2024

Predictive Validity and Cutoff Scores of Acadience Reading Assessments for Third Grade RISE Reading Assessments (#USBE240083IAA)

Prepared by the
Utah Education Policy Center
August 2024

THE UNIVERSITY OF UTAH
COLLEGE OF EDUCATION
UTAH EDUCATION POLICY CENTER

THE UNIVERSITY OF UTAH
COLLEGE OF EDUCATION

UTAH EDUCATION
POLICY CENTER

Bridging Research, Policy, & Practice

The Utah Education Policy Center (UEPC) is an independent, non-partisan, not-for-profit research-based center at the University of Utah founded in the Department of Educational Leadership and Policy in 1990 and administered through the College of Education since 2007. The UEPC mission is to bridge research, policy, and practice by conducting rigorous and comprehensive research and evaluations and providing expert and research-informed technical assistance and professional learning. We empower educators, policymakers, and leaders to make research actionable and impactful to transform education across early childhood education, K-12 schools, and higher education. We are committed to supporting the understanding of whether educational policies, programs, and practices are being implemented as intended, whether they are effective and impactful, and how they may be improved, scaled-up, and become sustainable.

Please visit our website for more information about the UEPC:
http://uepc.utah.edu

Andrea K. Rorrer, Ph.D., Director
andrea.rorrer@utah.edu

Cori Groth, Ph.D., Associate Director
cori.groth@utah.edu

Ellen Altermatt, Ph.D., Assistant Director for Research and Evaluation
ellen.altermatt@utah.edu

T. W. Altermatt, Ph.D., Assistant Director and Lead Data Scientist
bill.altermatt@utah.edu

Acknowledgments

The Utah Education Policy Center extends our gratitude to the administrative, research, assessment, and data specialists at USBE who worked quickly to supply us with the data and documentation needed for this study: Darin Nielsen, Aaron Brough, Maureen Rushing, Kim Rathke, and Cydnee McCarty.

Study Team

James Gallyer, University of Utah, Utah Education Policy Center
Bill Altermatt, University of Utah, Utah Education Policy Center
Andrea K. Rorrer, University of Utah, Utah Education Policy Center

Citation: Gallyer, A. J., Altermatt, T. W., & Rorrer, A. K., (2024). *Predictive Validity and Cutoff Scores of Acadience Reading Assessments for Third Grade RISE Reading Assessments (#USBE240083IAA).* Salt Lake City, UT: Utah Education Policy Center.

Table of Contents

List of Tables

List of Figures

Executive Summary

Study Overview

Utah Senate Bill 127 (2022) established a goal to have 70% of third grade students reading on grade level by June 2027. To measure progress toward this goal, the state uses a state-wide reading test called RISE (Readiness Improvement Success Empowerment). In addition to the RISE assessment, students in Utah also participate in the Acadience Reading assessment, which is administered to students in grades K-6 (with grades 4-6 being optional) as part of the state's benchmarking reading assessment (see Rule 277-406). Along with an overall score, the Acadience assessment has several subscores:

1. Oral reading fluency (i.e., "Fluency", the number of words correct)
2. Accuracy (during the Fluency task, the number of words correct divided by total words)
3. Reading comprehension by Retelling the passage that was just read (i.e., "Retell")
4. Performance on the cloze procedure (Acadience's MAZE), designed for students beginning in third grade, where students read a sentence in which some words are replaced with a box of three or more words and students must select the best word based on context, syntax, and background knowledge (Taylor, 1953).

The Utah State Board of Education (USBE) partnered with the Utah Education Policy Center (UEPC) to evaluate the predictive validity of the Acadience Reading assessment's four subscores (listed above), measured in first through third grade, in reference to two subscores on Utah's third grade RISE Reading assessment: Informational Text and Literature. In addition, the UEPC evaluated the Reading On Grade Level (ROGL) cutoff for the Acadience Reading comprehensive scores in first through third grades with regard to the benchmark for proficiency on the RISE English Language Arts Assessment in third grade. This report continues a series of research collaborations between the UEPC and USBE[1].

Research Questions

This study had two primary objectives: First, to understand the relationship between Acadience Reading subscores (in first, second, and third grade) and two RISE Reading subscores (Literature and Informational Text) given during third grade. Second, to evaluate the Reading On Grade Level (ROGL) cut scores for the Acadience Reading composite score in first to third grades to determine whether they are optimal for predicting third grade RISE ELA (English Language Arts) proficiency.

Research Questions on Acadience and RISE Subscores
1. What is the predictive validity of each of the Acadience subtests on their own to predict third grade RISE Reading Literature and Reading Informational Text subscores?

[1] To date, this collaboration has explored issues regarding educator preparation, pipeline, and working conditions (Acree et al., 2023; Auletto et al., 2020; Ni et al., 2017a; Ni et al., 2017b; Ni & Rorrer, 2018; Rorrer et al., 2020) as well as the characteristics of schools where students with disabilities have both higher achievement and higher levels of inclusion (Acree et al., 2023).

2. What is the predictive validity of combining the Fluency and MAZE subtests to predict RISE Reading Literature and Reading Informational Text subscores?
3. What is the predictive validity of combining the Retell and MAZE subtests to predict RISE Reading Literature and Reading Informational Text subscores?
4. Does the predictive validity of the Acadience subtests, and the combinations of subscores of interest, differ by gender, race or ethnicity, for students with disabilities, students learning English, or for students who are economically disadvantaged?

Research Questions on Reading on Grade Level Cut Scores

1. How well do the current Acadience cut scores for first, second, and third grades perform?
2. What are the optimal cut scores?
3. How would changing the cut scores impact the number of students predicted to fail, pass, and so on?

Report Organization

The report is divided into five sections. First, we describe each of the Acadience Reading subtests and the RISE ELA assessment. Then, we discuss the predictive validity of each Acadience subscore (Fluency, Accuracy, Retell, and MAZE) as well as the two combinations of interest (MAZE + Fluency and MAZE + Retell) for the two RISE reading comprehension subscores (i.e., Reading Literature and Reading Informational Text). The third section discusses whether these predictions differ across various demographic groups, including: gender, students receiving special education services, students learning English, students who are economically disadvantaged, and by race and ethnicity. The fourth section examines the USBE's current Acadience composite score Reading on Grade Level cutoffs and investigates their performance in predicting RISE Reading on Grade Level compared to three model-based optimal cutoffs. The final section discusses the implications of the results. While detailed findings are presented in each report subsection, we summarize the primary findings here.

Key Findings

Predictive Validity. For our first research question regarding the predictive validity of the Acadience subscores in predicting third grade RISE Reading Literature and Reading Informational Text subscores, we consistently found that all Acadience subscores were significantly positively related to both third grade RISE subscores. Across Acadience subscores in first, second, and third grade, the average percentage of variance in RISE subscores explained was 23% (ranging from 10% to 35%). However, the Acadience subscores were not equal in their predictive ability. **Fluency was consistently the strongest single predictor of both RISE reading subscores (Informational Text and Literature) in all three grades.** Predictive power was marginally improved if MAZE was used together with Fluency to predict RISE scores (percent of variance explained improved between 1.87 and 2.24 percentage points). This suggests that, of the Acadience subscores considered in this report, the most predictive was the combination of Fluency and MAZE in third grade (the first grade where MAZE is available), with Fluency alone being the best predictor in first and second grades.

Differences across Student Groups. Regarding differences in predictive ability across student demographics, we examined whether the predictive ability of each of the Acadience subscores differed by gender, race and ethnicity, for students with disabilities (defined as students receiving special education services), students learning English, or students who are economically disadvantaged (defined as students who qualify for free or reduced-price lunch) for every grade. Because results were widely the same regardless of which subscores we examined, we focused on reporting results for Fluency (the strongest predictor) and, in third grade, the combination of MAZE with Fluency or Retell. Predictive validity showed few differences across gender, economic disadvantage, and disability status. **However, five groups had predicted RISE scores that were more than 0.1 standard deviations higher than their actual RISE scores (i.e., they were consistently overestimated): English Language Learners, students who are Black or African American (hereafter "Black students"), Hispanic students, Native American students, and students who are Pacific Islander or Native Hawaiian (hereafter "Pacific Islander students").** This pattern is concerning because it suggests that forecasts made based on Acadience subscores for these students will tend to overestimate their actual RISE Reading scores, putting them at greater risk than their reference groups of being classified as proficient based on an Acadience test and then later not passing the RISE test (i.e., of being "False Passes" in the terms introduced in the next section).

Reading on Grade Level Cutoffs. We evaluated the Acadience Reading composite score cutoffs for first, second, and third grade in relation to the two ways that forecasts made from the Acadience to RISE could be errors: 1) the "False Pass" (a student's Acadience score is above the cutoff but they later fail the RISE), or 2) the "False Fail" (a student's Acadience score is below the cutoff but they later pass the RISE). Typically, both of these errors are simply added together when considering a cutoff score's "Accuracy": the percentage of correct forecasts divided by the total number of forecasts. However, we discuss how the costs of a False Pass (failing to provide interventions to students who need them) may be higher than the costs of a False Fail (providing interventions to students who may not need them). We present the Accuracy, False Fail Rate, and False Pass Rate of each of the current Acadience Reading composite cutoffs and compare those to three different model-based definitions of optimal cutoffs: (1) cut scores maximizing Accuracy, (2) cut scores weighing False Passes as twice as costly as False Fails, and (3) cut scores weighing False Fails as twice as costly as False Passes. Overall, the current cut scores outlined in Appendix G of the Utah Accountability Technical Manual for first through third grade perform relatively well at predicting which students will be categorized as Reading on Grade Level at third grade from the RISE ELA assessment, with all three showing Accuracy above 78%. However, consideration should be given not only to overall accuracy but also to the relative costs of False Passes and False Fails.

1 | Introduction to Reading Measures

This study uses both the Acadience Reading assessment and the RISE Reading assessment. Here we provide an overview of the measures of each assessment.

As outlined by Good III and colleagues (2011), the **Acadience Reading** assessment is designed to measure early literacy and reading ability for students in kindergarten through sixth grade. While it consists of several tasks, one of the key tasks involves students reading passages and then demonstrating their reading comprehension by retelling what the passage was about to a teacher or teacher's designee who serves as the test administrator. Specifically, for each passage, the test administrator says to the student, "I would like you to read a story to me. Please do your best reading. If you do not know a word, I will read the word for you. Keep reading until I say 'stop.' Be ready to tell me all about the story when you finish" (Good III et al., 2011). After the student begins, they have one minute to read the passage. While reading, the test administrator tracks which words the student either says incorrectly or does not know (i.e., the student doesn't read the word within 3 seconds). After reading the passage, the test administrator says, "Now tell me as much as you can about the story you just read." During this portion, the test administrator marks the number of words related to the story the student says within a minute.

Acadience reading results are reported in a composite score and several subscores: (1) **Fluency** (i.e., the number of words correctly read during the oral reading fluency task), (2) **Accuracy** (i.e., the number of words read correctly divided by the total number of words read during the oral reading fluency task), and (3) **Retell** (i.e., number of words associated with the passage that the student says during the retelling portion of the oral reading fluency task). In third grade, in addition to Fluency, Accuracy, and Retell, another subscore is used to assess reading comprehension: **MAZE**. In this task, which is based on the "cloze" procedure (Taylor, 1953), students are asked to read a passage silently. Within the passage are blank spaces containing three words. The student's task is to pick the word that best fits the space given the context of the sentence/passage. From this task, a MAZE adjusted score is calculated that compensates for randomly guessing the correct choice. From these measures, a composite Acadience Reading score is also calculated, which gives an overall measure of reading ability.

