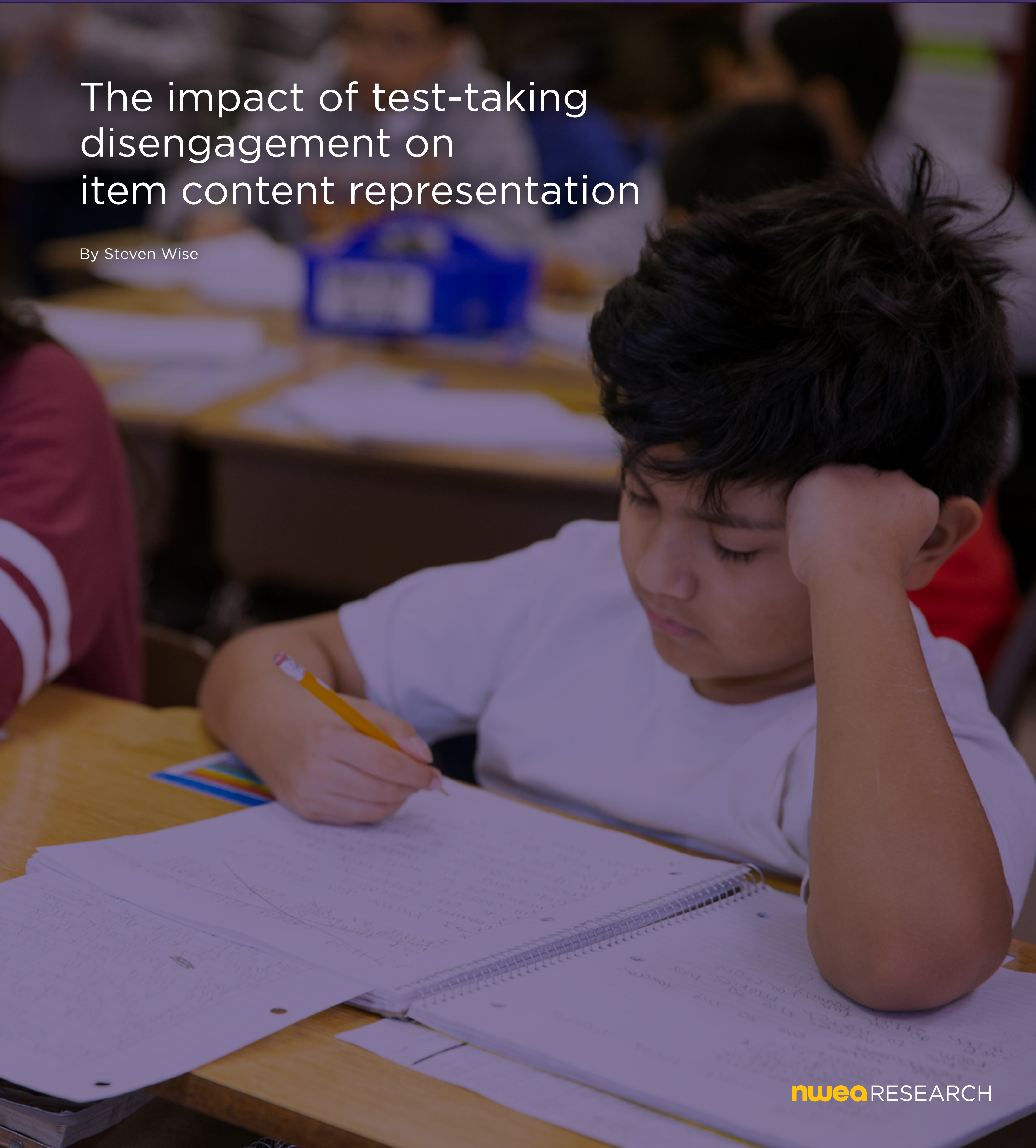
COLLABORATIVE FOR
STUDENT GROWTH

BRIEF



The impact of test-taking disengagement on item content representation

By Steven Wise

nwea RESEARCH

KEY FINDINGS

- **Rapid guessing was much more frequent in some content areas than in others.**
- **This differential rapid guessing often resulted in test events with meaningfully distorted content representation.**
- **Content misrepresentation was higher on tests that had higher levels of rapid guessing, and was more frequent in reading assessments than in mathematics.** In reading, over 31% of Reading 2–5 tests and 44% of Reading 6+ tests with high disengagement showed meaningful content misrepresentation. In mathematics, over 11% of Math 2–5 tests and nearly 17% of Math 6+ tests with high rates of rapid guessing showed meaningful content misrepresentation.
- **Differences in test taking engagement across content categories was primarily due to differences in how much reading items required.** Analysis showed that, after adjusting item reading load and depth of knowledge across different content areas, differences in rapid guessing rates across content categories were minor beyond what could be explained by reading load differences.

A subject like math or reading is complex. Math isn't just math, but rather encompasses many related areas of understanding and applying concepts in geometry, algebra, and statistics. Similarly, reading isn't just reading, but includes vocabulary and different aspects of understanding informational and literary texts. To provide consistent and valid insights into what students know and can do in a subject, when assessments are designed, they follow specific blueprints that balance questions across content areas within the subject. This ensures that the assessment questions and resulting test scores are representative of what students have learned and can do in the subject as a whole.

