the usefulness of particular assessment approaches (through interviews, polling, focus groups) should be a first step toward shaping reading assessments and their timing.

Second, for every mandated reading test, policymakers should ensure that teachers know *why* a test is being given—and how it will be used. Teachers should be told whether a reading assessment score is intended to support instructional decisions or whether it will be used for accountability or public reporting, and teachers should be informed about the frequency with which such a test can be administered and readministered usefully.

Third, policymakers should require that all compulsory reading tests provide validity evidence and reliability estimates. As well, policy should encourage out-of-grade-level assessment when appropriate (such as when a test score is to be used for instructional purposes).

State and district policymakers should periodically review mandated reading tests to continually improve the pool of useful reading scores.

Fourth, state and district policymakers should periodically review mandated reading tests to continually improve the pool of useful reading scores available to teachers and to minimize accountability assessment. Elementary reading instruction plays a critical role in the development of students' reading abilities, and strong reading ability is at the center of disciplinary learning. Some amount of reading assessment is vital to support quality reading instruction, to keep families informed about how their children are doing, and to allow for reasonable monitoring of instructional success. However, efforts are needed to minimize testing duplication and to instead focus on assessment only for clear purpose.

BIBLIOGRAPHY

Alvarez, L. (2014, November 9). States listen as parents give rampant testing an F. *The New York Times*. Retrieved from www.nytimes.com/2014/11/10/us/states-listen-as-parents-give-rampant-testing-an-f.html

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Association of Test Publishers. (n.d.). *Questions about testing in schools*. Retrieved from www.testpublishers.org/testing-in-schools

Freedman, M., & Houtz, J. (2004, March). *A glossary of terms used in educational assessment*. Retrieved from www.davidsongifted.org/Search-Database/entry/A10461

International Literacy Association. (2017). *Literacy assessment: What everyone needs to know*. [Literacy leadership brief]. Newark, DE: Author. Retrieved from www.literacyworldwide.org/docs/default-source/where-we-stand/literacy-assessment-brief.pdf?sfvrsn=efd4a68e_4

International Literacy Association. (2017). *The roles of standardized reading tests in schools* [Literacy leadership brief]. Newark, DE: Author. Retrieved from www.literacyworldwide.org/docs/default-source/where-we-stand/ila-roles-standardized-reading-tests-in-schools.pdf?sfvrsn=c6ada58e_4

International Literacy Association. (2018). *Beyond the numbers: Using data for instructional decision making* [Literacy leadership brief]. Newark, DE: Author. Retrieved from www.literacyworldwide.org/docs/default-source/where-we-stand/ila-beyond-the-numbers.pdf

International Literacy Association. (2018). *Exploring the 2017 NAEP reading results: Systemic reforms beat simplistic solutions* [Literacy leadership brief]. Newark, DE: Author. Retrieved from www. literacyworldwide.org/docs/default-source/where-we-stand/ila-exploring-the-2017-naep-reading-results.pdf

Layton, L. (2015, October 24). Study says standardized testing is overwhelming nation's public schools. *The Washington Post*. Retrieved from www.washingtonpost.com/local/education/study-says-standardized-testing-is-overwhelming-nations-public-schools/2015/10/24/8a22092c-79ae-11e5-a958-d889faf561dc_story.html?utm_term=.256028e54642

National School Boards Association. (n.d.). *Assessment 101: Supporting high quality assessment systems: Glossary*. Alexandria, VA: Author. Retrieved from www.achieve.org/files/Assessment101_Glossary.pdf

Ujifusa, A. (2015, July 8). Amid cries of overtesting, a crazy quilt of state responses. *Education Week*. Retrieved from www.edweek.org/ew/articles/2015/07/08/amid-cries-of-overtesting-a-crazy-quilt.html

International Literacy Association: Literacy Research Panel 2018–2019

Principal Authors

Jill Fitzgerald, Research Professor, The University of North Carolina at Chapel Hill
Timothy E. Shanahan, Distinguished Professor Emeritus, University of Illinois at Chicago

Panel Chair

Diane Lapp, San Diego State University

Panel Members

Dorit Aram, Tel Aviv University, Israel
Diane Barone, University of Nevada, Reno
Eurydice B. Bauer, University of South Carolina
Nancy Frey, San Diego State University
Andy Goodwyn, University of Bedfordshire, England
Jim V. Hoffman, University of North Texas
David E. Kirkland, New York University, Steinhardt
Melanie Kuhn, Purdue University College of Education
Maureen McLaughlin, East Stroudsburg University of Pennsylvania
Heidi Anne E. Mesmer, Virginia Tech
Donna Ogle, National Louis University
D. Ray Reutzel, University of Wyoming, Laramie
Alyson Simpson, University of Sydney, Australia
Jennifer D. Turner, University of Maryland
Amy Wilson-Lopez, Utah State University
Jo Worthy, University of Texas, Austin
Ruth Yopp-Edwards, California State University, Fullerton
Hallie Yopp Slowik, California State University, Fullerton

Kathy N. Headley, Clemson University, President and Board Liaison, International Literacy Association
Bernadette Dwyer, Dublin City University, Ireland, Immediate Past President, International Literacy Association
Stephen Peters, Laurens County School District 55, Vice President, International Literacy Association
Marcie Craig Post, Executive Director, International Literacy Association

This literacy leadership brief is available in PDF form for free download through the International Literacy Association's website: literacyworldwide.org/statements.

Media Contact: For all media inquiries, please contact press@reading.org.

Suggested APA Reference

International Literacy Association. (2020). *Making sense of elementary school reading scores* [Literacy leadership brief]. Newark, DE: Author.

About the International Literacy Association

The International Literacy Association (ILA) is a global advocacy and membership organization dedicated to advancing literacy for all through its network of more than 300,000 literacy educators, researchers, and experts across 146 countries. With over 60 years of experience, ILA has set the standard for how literacy is defined, taught, and evaluated. ILA's *Standards for the Preparation of Literacy Professionals 2017* provides an evidence-based benchmark for the development and evaluation of literacy professional preparation programs. ILA collaborates with partners across the world to develop, gather, and disseminate high-quality resources, best practices, and cutting-edge research to empower educators, inspire students, and inform policymakers. ILA publishes *The Reading Teacher*, *Journal of Adolescent & Adult Literacy*, and *Reading Research Quarterly*, which are peer reviewed and edited by leaders in the field. For more information, visit literacyworldwide.org.

@ILAToday
/ILAToday
/InternationalLiteracyAssociation
literacyworldwide.org