The **RISE (Readiness Improvement Success Empowerment)** ELA test is a computer adaptive assessment system used to assess English Language Arts criteria from third through eighth grade (*Assessments*, 2024). Like the Acadience reading measure, RISE ELA consists of a composite score and several subscores. The subscores of interest in this report reflect reading comprehension specifically and are the *Reading Literature* and *Reading Informational Text* subscores. As their name suggests, Reading Literature involves passages of fiction or literature, while Reading Informational Text involves nonfiction passages.

2 | Acadience Subscores' Predictive Validity

Background

After merging the RISE subscore and Acadience subscore data, resolving duplicates, and removing anomalous scores[2], the data consisted of 161,958 unique students who attended third grade and took the RISE ELA test during the school years 2018-2019 to 2022-2023, excluding the 2019-2020 school year due to the pandemic. To determine whether each Acadience subscore (i.e., Fluency, Accuracy, Retell, and MAZE) was a predictor of each RISE subscore (i.e., Reading Literature and Reading Informational Text), we constructed a series of multi-level models for each grade level when the Acadience test was taken (first, second, or third). Multi-level models are ideal when the data are nested, as is the case here with students nested within both the school in which they took the RISE test in third grade, and the school in which they took the Acadience test in first, second, and third grade. Accounting for the nested data structure is important because students who are in the same school are more likely to have scores that are similar to each other, and without a multilevel model this clustering can artificially inflate the likelihood of obtaining statistical significance. In these models, there is one outcome variable – either third grade RISE Reading Literature or third grade RISE Reading Informational Text – and one or more predictor variables: the subscores on the Acadience reading tests. For each predictor (except MAZE, which is only administered in third grade), there were six models corresponding to the three grade levels and two RISE subscores. For example, for the Fluency Acadience subscore alone (i.e., not combined with MAZE), the six models were (1) first grade Fluency predicting third grade Reading Literature, (2) first grade Fluency predicting third grade Reading Informational Text, (3) second grade Fluency predicting third grade Reading Literature, (4) second grade Fluency predicting third grade Reading Informational Text, (5) third grade Fluency predicting third grade Reading Literature, and (6) third grade Fluency predicting third grade Reading Informational Text.

Evaluating Predictive Validity

To evaluate how effective each Acadience subscore was at predicting RISE Reading subscores, we used two metrics: (1) the proportion of the variance in the RISE subscore explained by the Acadience subscore, called R^2, and (2) the width of a 66% prediction interval. The R^2 value ranges from 0% to 100%, corresponding to the percentage of the variance in the RISE subscore that the Acadience subscore accounts for. There is no universal rule for how much variance accounted for is considered "good" or "bad". However, given the nature of the data, we expected R^2 values to hover around 20%, with lower values when the prediction was from first grade and higher values when the prediction was

[2] RISE reading subscore data received from USBE displayed a largely normal distribution, except the lowest score (110 for both subscores) was unusually frequent, being nearly as common as the average score. Inquiries were made about whether these 110 scores had a special interpretation (such as indicating a blank test). USBE indicated there were no special cases or interpretations. Given that the rest of the scores were distributed normally, the probability that these scores reflected valid reading performance at the level indicated was vanishingly small. Thus, all participants who scored 110 on either RISE reading subscore were filtered out, as including them would have introduced noise and skewed results.

from third grade because of the tendency for prediction to be more accurate when it is made over shorter time periods. The 66% prediction interval is the range of scores around a predicted score in which 66% of actual third grade RISE scores will tend to be found[3]. Conversely, 34% of actual third grade scores will tend to be found outside this range. We selected 66% because of Mastrandrea et al.'s (2010) recommendation of 66% as a reasonable threshold for an outcome to be considered "likely." For example, consider a student with a first grade Fluency score of 67. According to our statistical models, the predicted third grade RISE Reading Informational Text Score is 332. The 66% prediction interval around 332 is ± 65.03, meaning that, on average, 66% of the actual RISE scores for students with a first grade Fluency score of 67 will tend to fall within the range 266.97 to 397.03. The key to our results is this: the better the predictor is, the smaller the width of the 66% prediction interval will be. For example, a 66% prediction interval width of ± 73.26 would indicate that if you know the Acadience subscore, the actual RISE subscore would be within 73.26 points of the predicted score 66% of the time. But a 66% prediction interval width of ± 56.11 would indicate that if you know the Acadience subscore, the actual RISE subscore would be within 56.11 points of the predicted score 66% of the time, a more precise estimate. Like the R^2 value, there is no rule for what constitutes a "good" or "bad" 66% prediction interval width. Instead, the 66% prediction interval will be useful for comparing the predictive ability of the different Acadience subscores and providing a concrete estimate of how good the prediction is in the units of the RISE test. For context, the narrowest 66% prediction interval obtained in the present study is ± 56.11, which covers 27% of the range of possible RISE Reading Literature scores, while the widest obtained 66% prediction interval of ± 73.26 covers 35% of the range of possible RISE Reading Informational Text scores. As expected, these prediction intervals are wide and indicate that making predictions for individual students involves a great deal of uncertainty. For further technical details about our models, see Appendix A.

[3] The 66% prediction interval is intended to convey the *precision* of the estimates generated by a model, in the units of the original outcome variable (RISE subscores, in this case). In contrast to a confidence interval, which describes the range of values within which a population parameter is likely to be found, a prediction interval is focused on the range of values within which actual individual outcomes are likely to be found. In this way, the prediction interval could be considered more concrete / less abstract than a confidence interval.

Table 1. 66% Prediction Interval Widths and R^2 Values for Each Predictor

Grade	Predictor	RISE Subscore	66% Prediction Interval Width (+/-)	R^2
First Grade	Fluency	Literature	59.83	29.34%
		Informational Text	65.03	26.93%
	Accuracy	Literature	63.52	20.05%
		Informational Text	69.57	15.91%
	Retell	Literature	64.24	16.53%
		Informational Text	69.61	14.61%
Second Grade	Fluency	Literature	57.41	34.89%
		Informational Text	65.30	29.16%
	Accuracy	Literature	65.09	15.50%
		Informational Text	73.04	10.54%
	Retell	Literature	63.10	19.98%
		Informational Text	70.68	16.04%
Third Grade	Fluency	Literature	57.60	35.35%
		Informational Text	64.61	30.02%
	Accuracy	Literature	66.46	13.64%
		Informational Text	73.26	9.51%
	Retell	Literature	63.84	20.47%
		Informational Text	70.38	16.59%
	MAZE	Literature	59.13	30.39%
		Informational Text	66.10	26.91%
	MAZE + Fluency	Literature	56.11	37.48%
		Informational Text	63.29	32.26%
	MAZE + Retell	Literature	57.03	34.78%
		Informational Text	64.07	30.08%

Predictive Validity of Individual Acadience Subscores

See Table 1 for the full results of the predictive validity of each of the Acadience Subscores. R^2 values ranged from 9.51% to 37.48%, and the 66% prediction interval widths ranged from ±56.11 to ±73.26. From first through third grade, Fluency emerged as the strongest single predictor of both Reading Literature and Reading Informational Text subscores, with R^2s ranging from 26.93% to 35.35% and 66% prediction interval widths ranging from ±57.41 to ±65.30. In third grade, MAZE emerged as the second best single predictor, with R^2s ranging from 26.91% to 30.39% and 66% prediction interval widths ranging from ±59.13 to ±66.10. The least predictive subscore was sometimes Retell and sometimes Accuracy. In first grade, Accuracy marginally outperformed Retell, but in second and third grade, Retell outperformed Accuracy.

Interestingly, among the three Acadience subscores that appeared in all three grades, Fluency was the only one whose predictive ability consistently increased from first to third grade. For example, for the Reading Literature outcome, R^2 increased from 29.34% to 35.35%, and 66% prediction interval widths decreased from ±59.83 to ±57.60. This further bolsters the conclusion that Fluency is the best predictor of third grade RISE reading subscores, given that if a measure is related to future reading ability, that measure should become more predictive as less time passes between when the measure is taken and when the future reading ability is measured. Together, these results indicate that a) all of the Acadience subscores were significantly positively related to both RISE Reading subscores, b) the subscores varied considerably in how well they predicted RISE Reading subscores, and c) Fluency was the best single predictor of RISE Reading subscores.

Predictive Validity of MAZE Combined with Other Subscores

Students encounter the MAZE test for the first time in third grade, so any consideration of MAZE as a predictor of RISE should be made with the understanding that any predictions involving MAZE are predictions from third grade to third grade, rather than across grade levels as was possible for the other Acadience subscores. In third grade, we examined the predictive validity of combining MAZE with Fluency, and MAZE with Retell. The results are presented at the bottom of Table 1.

MAZE and Retell

The results in Table 1 show that third grade MAZE alone is a much better predictor than third grade Retell alone of third grade RISE scores. Whereas third grade Retell has an R^2 of 16.59% for Informational Text and 20.47% for Literature, third grade MAZE has an R^2 of 26.91% for Informational Text and 30.39% for Literature. As a result, the combination of MAZE and Retell is a dramatic improvement over Retell alone, increasing R^2 by about 15 percentage points and shrinking the prediction interval by about 7 points. However, the combination of MAZE and Retell is only a small improvement over MAZE alone, increasing R^2 by only about 4 percentage points and shrinking the prediction interval by about 2 points.

MAZE and Fluency

Table 1 shows that third grade Fluency alone is a better predictor than third grade MAZE alone of third grade RISE reading scores. The R^2 for Fluency is about 5 percentage points higher and the prediction intervals 2 points narrower than they are for MAZE. As a result, predictions based on a combination of Fluency and MAZE show a more considerable improvement over MAZE alone than they do over Fluency alone. The combination of Fluency and MAZE improves R^2 over MAZE alone by about 7 percentage points but improves R^2 over Fluency alone by only 2 percentage points. The combination of Fluency and MAZE narrows prediction intervals over MAZE alone by 3 points but narrows prediction intervals over Fluency alone by only 1 point.

Conclusions about Combining MAZE with Fluency or Retell

These results suggest that predictions based on a combination of subscores are more accurate than predictions based on individual subscores. However, not all combinations were equal: MAZE and Retell together were similar to the predictive ability of Fluency alone, and adding MAZE to Fluency only marginally improved on the predictive ability of Fluency alone. It is important to note that the use of multiple subscores in making predictions for individual students does not involve simply averaging those subscores together. Using multiple Acadience subscores in practice would require entering each subscore into a formula derived from a statistical model that would then produce a predicted RISE subscore.

3 | Acadience Subscores' Predictive Validity by Demographics

Background

In this section, we examine whether the ability of Acadience subscores to predict RISE subscores varies across student demographic groups. The central question is whether the predictive validity of the Acadience measures is lower for some student groups than for others. The demographic variables examined were gender, race and ethnicity, disability status as indicated by the receipt of special education services, English Language Learner status, and economic disadvantage as indicated by eligibility for free or reduced-price lunch.

We consider four ways to evaluate how predictive validity varies across groups. The first way starts with the models from section 2 of this report, which predict RISE Reading subscores from Acadience subscores. Those models produce predicted RISE Reading subscores for each student, and the difference between the actual RISE subscore and the predicted RISE subscore for each student is called a *residual*. When the predicted score is greater than the actual score, the student is being overestimated and the residual will be negative. When the predicted score is less than the actual score, the student is being underestimated and the residual will be positive. These residuals can be analyzed for each subgroup of students to answer two questions.