However, research has shown that test takers do not always answer questions effortfully, and that test taking disengagement can distort scores to the extent that results do not accurately reflect what students know and can do. Rapid guessing—when students answer a question so quickly that they could not have understood the question's content—can negatively impact measurement quality in several ways. It can distort test performance by lowering scores, sometimes by a large amount , since the accuracy of rapid guesses is typically much lower than that of effortful responses. Rapid guessing also decreases the precision of scores. For example, if a test taker rapidly-guesses on 10 items on a 40-item test, from a measurement standpoint, only 30 of the item responses provide useful information about the test taker's achievement level. This decrease in precision is often hidden: because standard errors are typically calculated from all item responses, and not just the ones for which the test taker was engaged, when rapid guessing occurs, reported standard errors tend to overestimate score precision. Researchers have developed approaches to decrease the impact of rapid guessing by filtering out or de-emphasizing rapid guessing in scoring[ii,iii]. A third potential measurement cost of rapid guessing, though, has received little research attention and cannot be addressed by such adjustments to scoring: if students rapidly-guess more in some content areas than in others on a test, then the test's content representation may be distorted so the score provides a less valid measure of what a student knows and can do in the subject as a whole. Using two studies, this work sought to explore the relationship between item content and rapid-guessing behavior.

The first study used data from MAP® Growth™ adaptive assessments from about 250,000 students in grades 2 through 10 in a single state in math and reading, to identify and quantify rapid guesses in different content areas. MAP Growth assessments include a balance of questions from instructional areas aligned to state content standards. In each subject, two different assessments were used to provide suitable questions across the grade span: *Reading 2–5* and *Reading 6+*, and *Math 2–5* and *Math 6+*. The assessments included four instructional areas in mathematics and five instructional areas in reading.

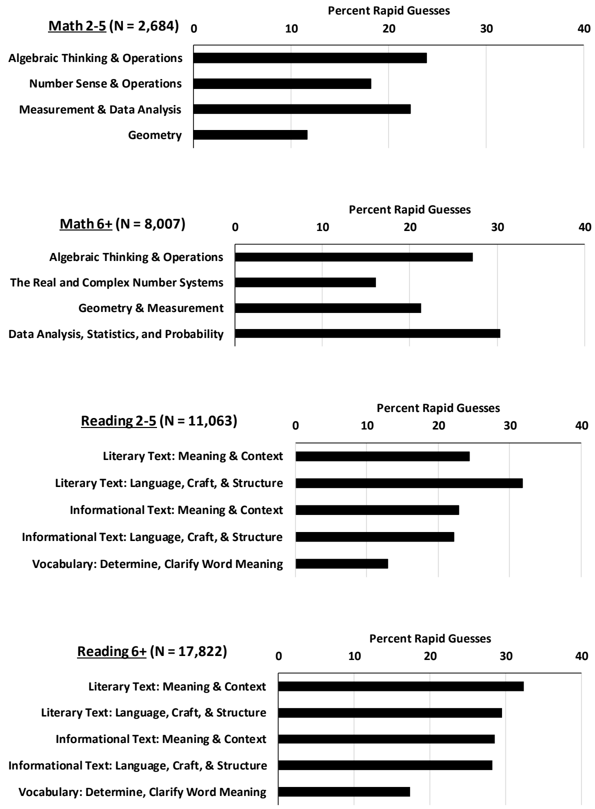
The first study addressed two questions:

1. How do the rates of rapid guessing vary across content categories?
2. If rapid guessing rates differ, how frequently does this meaningfully distort the content representativeness from individual test events?

Rates of rapid guessing differed markedly across instructional areas.

In most tests, rapid guessing rates were fairly low, though, consistent with other research, were higher in reading than in mathematics and more frequent in higher grades. Response time effort (RTE), a measure that shows the proportion of questions on a test on which a student's response times suggest he or she answered questions effortfully rather than rapidly guessing, was 0.90 or above in about 90 percent of Math 2–5 assessments, 76 percent of Math 6+ assessments, 71 percent of Reading 2–5 assessments, and 58 percent of Reading 6+ assessments.

However, for test events with moderate to high levels of rapid guessing, in both math and in reading, rapid guessing rates were much higher in some instructional areas than in others. For the Math 2–5 assessment, rapid guessing to Geometry questions occurred about half as often as to questions in Algebraic Thinking and Operations. For Math 6+, rapid guessing to Data Analysis, Statistics, and Probability questions occurred at nearly twice the rate to The Real and Complex Number Systems questions. For both Reading 2–5 and Reading 6+ assessments, rapid guessing was less frequent with Vocabulary questions than those from other instructional areas.



Mean percentages, by instructional area, of responses that were rapid guesses during MAP Growth test events for which disengagement was moderate to high (i.e., RTE <0.90).

Differential rapid guessing often meaningfully distorted test content representation.

The study next examined how these differences in rapid guessing rates across instructional areas impacted content representation in individual test events using Cramer's V. In test events where a student's rapid guesses were concentrated in particular instructional areas, Cramer's V and content misrepresentation increased. When Cramer's V was greater than 0.40 the test's content representation was considered to be meaningfully distorted. Illustrated below is an example in which a student rapidly guessed on more than half of the questions in one instructional area and at lower rates or not at all in others, yielding V slightly above 0.40.

	Instructional area					
	1	2	3	4	5	
Rapid guessing	2	5	2	3	0	12
Solution behavior	5	4	6	4	9	28
All item responses	7	9	8	7	9	40

An example of a 40-item MAP Growth Reading test event exhibiting content imbalance resulting from differential rapid guessing. The bottom row shows the frequencies of administered items from each instructional area, while the shaded cells show the corresponding frequencies of the engaged responses. For this test event, Cramer's V=0.427, slightly higher than the 0.40 criterion for meaningful content misrepresentation.

The results showed that the percentage of test events with meaningful content misrepresentation increased with the rapid guessing rate on the assessment, was higher in reading than in mathematics, and was often quite high. On the Math 2–5 test, for example, content representation was meaningfully distorted on 1.4% of tests with low disengagement, on 6.8% of those with moderate disengagement, and on 11.2% of tests with high disengagement. On the Reading 2–5 test, the misrepresentation was considerably higher, with meaningful content distortion on 10.4% of tests with low disengagement, on 27% of those with moderate disengagement, and on 31.2% of tests with high disengagement.