First, the mean value of the residuals for a subgroup can indicate whether that subgroup is being consistently over- or under-estimated. If most members of a subgroup have predicted values that are greater than their actual values, their mean residual value will be negative. By dividing a mean residual by the standard deviation of the RISE subscore, it is placed on a scale where -1 indicates an overestimate of one standard deviation. We consider any subgroups with a mean residual greater than 0.1 standard deviations to be underestimated and any subgroups with a mean residual less than –0.1 standard deviations to be overestimated[4].

The overestimation scenario is described in Figure 1, which displays three scenarios based on simulated (artificial) data. Plot **a** shows best-fitting regression lines for two simulated subgroups. Assume that the subgroup represented by the black line is a majority of students, such as students who are not learning English, while the subgroup represented by the red line is a minority of students, such as students who are learning English. Because of the different sizes of the subgroups, a model that is based on data from all students will be more consistent with the black line than the red line. Students who are in the subgroup represented by the black line will tend to be well represented by that model. However, students in the subgroup represented by the red line will tend to be overestimated. Their actual RISE scores will tend to be lower than the scores predicted by a single model, as suggested by the fact that the red line is below the black line. For example, a model based

[4] Our threshold of 0.1 standard deviations represents a difference of around 8 points on the scale of the RISE Reading Informational Text or Reading Literature subscores. We consider this to be a “small” but not "trivial” magnitude.

mostly on the black line will predict that a student with a Fluency score of 40 should receive a RISE Literature score of 110[5]. If that prediction is made for a student represented by the red line, it will be 20 points too high because the red line is approximately 20 points below the black line.

Overestimation is a problem if remedial assistance is triggered by a low score on an Acadience test that is designed to forecast performance on the RISE test. Students whose RISE Reading performance is systematically overestimated may be denied interventions they need to improve reading performance[6]. If a subgroup is being systematically *under*estimated, then these students may be incorrectly identified as needing a reading intervention.

A second way that residuals can be used is by computing R^2 for each subgroup. This R^2 value indicates how well a model based on all students predicts a particular subgroup of students. Models whose predictions are farther away from the actual values, whether that is because of a systematic over- or underestimation or simply because of more noise in the prediction, have lower R^2 values. If the R^2 for subgroup A is more than 10 percentage points lower than the R^2 value for subgroup B, we consider the predictive validity of the model to be worse for subgroup A than for subgroup B.

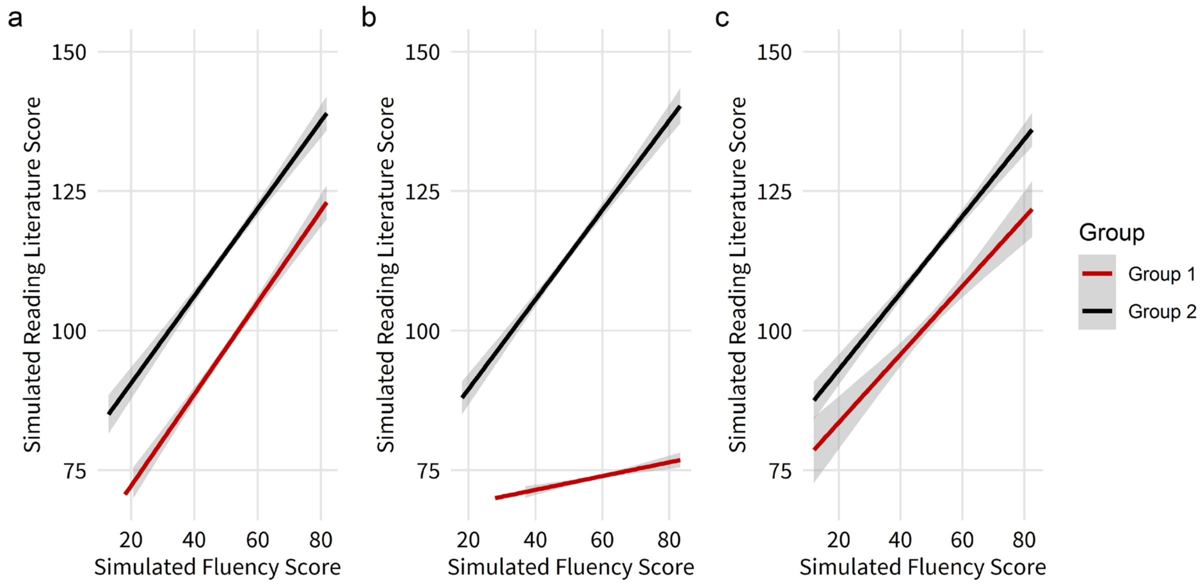
A third way we can evaluate predictive validity is by a difference in the **slopes** of the best-fitting regression lines that describe the relationship between an Acadience predictor and a RISE outcome. For example, in the simulated data in Figure 1 plot **b**, the slope of the line for Group 2 (the black line) is steeper than the slope of the line for Group 1 (the red line). This indicates that Fluency is more strongly related to Reading Literature for Group 2 than Group 1, because as Fluency increases for Group 2, the predicted Reading Literature Score does not increase as much as it does for Group 1. A difference in slopes is important because it indicates that the degree of over- or under-estimation changes depending on the predictor (e.g., Fluency). There could be a very small amount of overestimation at low levels of the predictor but large amounts of overestimation at high levels of the predictor.

Fourth, we can consider the R^2 we would obtain if we fit models separately for each subgroup, rather than fitting one model for all students and then examining the residuals separately by subgroup. R^2 from separate models for each subgroup provides a best-case scenario for how well RISE Reading subscores *could* be explained by a predictor. If the R^2 values from separate models are more than 10 percentage points higher than the R^2 subgroup values from a single model, it would be evidence that there is a systematic relationship between the Acadience predictor variable and the RISE outcome variable for the subgroup that is not well described by the model for the whole group (e.g., the two subgroups could be better described by lines with different slopes, as discussed above). In Figure 1 plot **c**, the black line shows a higher R^2 value than the red line because of the larger amount of error around the red line (indicated by the gray shading around the lines). Figure 1 plot **c** suggests that the poorer fit of the model for the red line group is not simply due to overestimation or underestimation

[5] The data used in Figure 1 are simulated data designed only to illustrate the possible ways that predictive validity could vary across groups. The relationship between a Fluency score of 40 and a RISE score of 110 is purely fictional and should not be interpreted as an actual result.
[6] This assumes that interventions are triggered by scoring below proficiency. The degree to which this occurs in practice is an important question and a possible topic for future research.

(which does not contribute to R^2 when the two subgroups are modeled separately) but rather due to more noise in the prediction. This noise could be due to greater error in the measurement of the predictor or outcome or due to more influential unmeasured variables for the red line group. Either way, the predictions for the group represented by the red line in Figure 1 plot **c** are less accurate than the predictions for the group represented by the black line.

Figure 1. Simulated data showing how slopes, mean levels, and R^2 can differ between groups



We used the same multi-level modeling approach that we used for the predictive validity analyses in the previous section to answer two questions using residuals: 1) over- or under-estimation of subgroups and 2) R^2 by subgroup. To answer the third question (differences in slopes), we added one demographic variable (e.g., gender) at a time as both a main effect and in an interaction with a given predictor to test for differences in slopes[7]. We then analyzed each group (e.g., male or female) separately to obtain R^2 values specific to each group[8]. Due to the combination of predictors, grades, outcomes, and demographic groups, there were 100 models in total. When there are so many models, there is a higher probability that some statistically significant results are "spurious," that is, occurring just by chance because of the sheer number of analyses. To mitigate (but not eliminate) this problem, in addition to requiring a threshold of statistical significance of $p < .05$, we also implemented requirements that the magnitude of an effect exceed a threshold of "practical significance" to be

[7] For example, to examine whether the slopes describing the relationship between first-grade Fluency and RISE Reading Informational Text differed between English Language Learners (ELL) and students not learning English, we fit a model that predicts RISE Reading Informational Text from ELL (a binary variable), Fluency (a continuous variable), and an interaction term that is equal to ELL multiplied by Fluency: RISE ~ ELL + Fluency + (ELL * Fluency). The interaction term captures the degree to which the slopes for the two groups are different.

[8] Note that this indicates the percentage of variance explained when a separate predictive model is developed for each student group. This reflects the degree to which predictors and outcomes are related for a given group under somewhat ideal conditions given that in practice, a single predictive model will be used for different groups.

noteworthy: differences in the mean residuals of subgroups greater than 0.1 standard deviations (using the standard deviation of the outcome variable, not the residuals), R^2 differences greater than 10 percentage points, and differences in slopes greater than 0.1 standard deviations.[9] Last, for the two combination predictor models (i.e., MAZE with Fluency and MAZE with Retell), we excluded the analysis of slopes, as there is not a single interaction term that can illustrate differences in slopes for a multiple predictor model. To see more technical details for these analyses, see Appendix B.

When we examined the results, we determined that the pattern of results did not meaningfully differ among Fluency, Accuracy, and Retell. Thus, to reduce complexity and maintain readability in the main body of this report, the focus will be only on Fluency (since it was the strongest single predictor) and the two combination predictor models (MAZE + Fluency and MAZE + Retell). For a table of all the results, including results for Accuracy and Retell, see Appendix C.

Differences in Predictive Validity of Acadience Subscores

Gender

Across all grades, there were almost no practically significant differences between boys and girls in the predictive validity of Acadience subscores. The R^2 values for boys and girls when both were fit by a single model were within two percentage points of one another and were within one percentage point of one another when they were fit by separate models. This indicates very little difference in the overall accuracy of predictions for boys and girls. None of the slopes of the best-fitting regression lines describing the relationship between Fluency and RISE subscores differed by more than .01 standard deviations. The mean residuals for boys and girls were very similar when predicting RISE Reading Informational Text (within .02 points of zero), indicating no overestimation or underestimation for that RISE subscore. The only practically significant difference observed between boys and girls was between the mean residuals for the RISE Reading *Literature* subscore in first grade, when boys were overestimated by 0.11 standard deviations while girls were underestimated by an equal amount. This is equivalent to an overestimation for boys of approximately 8.3 points on the RISE Reading Literature subscore. In second and third grade, there was a similar pattern, but it did not exceed the 0.1 standard deviation threshold for practical significance: boys were overestimated by between .08 and .09 standard deviations, while girls tended to be underestimated by an equal amount. Not surprisingly, the results for MAZE + Fluency and for MAZE + Retell in third grade were the same: boys were overestimated and girls underestimated by .09 standard deviations but only for Reading Literature, not Reading Informational Text, while R^2 values for the two groups were within .01 of each other. Together, these results indicate that there is little difference by gender in the predictive validity of the Acadience subscores to predict RISE Reading subscores.

[9] There is no official standard for how small an effect must be to be considered “trivially small” because it depends on the context. The idea of “practical” significance goes beyond statistical significance (which indicates that a pattern is more prominent than would be expected by chance) to also require that the magnitude of that pattern be strong enough to be meaningful in the real world. In this case, our thresholds are designed to detect effects that are small but not trivial.

Students Who are Economically Disadvantaged

For our analyses, students were categorized as “economically disadvantaged” if they qualified for free or reduced-price lunch (Utah Code 35A-15-102). The mean of the residuals from the Fluency, MAZE + Fluency, and MAZE + Retell models for economically disadvantaged students were close to but did not cross, the threshold for practical significance in first, second, and third grade for both RISE Reading subscores. Overestimations for this subgroup ranged from 0.08 to 0.09 standard deviations (equivalent to an overestimation of only 6 to 8 points on the RISE Reading subscore scales), which did not meet our threshold for practical significance. The R^2 values for each subgroup were within four percentage points of one another both when the subgroups were fit by a single model and when they were fit by separate models. None of the differences between students who were and were not economically disadvantaged in the slopes of the best-fitting regression lines describing the relationship between Fluency and RISE subscores were practically significant. These results indicate that there are no practically significant differences between economically disadvantaged students and students who are not economically disadvantaged in the predictive validity of the Acadience subscores to predict RISE Reading subscores.

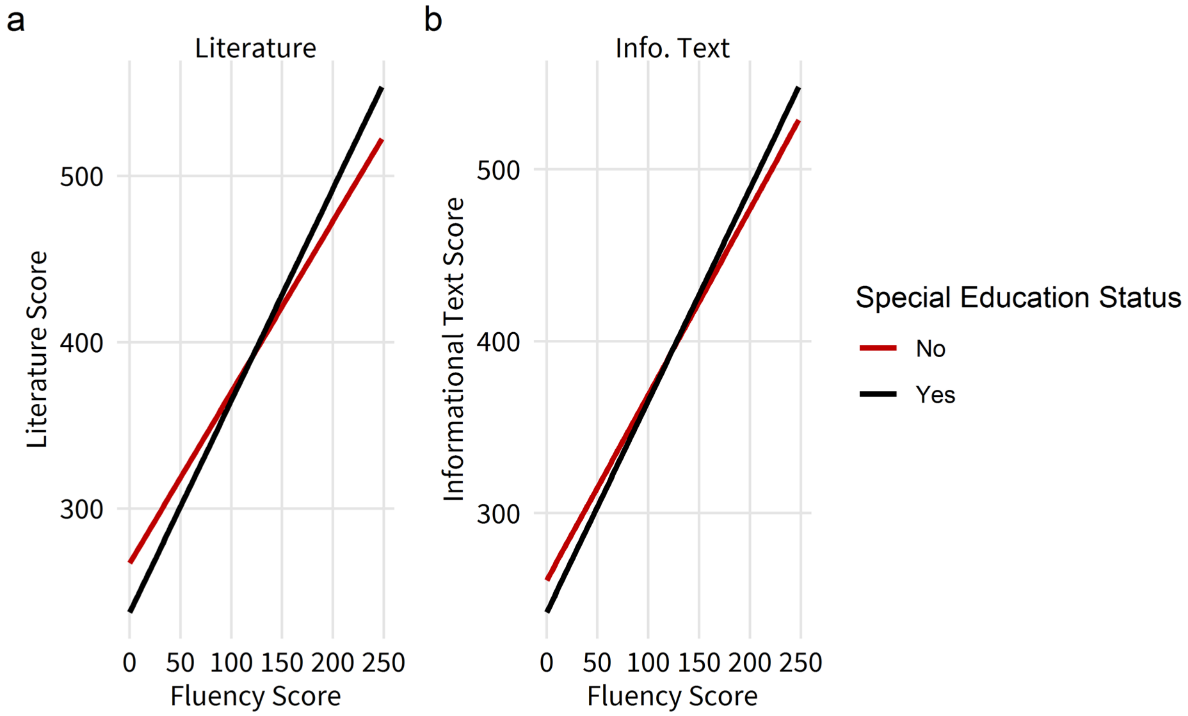
Students with Disabilities

Students with disabilities (SWD, defined by students receiving special education services) showed some signs of being overestimated that were most pronounced in first grade. In first grade, a model based on Fluency subscores for all students overestimated the RISE Reading Informational Text subscores of SWD by 0.11 standard deviations (9 points on that RISE subscore) and overestimated their Reading Literature subscores by 0.18 standard deviations (14 points on that RISE subscore). For Reading Informational Text, the overestimation was below the level of practical significance at second and third grades (0.04 standard deviations), but for Reading Literature, it was right at the threshold of practical significance: 0.10 standard deviations in grades 2 and 3. The combined model of MAZE + Fluency also showed overprediction of students with disabilities at the threshold for practical significance (0.10 standard deviations), but only for Reading Literature, not Reading Informational Text. The overprediction was slightly greater for the MAZE + Retell combination model (0.14 standard deviations), but again only for Reading Literature.

When Fluency was used as the single predictor for RISE Reading subscores, R^2 was slightly higher for SWD than for students without disabilities, but the difference did not exceed our ten percentage-point threshold for practical significance. This was true both when R^2 was calculated for subgroups from a single model and when R^2 was calculated based on separate models for each subgroup. A practically significant difference in R^2 values between SWD and students without disabilities was obtained when RISE Reading subscores were predicted from combinations of MAZE and either Retell or Fluency, but only when the RISE Reading subtest was Literature, not when it was Informational Text. The R^2 value for SWD for MAZE and Retell predicting Reading Literature in third grade was 42%, while for students without disabilities it was 31%, a difference of eleven percentage points. For Reading Informational Text, the difference was only five percentage points, with an R^2 value for SWD of 33%, compared to 28% for students without disabilities. The same pattern emerged when MAZE and Fluency were predictors. The R^2 value for SWD for MAZE and Fluency predicting Reading Literature in third grade was 43%, while for students without disabilities it was 33% (a difference of ten percentage points). For Reading Informational Text, the R^2 value for SWD was 34%, while for students without disabilities it was 30% (a difference of only four percentage points). In summary, the R^2 results indicate that predictions of RISE subscores for students with disabilities tend to have an error level that is generally similar to students without disabilities but, for Reading Literature only, sometimes *less than* the error level for students without disabilities.

A practically significant difference in slopes for SWD was identified in only one grade level: first grade. Plot **a** in Figure 2 shows that the slope of the best-fitting regression line for SWD was steeper than the slope of the best-fitting regression line for students without disabilities, indicating that there was a stronger relationship between Fluency in first grade and third grade RISE Reading Literature scores for SWD than for students without disabilities. Note that plot **b** does not show a practically significant difference in slopes for the Reading Informational Text outcome, though it was close, having a difference in slopes of 0.07 (0.03 points away from our cutoff of 0.10 for practical significance). The pattern for SWD shows a crossover between the two regression lines that occurs at the midpoint of the Fluency score range. This crossover pattern indicates that SWD will show some overestimation and some underestimation depending on their level of Fluency. Below first grade Fluency scores of 125, the third grade RISE Reading Literature scores for SWD tend to be below the third grade RISE Reading Literature scores of students without disabilities (indicating that SWD with low Fluency scores are being overestimated). Above first grade Fluency scores of 125, the third grade RISE Reading Literature scores for SWD tended to be above the third grade RISE Reading Literature scores of students without disabilities (indicating that SWD with high Fluency scores are being underestimated). Based on the mean residuals for SWD discussed above, the net effect is an overestimation of RISE scores, but it is worth knowing that this overestimation will be stronger for SWD with low Fluency scores.

Figure 2. Differences in Fluency's predictive ability based on receipt of special education services



The fact that the difference in slopes between SWD and students without disabilities only occurred in first grade and only for one outcome (Reading Literature) suggests caution in interpretation. It may be that, even though we used a stringent criterion for practical significance, this result may be spurious (a false positive).

Overall, the predictive validity of Acadience subscores for SWD was similar to that for students without disabilities. Students with disabilities tended to have RISE subscores that were slightly lower than what were predicted from Acadience (i.e., they tended to be overestimated), but this was only practically significant in first grade and was more pronounced for Reading Literature than for Reading Informational text.

Students Learning English (i.e., English Language Learners)

Students learning English showed practically significant levels of overestimation at every grade level and for both RISE subscores, ranging from 0.18 to 0.23 standard deviations (equivalent to an overestimation of 14 to 18 points on the RISE subscores). This means that a student learning English who has a given Acadience score will receive an actual score on the RISE that is 14 to 18 points lower than the score that was predicted for them based on their Acadience score. As a result, English Language Learners are at greater risk of being incorrectly identified as proficient based on their Acadience score but later not achieving proficiency on the RISE Reading test. This overestimation was accompanied by correspondingly lower R^2 values for students learning English that was more severe for Reading Informational Text than for Reading Literature. Whereas R^2 values from a single model for students *not* learning English ranged from 31% to 42% across both subscores and across all models (Fluency, MAZE + Fluency, and MAZE + Retell), R^2 values from a single model using Fluency alone for students learning English were between 11% and 14% for Reading Informational Text (higher for the combination models in third grade: 19% and 20%), and between 19% and 28% for Reading Literature (34% for each combination model in third grade).

When students learning English and students not learning English were modeled separately, the R^2 for a model predicting Reading Informational Text from Fluency for students learning English was 16% in first grade, five percentage points higher than when students learning English were examined as a subgroup under a single model for all students. This improvement suggests that the relationship between Fluency and Reading Informational Text for students learning English may be different in kind (e.g., they may have a different slope), which is explored in Figure 3.

Figure 3 shows that there were differences between English learners and students not learning English in the slopes of the best-fitting regression lines describing the relationship between Fluency and RISE Reading Informational Text. These significant differences in slopes can be seen in the left column of Figure 3 (i.e., plots **a**, **c**, and **e**), with each of those panels demonstrating a practically significant difference in slopes. Plots **b**, **d**, and **f** in the right column of Figure 3 show that the slopes of the lines describing the relationship between Fluency and Reading Literature were not significantly different in slope.

Figure 3. Differences in Fluency's predictive ability based on English-Language Learner status

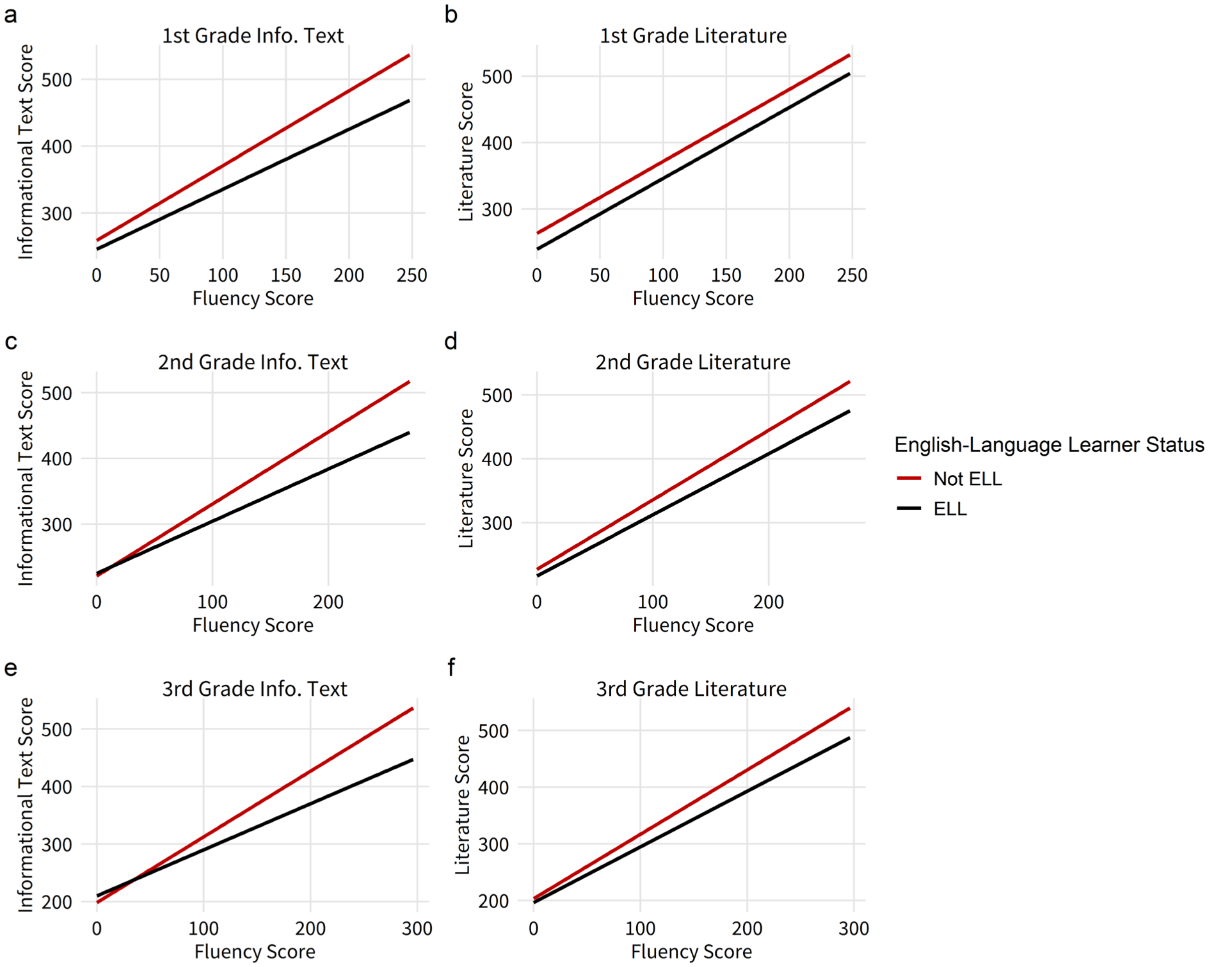


Figure 3 supports the results from mean residuals that the overestimation of students learning English occurs for both Reading Informational Text (left column) and Reading Literature (right column) because the red lines are above the black lines. However, the difference in slopes observed in the left column of Figure 3 shows that the overestimation for Reading Informational Text is close to zero when Fluency scores are low and gets worse at high Fluency scores. Thus, students learning English are at greater risk of being overestimated when their Acadience Fluency scores are high than when those scores are low.