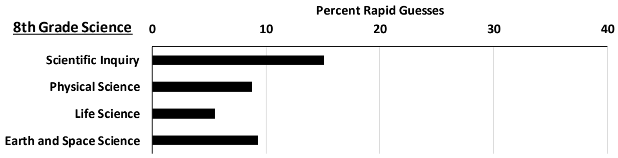
Test	Disengagement	Tests (#)	Percent of tests with meaningful content misrepresentation (V>0.40)
Math 2–5	Low	23,014	1.4%
	Moderate	2,353	6.8%
	High	331	11.2%
	Overall	25,725	2.0%
Math 6+	Low	25,063	1.6%
	Moderate	5,995	8.4%
	High	2,012	16.7%
	Overall	33,070	3.8%
Reading 2–5	Low	26,744	10.4%
	Moderate	8,778	27.0%
	High	2,285	31.2%
	Overall	37,807	15.6%
Reading 6+	Low	24,807	8.8%
	Moderate	12,155	21.6%
	High	5,667	44.0%
	Overall	42,629	14.8%

Descriptive statistics for content representation (Cramer's V) of MAP Growth test events, by level of student disengagement. Disengagement was classified as Low if 0.90 ≤ RTE < 1.0, Moderate if 0.70 ≤ RTE < .90, and High if RTE < 0.70. The Overall group was comprised all students exhibiting at least one rapid guess during their test event.

To expand upon the findings of the first study, the second study sought to identify which characteristics of test questions were correlated with rapid guessing. Previous research had found that item position, reading load, the inclusion of tables, figures, or other additional reading materials, item difficulty, and other factors all may impact rapid guessing[iv,v,vi]. Because MAP Growth assessments have a very large item pool, they are not well-suited for analysis of these item characteristics. Rather, the second study examined responses and response times and item characteristics on a fixed-form, computer-based science test administered to over 23,000 U.S. eighth-grade students in the spring of 2018. The assessment included items from four content categories. For each of the science assessment items, five characteristics were identified: item position, item difficulty, the item's depth of knowledge (DOK) category, the item's reading load, and whether or not the item contained complex or scientific artwork. Each response was classified as either a rapid guess or as an engaged response.

Much of the variation in rapid guessing rates across content areas was explained by differences in reading load of items.

While the percent of science tests with rapid guesses that had meaningful content misrepresentation was lower than on the MAP Growth reading and mathematics assessments examined in the first study, 2,379 test events had at least one rapid guess. As in the first study, rapid guessing rates differed across content categories, with a rate three times higher for Scientific Inquiry items than for Life Science items.



Mean percentages, by content category, of responses to the Science assessment that were rapid guesses during test events for which at least one rapid guess occurred.

An analysis of the item characteristics showed that the differences in rapid guessing rates were largely explained by differences in item reading load and, to a lesser extent, DOK. Other characteristics, including item position, difficulty, or inclusion of scientific artwork, did not show a significant correlation with rapid guessing rates. There were also differences in item characteristics across content categories, with higher item reading load and DOK in Scientific Inquiry items than in other categories. Analysis showed that, after adjusting for these differences, that the differences in rapid guessing rates across content categories were minor beyond what could be explained by reading load differences. Taken together, these studies show that when a question is administered to a student, relatively superficial item features, such as how much reading the item required or how mentally taxing the item would be to solve, rather than the actual content of the items, affects whether students rapidly guess or if they make effortful responses.

RECOMMENDATIONS

To improve student test engagement and test validity, test developers should work to better understand factors that tend to elicit rapid-guessing behavior and be mindful of them when designing tests.

When disengaged test taking occurs, measurement is compromised. Score accuracy and precision can be distorted, and, as this study shows, content representation can also be distorted when students rapidly-guess more in some content categories than in others. And while excluding rapid guesses from scoring can improve score accuracy and measurement precision, it does not address the problem of content non-representation. However, with better understanding of factors that may contribute to rapid-guessing behavior, test developers may be able to mitigate this problem. For example, items might be developed or chosen to avoid differential item reading load across content categories. By balancing the factors related to rapid guessing, the problem of content nonrepresentation should be mitigated, even when sizable numbers of rapid guesses are present during a test event.

Innovative approaches in computer based tests may improve content representation by re-balancing content or decreasing rapid-guessing.