Taken together, the results suggest that Acadience subscores are not as effective at predicting RISE Reading subscores for students learning English. These students will tend to perform 0.18 to 0.23 standard deviations (14 to 18 points) lower on the RISE subtests than a model based on all students would predict. This difference was especially pronounced for Reading **Informational Text**, which showed a much lower R^2 value and a difference in slopes indicating that overestimation on Informational Text will be more severe among English learners with higher Fluency scores. The differential effects across RISE subscores (Literature vs. Informational Text) was unexpected and invites further exploration of the differences between the tasks found in the two RISE Reading subtests. Reading literature may offer a more consistent set of conventions (character, plot, etc.) and opportunities to induce meaning from context than informational text, which may depend more heavily on vocabulary size, but this remains to be formally examined in the context of the RISE Reading assessments.

Race and Ethnicity

So far, the demographics we have examined have only included two groups (e.g., boys/girls, students with disabilities/students without disabilities). Race and ethnicity, however, contain the seven federal categories White, Hispanic, Black, Asian, Native American, Pacific Islander, and Multiracial. There are typically two approaches to handling multiple comparisons like this: (1) comparing each group to a reference group or (2) comparing each group to every other group. Because the second approach would result in 42 comparisons (the number of possible pairs of 7 racial and ethnic groups, irrespective of order, multiplied by 2 outcomes), we chose to use White students as the reference group. This approach prioritizes the sensitivity of detecting differences between the majority student group and students of color.

Fluency Overpredicts RISE Reading Subscores for Black, Hispanic, Native American, and Pacific Islander Students

Analysis of the residuals from a model based on all students indicates that the RISE subscores for Black, Hispanic, Native American, and Pacific Islander students are lower by more than 0.1 standard deviations from the scores predicted by Fluency. This overestimation was consistent across both Reading Informational Text and Reading Literature and fairly consistent across grade levels, but varied in intensity across racial and ethnic groups. For Hispanic students, the overestimation was between 0.12 and 0.14 standard deviations (about 10 to 11 points on the RISE subscores). For Black students, the overestimation was between 0.15 and 0.24 standard deviations (about 12 to 19 points on the RISE subscores), with more overestimation occurring for Informational Text (0.19 to 0.24 standard deviations) than for Literature (0.15 to 0.19 standard deviations). For Native American students, the overestimation was between 0.14 and 0.25 standard deviations (about 12 to 19 points). For Pacific Islander students, the overestimation was between 0.20 and 0.26 standard deviations (about 16 to 20 points). Similar levels of overestimation were observed when MAZE + Fluency or MAZE + Retell were the predictors.

When a single model using Fluency as the predictor was fit for all students, the R^2 values were in line with the overestimation findings above: subgroups with greater overestimation tended to have lower R^2 values. Controlling for the effects of RISE subtest and grade level and using the R^2 values for White students as the reference level, the R^2 values for Hispanic students were 3 percentage points lower, for Native Americans were 7 percentage points lower, for Black students were 8 percentage points lower (although this was very different depending on RISE subtest, with higher R^2 values for Reading Literature than for Reading Informational Text), and for Pacific Islander students were 13 percentage points lower[10]. However, when separate models were used for White students and students of color, the differences between the R^2 values were lower. Compared to White students, the R^2 for Black students was only 4-6 percentage points lower and for Pacific Islander students it was only 7-9

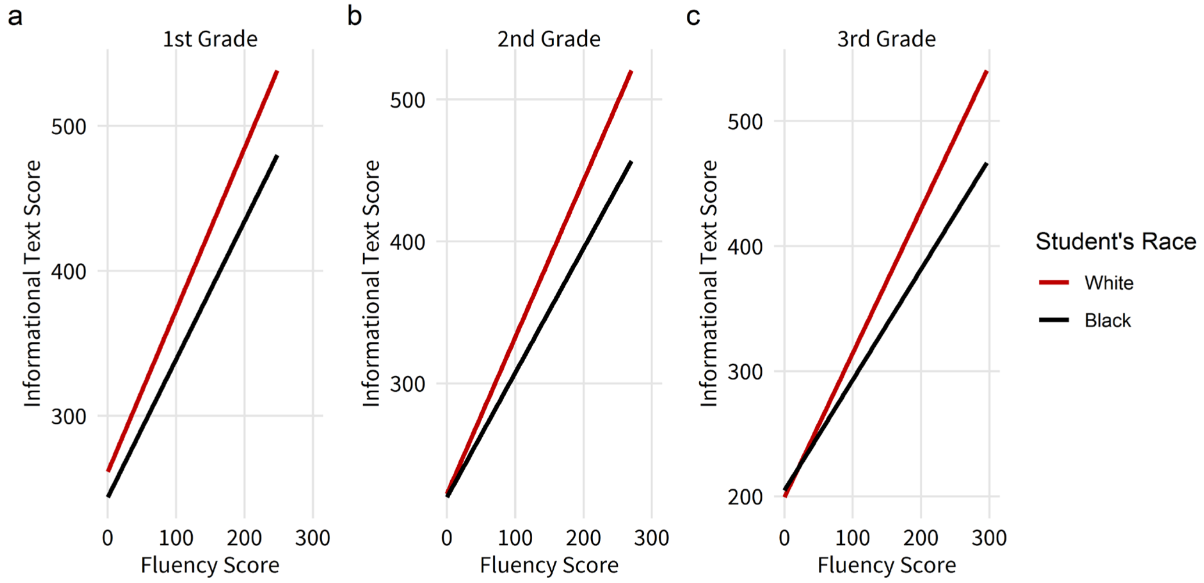
[10] The pattern for MAZE + Fluency and for MAZE + Retell were similar to the pattern for Fluency with the exception that the R^2 for Native American students was less impacted in the combination models than in the Fluency-only model.

percentage points lower[11]. This discrepancy between R^2 values when a single model or separate models are used suggests that the subgroups may have different slopes, which is addressed in the next section.

Overprediction of Reading Informational Text is Greater at Higher Levels of Fluency for Black and Pacific Islander Students

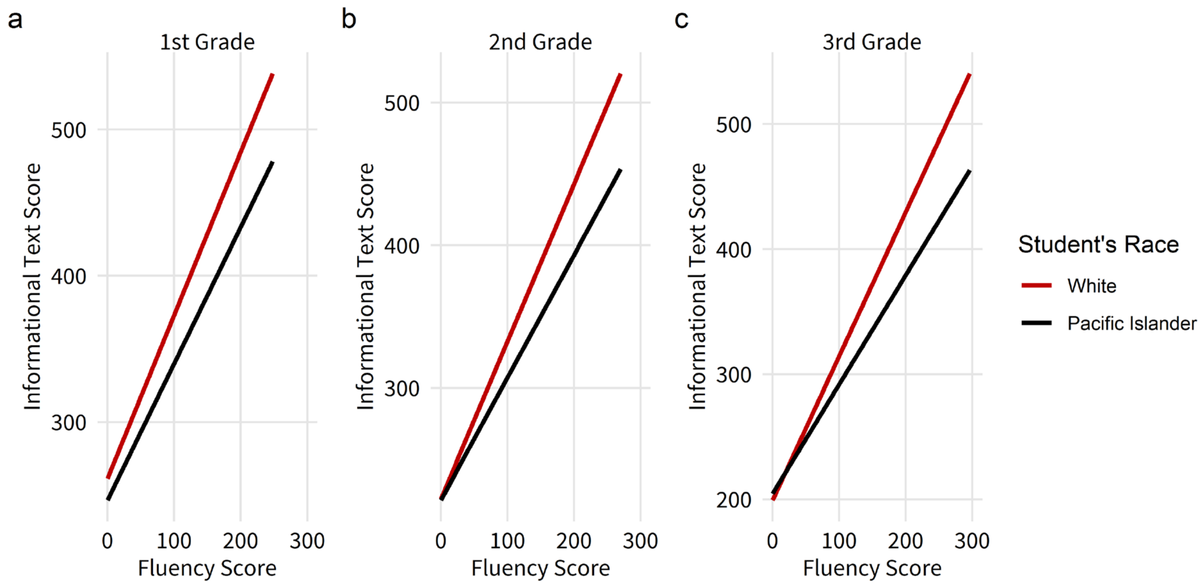
The slope describing the relationship between Fluency and third grade RISE Reading Informational Text was shallower for Black and Pacific Islander students than for White students but was only practically significant in second and third grades (see Figure 4 and Figure 5). The pattern of these results is similar to that for students learning English, who had significantly shallower slopes between Fluency and Reading Informational Text than students not learning English. As with students learning English, this pattern indicates that overestimation will be greater for students with higher scores on the Fluency subscore than for students with lower scores. Also consistent with the findings for students learning English is the fact that a difference in slopes was not observed for RISE Reading Literature, only for Informational Text.

Figure 4. Fluency's predictive ability for Black Students and White Students



[11] The R^2 values for the combination models fit separately to White, Black, and Pacific Islander students were similar to the models using Fluency only, with all differences in the R^2 values between White and Black students and between White and Pacific Islander students being less than 10 percentage points.

Figure 5. Fluency's predictive ability for Pacific Islander and White students



Overprediction of Reading Informational Text is Greater at Higher Levels of Fluency for Native American Students in Second Grade

The only other practically significant difference in slopes was for Native American students in second grade. Specifically, the slope of the relationship between Fluency and Reading Informational Text was lower for Native American students than for White students, following the pattern noted above in Figure 4 and Figure 5. The difference in R^2 values between White and Native American students when those groups were modeled separately was not practically significant at 7 percentage points. Because this was the only grade level in which we found this difference, we are not as confident that the pattern observed for English Language Learners and students who are Black or Pacific Islander should be extended to Native American students. While it is possible this is a true difference, there is also a chance that this is a false positive.

For all other racial and ethnic groups across all three grades, we did not find any practically significant differences in the predictive validity of Acadience subscores compared to White students.