Computer based tests that can identify and adapt to rapid-guessing behavior during the test event may provide additional solutions to improve content representation. For example, if the pattern of rapid guessing during the first two thirds of a test event showed meaningfully distorted content representation, the last third of the test could purposefully select items whose content would re-balance content representation. Alternately, during the test itself, the computer could notify students or proctors when rapid guessing occurs. Research has shown that such notifications and intervention by a proctor can curtail subsequent rapid guessing[vii]. By decreasing the overall number of rapid guesses that occur, the problem of content nonrepresentation should also be mitigated.

Each of these remedies are feasible only when computer based assessments that can identify rapid-guessing behavior are used. But disengaged test taking occurs on all assessments, both on computer based tests and traditional paper-and-pencil tests, suggesting that all are vulnerable to each of the psychometric costs associated with rapid guessing. Computer based tests offer an advantage, since they allow us to identify when rapid guessing behavior occurs, better understand its dynamics, and develop strategies for mitigating its impact.

[i] Wise, S.L. (2015). Effort analysis: Individual score validation of achievement test data. *Applied Measurement in Education*, 28, 237-252.

[ii] Guo, H., Rios, J. A., Haberman, S., Liu, O. L., Wang, J., & Paek, I. (2016). A new procedure for detection of students' rapid guessing responses using response time. *Applied Measurement in Education*, 29, 173-183.

[iii] Wise, S. L., & DeMars, C. E. (2006).An application of item response time: The effort-moderated IRT model. *Journal of Educational Measurement*, 43, 19-38.

[iv] Setzer, J. C., Wise, S. L., van den Heuvel, J. R., & Ling, G. (2013).An investigation of test-taking effort on a large-scale assessment.Applied Measurement in Education, 26, 34-49.

[v] Wise, S. L., Pastor, D. A., & Kong, X. J. (2009).Correlates of rapid-guessing behavior in low-stakes testing: Implications for test development and measurement practice. *Applied Measurement in Education*, 22, 185-205.

[vi] Goldhammer, F., Martens, T., & Lüdtke, O. (2017). Conditioning factors of test-taking engagement in PIAAC: An exploratory IRT modelling approach considering person and item characteristics. *Large-scale Assessments in Education*, 5:18. doi: 10.1186/s40536-017-0051-9

[vii] Wise, S., Kuhfeld, M., & Soland, J. (2019). The effects of effort monitoring with proctor notification on test-taking engagement, test performance, and validity. *Applied Measurement in Education 32* (2), 183-192.

This brief describes research documented in:

Wise, S. (2020). The impact of test-taking disengagement on item content representation. *Applied Measurement in Education*, 3(2), 83-94. https://doi.org/10.1080/08957347.2020.1732386

Suggested citation:

Wise, S. (2020). The impact of test-taking disengagement on item content representation. (The Collaborative for Student Growth at NWEA Research Brief).

ABOUT THE AUTHOR

Dr. Steven Wise is a Senior Research Fellow at the Collaborative for Student Growth at NWEA, serves on the editorial board of several academic journals, and has provided psychometric consultation to organizations including state departments of education in Maryland, Virginia, and Nebraska; the National Assessment Governing Board; and the GED Testing Service. Wise's current research focus is primarily on practical methods for effectively dealing with the measurement problems posed by low examinee engagement on achievement tests. He holds a PhD in Educational Psychology, Measurement and Statistics from the University of Illinois at Urbana-Champaign.

ABOUT THE
COLLABORATIVE FOR STUDENT GROWTH

The Collaborative for Student Growth at NWEA is devoted to transforming education research through advancements in assessment, growth measurement, and the availability of longitudinal data. The work of our researchers spans a range of educational measurement and policy issues including achievement gaps, assessment engagement, social-emotional learning, and innovations in how we measure student learning. Core to our mission is partnering with researchers from universities, think tanks, grant-funding agencies, and other stakeholders to expand the insights drawn from our student growth database—which is one of the most extensive in the world.

NWEA is a not-for-profit organization that supports students and educators worldwide by providing assessment solutions, insightful reports, professional learning offerings, and research services. Visit NWEA.org to find out how NWEA can partner with you to help all kids learn.