Differences in Predictive Validity: Summary

In this section, we tested whether the predictive validity of Acadience subscores for predicting RISE subscores differed across the student demographic categories of gender, economic disadvantage, students with disabilities, students learning English, and race and ethnicity. One way that predictive validity can vary across subgroups is when predictions are systematically above or below the actual values for a subgroup. These overestimations and underestimations are plotted in Figure 6, which shows the mean residuals by subgroup based on a single model predicting RISE scores from Fluency that included all students. Figure 6 averages across grade level and RISE subtest (Informational Text or Literature) to focus on the general pattern for a demographic category. The dashed red lines indicate our threshold for overestimates or underestimates exceeding practical significance: 0.1 standard deviations.

Figure 6. Systematic Overestimation or Underestimation of Subgroups

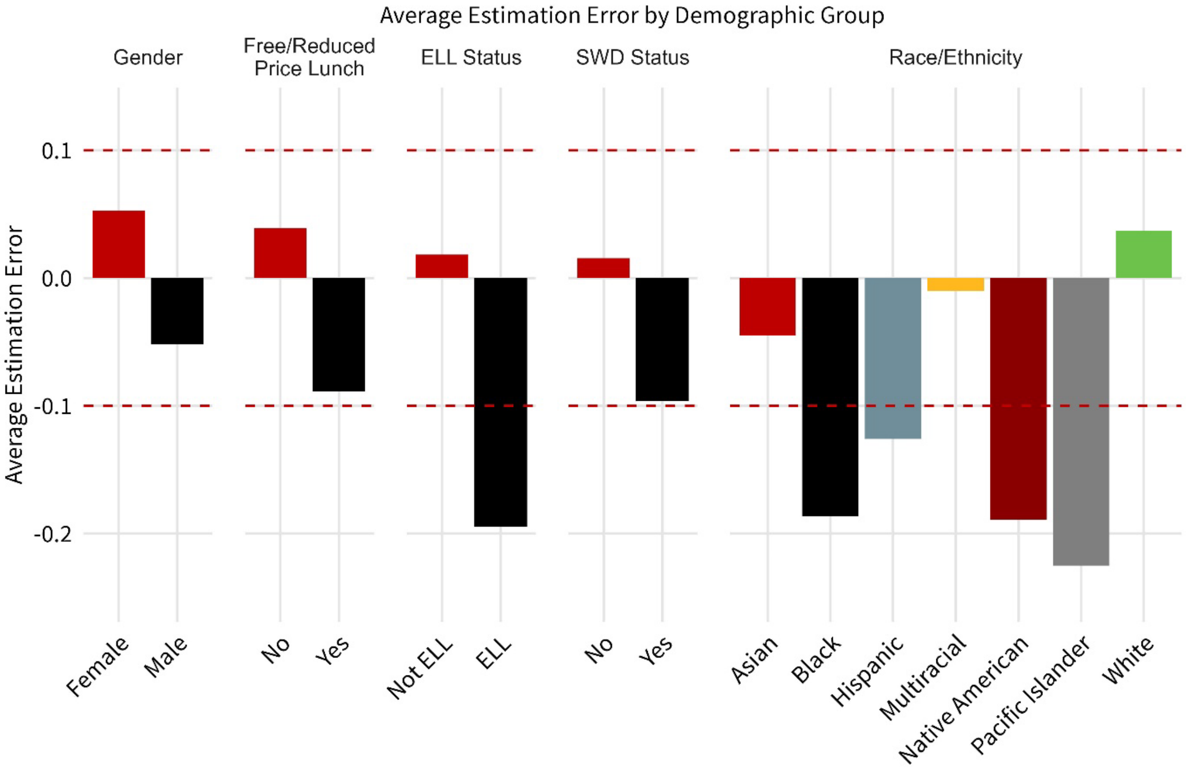


Figure 6 shows that for gender, economic disadvantage, and disability status, predictions made from Fluency did not exceed our 0.1 standard deviation threshold. Across those variables, the predictions made for each subgroup were not consistently above or below their actual RISE scores. The story is different for students learning English and for some race and ethnicity categories. The RISE scores of students learning English and students who were Black, Hispanic, Native American, or Pacific Islander tended to be lower than the scores predicted for those groups based on their Fluency score. This overestimation is problematic because students who are overestimated are more likely to be classified as proficient based on their Acadience score but later fail to meet the benchmark for proficiency for the RISE Reading assessment. This pattern contributes to the rate of "False Passes" discussed in the next section on Reading on Grade Level cut scores.

A second way that predictive validity can vary across subgroups is when there are differences in the accuracy of predictions across subgroups. Overestimation and underestimation occur when inaccuracies are biased in one direction: either consistently below or consistently above the actual values. However, it is possible for inaccuracy to be greater for one group even when there is not consistent over- or underestimation but rather just more random error. This can be measured using R^2 values for each subgroup based on a single model that included all students. The R^2 values for each subgroup are presented in Figure 7.

Figure 7. Average R^2 by Demographic Group

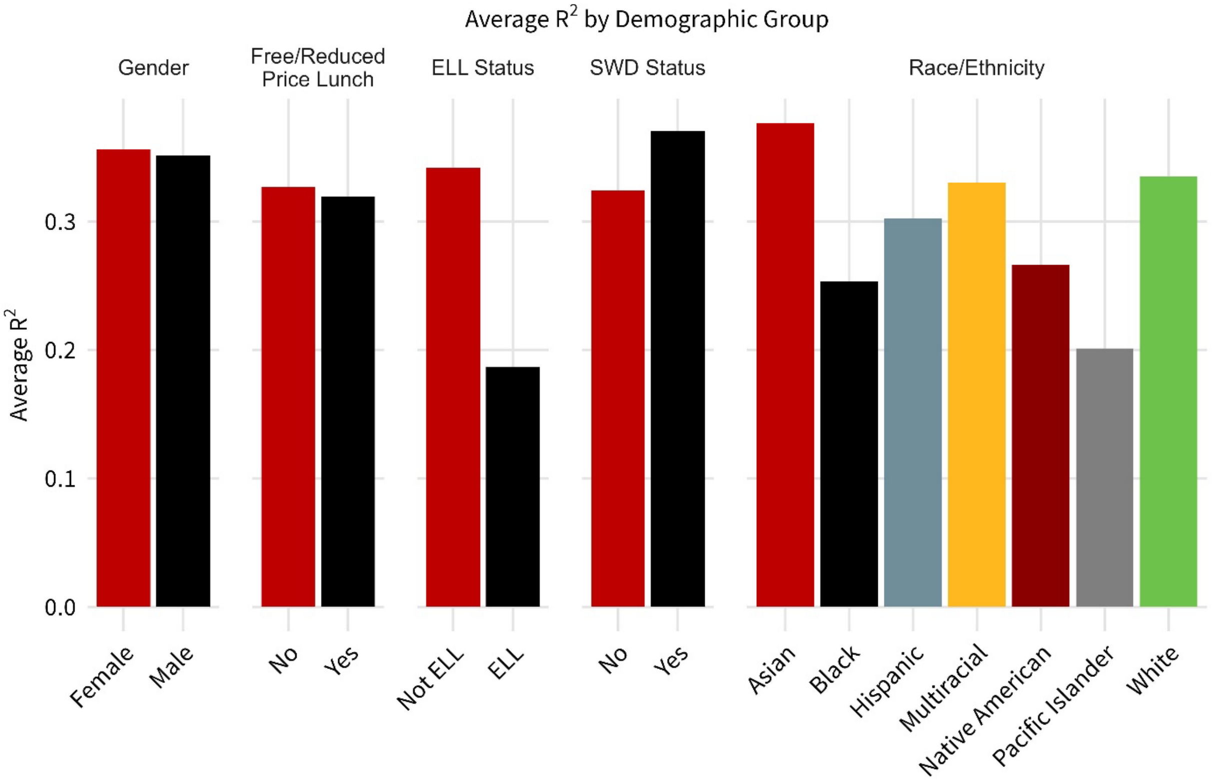


Figure 7 shows only small differences in R^2 values within the categories of gender, economic disadvantage (Free/Reduced Price Lunch), and disability status (SWD). However, there were large differences within the categories of English Language Learner status and race and ethnicity. Specifically, there was more error in the predictions for English Language Learners (relative to students not learning English), and in the predictions of Black, Native American, and Pacific Islander students (relative to White students). These are the same groups who were overestimated by the model in Figure 6. This pattern suggests that the major source of error in prediction is the overestimation of these subgroups.

An additional pattern observed for several of the subgroups who were overestimated (students learning English, Black students, Pacific Islander students, and (for one grade level only) Native American students) is that the level of overestimation increased with Fluency score. Students in those groups who have higher Fluency scores will show a greater degree of overestimation than students in those groups with lower Fluency scores.

4 | Reading on Grade Level Cut Scores

Background

The key to understanding our cutoff score findings is to understand the four different outcomes that can occur when using the Acadience reading composite cut score to predict whether a student will be above or below the proficiency cutoff score on the RISE ELA composite. Those four outcomes, which are illustrated below in Table 2, are:

1. The student scores above the Acadience cut score, and thus the student is predicted to pass the RISE, and the student goes on to pass the RISE (i.e., **True Pass**);
2. The student scores below the Acadience cut score, and thus the student is predicted to fail the RISE, and the student goes on to fail the RISE (i.e., **True Fail**);
3. The student is predicted to pass the RISE, but the student goes on to fail the RISE (i.e., **False Pass**);
4. The student is predicted to fail the RISE, but the student goes on to pass the RISE (i.e., **False Fail**).

Table 2. Four Possible Outcomes of Predicting Pass or Fail and Student Actually Passing or Failing

	Actually Passes	Actually Fails
Predicted to Pass	*True Pass*	*False Pass*
Predicted to Fail	*False Fail*	*True Fail*

In other words, there are two circumstances where the prediction is correct (True Pass and True Fail), and two where the prediction is incorrect (False Pass and False Fail). For the incorrect outcomes, an important decision to be made by decision makers is whether one of those errors is more costly: predicting a student is going to pass when they go on to fail (**False Pass**) or predicting a student is going to fail when they go on to pass (**False Fail**). A third possibility is that these two errors are equally costly.

Costs of a False Pass

If scoring above the benchmark for proficiency on an early test reduces the likelihood that students will receive remedial assistance, then the costs of a False Pass include students not receiving interventions or the help that they need to be proficient in reading by third grade. The early test results indicated they would pass, so no interventions were provided, and the student later failed the third grade RISE test. The costs of a False Pass error could be considerable if the student is now at greater risk of falling behind. Interventions to catch the student up may be more extensive and expensive than they would have been had the student been correctly identified earlier.

Costs of a False Fail

The costs of a False Fail include students receiving help that they may not need. However, the students in the False Fail scenario are not all the same. The dynamics of the False Fail are more complex than the False Pass because failure on the earlier test may trigger remedial interventions, which may be responsible for the student later passing the third grade RISE test. Thus, the apparent inconsistency between failing the earlier test and passing the later test may not always indicate a failure of prediction, but rather sometimes indicate an effective system of intervention. Thus, False Fail "errors" include not only students who received unnecessary interventions and would have passed the RISE later anyway, but also students who benefited from the intervention and later passed because of the intervention. In short, the "costs" of False Fail errors are smaller than they appear.

"Accuracy"

In evaluating how well an earlier test predicts a later test, it is common to combine the false pass and false fail events and treat both as "incorrect" predictions to be contrasted with the times when a student was "correctly" predicted to pass or fail. When the number of correct predictions is divided by the total number of predictions (both correct and incorrect), the technical term for the result is "Accuracy." However, it's important to keep in mind that accuracy treats the two types of errors (False Pass and False Fail) as equally important.

Optimal Cutoffs

When the costs of a False Pass are considered equal to the costs of a False Fail, then the optimal cutoff score will be the one that maximizes "Accuracy" as defined above (the highest ratio of correct predictions to total predictions). However, it is possible to adjust this cutoff so that it is sensitive to perceived differences in the relative costs of a False Pass and False Fail. In those cases, the optimal cutoff is the test score at time 1 (i.e., an Acadience score) that divides students into two groups (proficient and not proficient) such that the overall expected costs of False Passes and False Fails at time 2 (i.e., based on proficiency on the third grade RISE Reading test) is minimized. In assessing the optimal cutoff for each grade, we examined three scenarios:

1. Treat a **False Pass** as twice as costly as a **False Fail.** This approach prioritizes reducing the number of future fails by erring on the side of over-diagnosing them in earlier grades (and perhaps applying an intervention). This approach would raise the cutoff score, resulting in more students in early grades being classified as not proficient.

2. Treat a **False Fail** as twice as costly as a **False Pass**. This approach prioritizes reducing the number of students who are incorrectly identified as failing in early grades. This approach would lower the cutoff score, resulting in more students in early grades being classified as proficient.

3. Treat each error as equally costly. The third scenario can essentially be characterized as maximizing **Accuracy**, which is the percent of predictions that were correct (True Pass + True Fail) out of all the predictions made.

The choice of making one error twice as costly as another is arbitrary and is not meant as a recommendation for the relative importance of those errors. The choice of making one error twice as costly as another was selected simply to illustrate how different costs can affect cutoff scores. The choice of the relative costs must be made by decision makers based on an understanding of the true costs (in time, money, missed opportunities, etc.) that result from the two errors.

For all three grade levels of Acadience scores, we used a procedure that assessed the performance of all possible cut scores according to all three of our optimal cut score criteria (equal cost, 2x cost for false fail, 2x cost for false pass). The procedure then identified the optimal cut score for each grade level for each of our three optimal cut score types. After each optimal cut score was determined, we used a bootstrapping procedure to estimate a range of possible optimal cut scores around that estimate (i.e., 95% confidence interval). This confidence interval conveys the precision of the cut score estimate, with narrower confidence intervals indicating more precise cut score estimates. For technical details on these analyses, please see Appendix D.

First Grade

According to the USBE's Accountability Technical Manual Appendix G and personal communication with USBE officials, the cut score on the Acadience Reading test in first grade is 208+. To evaluate this cut score and how it compares with optimal cut scores, we used data from all students with a RISE ELA score in third grade during the 2018-2019, 2020-2021 or 2022-2023 school years. 2019-2020 was excluded because students did not have third grade RISE scores in that year due to the pandemic, and 2021-2022 was excluded because students in third grade during that year would have attended first grade in 2019-2020, which has the same problem with missing assessment data. The accuracy of this cut score was 78.22%. If you summed together the two "correct" outcomes – 1) the number of students who scored 208 or higher in first grade and also scored above the threshold for proficiency on the third grade RISE Reading test, and 2) the number of students who scored below 208 in first grade and also scored below the threshold for proficiency on the third grade RISE Reading test – and divided that sum by the total number of students who took both tests, you would obtain 78.22%. For all first grade optimal cut scores and their corresponding 95% confidence intervals, see Table 3.

Table 3. First grade USBE Acadience cut score and model-estimated optimal cut scores

	Cut Score	95% CI Lower Bound	95% CI Upper Bound	Accuracy	False Pass Rate	False Fail Rate
USBE	208	-	-	78.22 %	13.11 %	8.68 %
Equal weight (Accuracy)	202.75	198.48	207.55	78.26 %	13.81 %	7.93 %
False Pass 2x cost	237.69	235.38	246.16	76.04 %	8.81 %	15.15 %
False Fail 2x cost	160.29	153.22	162.98	76.71 %	19.68 %	3.62 %

As can be seen in Table 3, the USBE's cut score of 208 is slightly higher than the cut score optimized for Accuracy (the "Equal Weight" model, with cutoff of 202.75) but it has almost the same level of overall accuracy (78.22%, compared to 78.26% for the Equal Weight model) and has a slightly lower False Pass Rate (13.11%, compared to 13.81% for the Equal Weight model). Given the costs associated with False

Pass noted above (students not getting the help they need), the USBE's current cut score could be considered marginally better than the Equal Weight model. By raising the cut score to 237 (the score identified when False Pass has double the cost), the rate of False Pass could be reduced by 4 percentage points, from its current level of 13% down to 9%. Unfortunately, doing so would raise the False Fail Rate (of the students who actually passed the RISE Reading test, the percentage who were marked not proficient on the Acadience test) even more: by about 6.5 percentage points, from about 8.5% up to 15%. Decision makers may consider whether it is worth exploring cut scores between the current level and the 2x cost scenario to see if there is a balance between False Pass and False Fail rates that would be more desirable.

Lowering the cut score toward 160, the optimal score when False Fails have twice the cost of False Passes, is not recommended. Although this would pull the False Fail rate from its current level of 9% down to 4%, it would push the False Pass rate up to 20%. This would mean that one out of every five students who are marked as proficient at first grade would go on to fail the third grade RISE test.

Second Grade

The USBE's cut score on the Acadience reading composite for second grade is 287+. We evaluated the performance of this cut score using data from all students with a RISE ELA score in third grade during the 2018-2019, 2021-2022 or 2022-2023 school years. We did not include students who attended third grade during the 2019-2020 school year because of the absence of third-grade assessment data that year due to the pandemic, and we excluded students who attended third grade in the 2020-2021 school year because they would have attended second grade in 2019-2020. The accuracy of this cut score was 80.09%. For all second grade optimal cut scores and their corresponding 95% confidence intervals, see Table 4.

Table 4. Second grade USBE Acadience cut score and model-estimated optimal cut scores

	Cut Score	95% CI Lower Bound	95% CI Upper Bound	Accuracy	False Pass Rate	False Fail Rate
USBE	287	-	-	80.09 %	11.81 %	8.1 %
Equal weight (Accuracy)	283.06	279.26	285.64	80.16 %	12.42 %	7.42 %
False Pass 2x cost	311.88	308.67	314.46	78.03 %	7.82 %	14.15 %
False Fail 2x cost	246.91	244.54	253.67	78.65 %	18.22 %	3.13 %

As shown in Table 4, the USBE's cut score of 287 is just slightly above the cutoff identified by the Equal Weight model. Despite this difference, they have essentially the same level of accuracy in the currently considered data set (80%). By raising the cut score to 312 (the optimal score when False Pass has twice the cost of False Fail), the rate of False Pass could be reduced by 4 percentage points, from its current level of 12% down to 8%. However, doing so would push the False Fail rate (of the students who later pass the third grade RISE test, the percent marked as not proficient in second grade) up 6 percentage points, from 8% to 14%.

Third Grade

In third grade, the USBE's cut score for the Acadience reading composite is 405+. We evaluated the performance of this cut score using data from all students with a RISE ELA score in third grade during the 2018-2019, 2020-2021, 2021-2022 or 2022-2023 school years. We did not include students who attended third grade during 2019-2020 due to lack of data from the pandemic. The accuracy of this cut score was 81.49%. For all third grade optimal cut scores and their corresponding 95% confidence intervals, see Table 5.

Table 5. Third grade USBE Acadience cut score and optimal cut scores

	Cut Score	95% CI Lower Bound	95% CI Upper Bound	Accuracy	False Pass Rate	False Fail Rate
USBE	405	-	-	81.49 %	11.77 %	6.74 %
Equal weight (Accuracy)	407.25	405.35	410.59	81.49 %	11.49 %	7.03 %
False Pass 2x cost	447.13	443.16	451.21	79.34 %	6.96 %	13.69 %
False Fail 2x cost	368.19	363.96	372.64	80.01 %	16.69 %	3.3 %

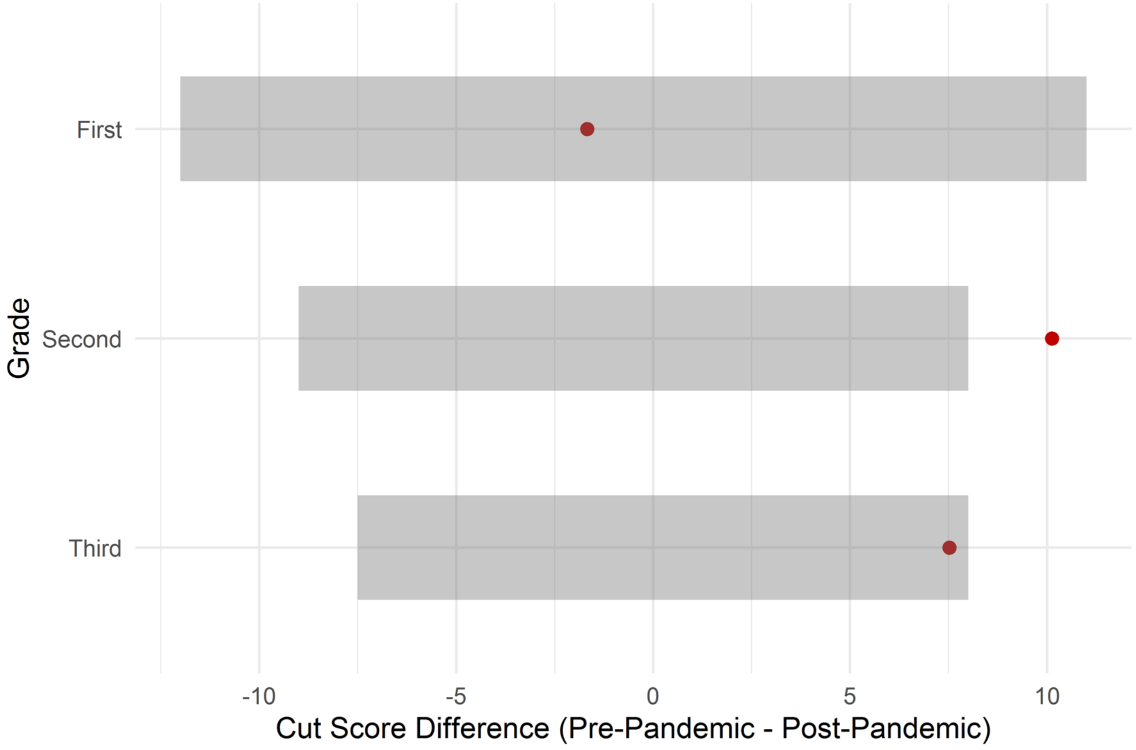
USBE's cut score of 405 is very close to the optimal cut score for maximizing accuracy, being only 2.25 points away from the optimal cut score. By raising the cut score from 405 to 447, the False Pass rate could be cut by 5 percentage points, from 12% down to 7%. However, this would come at the cost of increasing the False Fail rate by 7 percentage points, from 7% up to 14%. Decision makers might consider whether it is worth exploring cut scores between 405 and 447 to see whether there is a value that strikes a better balance between False Pass and False Fail rates than the current cut score.

Pre-Pandemic vs. Post-Pandemic

One question that arose during the UEPC analysis was whether the optimal cut score differs based on the years of data we included. To answer this question, we reran our cut score analysis maximizing Accuracy twice for each grade: (1) using data only from the year before the pandemic (2018-2019) and

(2) using data only from years after the pandemic began (2020-2021, 2021-2022 and 2022-2023). We calculated the difference between these two cut scores, and then used a simulation procedure to estimate the 95% confidence interval for this difference (the shaded bars in Figure 7, see Appendix D for details). If the difference (the red circle in Figure 7) overlaps with the 95% confidence interval range, then this indicates that the optimal cut scores do not change based on the years of data used. If the difference (red circle) is outside the 95% confidence interval range (shaded rectangle), then this indicates that the optimal cut scores do change depending on the years of data used. Our analyses revealed that the optimal cut scores for first grade and third grade do not differ between pre-pandemic and post-pandemic years. However, the optimal cut score for second grade does differ between pre-pandemic and post-pandemic years, with the pre-pandemic optimal cut score (286.45) being significantly higher than the post-pandemic cut score (276.33).

Figure 8. Difference Between Pre- and Post-Pandemic Cut Scores



Overall, the pattern in Figure 7 indicates that, if the goal of setting an Acadience cut score is to maximize the percentage of students who are accurately predicted to pass or fail the third grade RISE Reading test, then this cut score will need to be recalculated over time.

5 | Conclusions, Policy Implications, Limitations, and Future Directions

This study had three main goals. First, this study sought to determine whether Acadience Reading subscores were valid predictors of RISE Reading subscores. Second, this study sought to determine whether these predictions differed by any student demographics. Third, this study sought to evaluate the current Acadience Reading composite Reading on Grade Level cut scores used by USBE. Below, we summarize the findings for each of these three goals and discuss policy implications, limitations, and future research directions.

Predictive Validity

Our predictive validity results indicate that all the Acadience subscores at first, second, and third grade were significantly positively related to both RISE Reading subscores. However, **Fluency was much more predictive than Accuracy or Retell**. The percentage of variance in RISE Reading scores that could be explained from Fluency scores in Grade 1 through 3 ranged from 27% to 35% as measured by R^2. Although these R^2 values are impressively high, it is also worth considering the precision of any prediction made from Acadience to RISE. One way of expressing this precision is through a "prediction interval," which expresses the "plus or minus" around a predicted value that contains a given percentage of cases. The 66% prediction intervals for Fluency ranged from ± 58 to ± 65 points for third grade RISE Reading subscores, which accounts for 27% to 35% of the total range of RISE Reading subscores. Thus, even though Acadience Fluency is able to explain about one-third of the variance in RISE Reading subscores, it is still making predictions that have a fairly wide margin of error for individual students.

Although the MAZE Acadience subscore was available only for third grade, it was almost as predictive as Fluency for that grade. The superior performance of Fluency to the other Acadience subscores is in line with a previous meta-analysis that compared oral reading fluency to MAZE with regard to their prediction of reading achievement tests and found that oral reading fluency was the stronger of the two predictors (Shin & McMaster, 2019).

Of all the Acadience subscores, Accuracy was the least predictive, being the worst predictor in second and third grade and posting the worst scores on our two metrics (R^2 and prediction interval width) of all the predictors we examined. The Accuracy measure is the Fluency measure (number of words read correctly) divided by the total number of words read. Its performance was low enough that while our models did demonstrate a significant positive relationship between Accuracy and both of the RISE Reading subscores, we would not recommend its use for predicting future performance. One reason that Accuracy may have performed so poorly relative to the other subscores is that most students had very high Accuracy. The restricted variability in Accuracy means that there is simply less information available from that measure, diminishing its effectiveness as a predictor (a problem sometimes labeled "restriction of range").

In addition to single scores, in third grade, we also examined the combinations of MAZE with Retell and MAZE with Fluency. As you may expect, given our individual predictor results, these combinations were also significantly and positively related to RISE Reading subscores. However, because Fluency was so much more predictive than Retell, **the combination of MAZE and Fluency was much more predictive than the combination of MAZE and Retell.**

Policy Recommendations

For prediction purposes, this study demonstrates that more weight be given to Fluency than to other Acadience subscores for predicting third-grade RISE Reading subscores. MAZE was a close second in predicting RISE, but it is only available in third grade. While using MAZE and Fluency together in third grade would technically provide the best prediction of the models we used, there are pragmatic reasons to use Fluency alone. Most notably, using MAZE and Fluency together would require users to enter both scores into a formula to get an effective prediction. You could not, for example, simply take the average of the two scores to get a good estimate of the student's third grade RISE Reading scores. In contrast, with Fluency alone, a single lookup table could be constructed that matched each Fluency score with an expected RISE score (and ideally a prediction interval to add some understanding of the precision of the estimate).

Limitations

These results are strictly related to the predictive validity of the Acadience Reading subscores in predicting RISE Reading subscores, and therefore our results should not be used to draw conclusions about other forms of validity or reliability. For example, the results of this study do not directly touch on construct validity (i.e., do the Acadience Reading subtests measure what they are supposed to measure?) or test-retest reliability (e.g., if a student took the test twice three days apart and assuming the student's reading ability has not changed, do they get close to the same score?). Evidence for Acadience Reading's strong psychometric properties can be found by browsing the publications at Acadience Reading's website: https://acadiencelearning.org/resources/presentations-publications/.

Future Directions

In this study, we examined only two combinations of Acadience subscores: MAZE + Fluency and MAZE + Retell. Future research may examine whether other combinations of subscores, such as Fluency + Retell + Accuracy, meaningfully improve the prediction of RISE Reading scores over Fluency alone. It may also be of interest to see whether early grade scores are predictive of much later outcomes, such as ACT scores. There is some evidence to suggest that oral reading fluency is also related to arithmetic ability (Balhinez & Shaul, 2019). Therefore, future work may also consider examining whether Acadience Reading subscores are related to future achievement in mathematics. Last, given that Accuracy had the worst performance of the subscores examined, future research might explore whether there are some circumstances or student groups for which the Acadience Accuracy subscore is better suited.

Differences Across Student Groups

Predictions of RISE Reading subscores from Acadience Reading subscores did not significantly differ in accuracy for many of the student groups we evaluated. Differences by gender, economic disadvantage, or special education status tended to be small. **However, RISE subscores predicted from Acadience subscores tended to be overestimated by 0.1 standard deviations or more for five groups: students learning English (vs. Students not learning English), Black students (vs. White**

students), Hispanic students (vs. White students), Native American students (vs. White students), and Pacific Islander students (vs. White students). This is a problem if the predicted score erroneously indicates a student will pass a later test, in which case interventions may be less likely and the student would be at greater risk of falling behind. The intensity and consistency of this pattern varied across these groups, with the greatest consistency among students learning English. Students learning English also showed the highest level of overall error in their predicted RISE scores (as measured by R^2). Finally, three groups – students learning English, Black students, and Pacific Islander students – also showed a pattern where the degree of overestimation was greatest for students with high scores on Fluency and were close to zero when Fluency scores were zero.

Policy Recommendations

An important implication of the overprediction we describe above is at the school level: **Schools with a higher percentage of students who are English Language Learners, Black, Hispanic, Native American, or Pacific Islander will tend to show a pattern in which RISE Reading scores will tend to be lower than expected based on the earlier Acadience scores.** This should not be interpreted as a decline in performance from Acadience to RISE but rather as simply an artifact of using a model that tends to over-predict these student groups relative to their reference groups. In addition, the dramatically impaired predictive validity of Acadience subscores for students learning English should make administrators extremely cautious in making forecasts for this group. **Given the poor ability of Acadience subscores to predict RISE subscores for English Language Learners, decision makers might consider whether there is a supplemental test which, when used in combination with Acadience Reading subscores, could improve the precision of forecasts for English Language Learners.**

Limitations

Our results are restricted to the years of data we analyzed and the groups we examined. **Our work cannot be used to make inferences about students who fall within more than one of the demographic groups we examined**. Students in such “intersectional” categories are not always simple sums of the effects associated with their groups. Further research would be needed to explore the “non-additive” nature of each combination of traits.

Future Directions

Further research is needed to explore why some groups of students with the same score on an Acadience measure as their reference groups tend to systematically underperform on the RISE Reading tests. To examine this question, researchers could test whether a similar overprediction is observed from first grade Acadience scores to third grade Acadience scores. If no overprediction is observed, then the overprediction observed when RISE is the outcome may be partly the result of the difference between the Acadience and RISE testing formats.

A second line of inquiry could explore the implications of overprediction for the Reading on Grade Level cut scores. Students who are being overestimated are at greater risk of being “False Passes”: appearing not to require any assistance based on their Acadience score, but later failing the RISE test.

Future research could explore the magnitude of this risk: how *much* more at risk are these groups of students than their reference groups?

Lastly, future research could examine how the relationship between Fluency and RISE changes depending on a student's specific special education disability label such as Specific Learning Disability, Autism, Speech/Language Disability, etc. Students receiving Special Education services are a diverse group and it is likely that the results will vary across disability labels.

Reading on Grade Level Cutoffs

We examined the USBE's cut scores for the Acadience Reading composite score for first, second, and third grade and compared them to three different model-based definitions of optimal cutoffs: (1) a cut score maximizing Accuracy, (2) a cut score weighing false positives as twice as costly as false negatives, and (3) a cut score weighing false negatives twice as costly as false positives. Results indicate that **all three USBE cut scores perform relatively well, with all three having accuracies above 78%, and all three were relatively close to our optimal cut scores maximizing Accuracy**.

Policy Recommendations

Although a primary function of Acadience Reading cutoff scores is to accurately forecast whether a student will score above or below the benchmark for proficiency on the RISE Reading test, **decision makers should also consider whether it might be in students' interests to raise the Acadience cutoff score above the level that optimizes Accuracy with the goal of reducing the rate of False Passes** (students who appear to be proficient on an Acadience test but later fail the RISE test). Making this change would only be warranted if one could be reasonably confident that students identified as "not proficient" on an Acadience test will receive remedial assistance. The degree to which interventions after failing to meet benchmark are widespread and applied with fidelity is an important question and may be a worthwhile topic for future investigation. If students scoring below the benchmark for proficiency receive more assistance, then the cost of making a False Pass error is higher because False Pass students are being deprived of needed assistance. That higher cost could be offset by raising the Acadience cutoff score, which lowers the rate of False Passes. Although doing so would result in an Acadience pass rate that is lower than the RISE pass rate, leading some to say that the Acadience results are overly pessimistic, decision makers could communicate that this pattern is a necessary consequence of a process designed to protect students from being left behind. If scoring below the benchmark for proficiency results in interventions, then raising the Acadience cutoff score will not only direct help to the students who need it but will also contribute to Utah Senate Bill 127's (2022) goal of 70% of third grade students reading on grade level by June 2027. In considering changes to the cut score, decision makers may want to examine a range of cut scores around the current level, examining how each one corresponds not only to the False Pass rate but also to the False Fail rate and overall Accuracy.

Limitations

We used several years of data to calculate optimal cut scores because doing so reduces the likelihood that results will be overly influenced by a single unusual year. However, optimal cut scores should not be viewed as valid in perpetuity. Our finding that the second grade cut score changed significantly from pre- to post-pandemic indicates **a need to reevaluate cut scores every few years using the most recent data because the optimal cut score can shift over time.**

Another limitation concerns our choice of weighting each type of error – False Passes and False Fails – as either twice as costly as the other or as equally costly. These weights were arbitrary and intended to illustrate how different weighting schemes can affect the choice of an "optimal" cut score. It is unlikely that a False Fail would be exactly twice as costly as a False Pass. Therefore, if USBE is interested in using an approach other than maximizing accuracy, future analyses will need to be done to identify the compromise between False Pass and False Fail rates that best serves student interests.

Future Directions

As mentioned above under Limitations, future work could explore how a wide range of possible cut scores are related to different False Pass and False Fail rates. Rather than attempting to estimate the relative weight to assign to the costs of these two errors, it may be easier to deal directly with specific cut scores and how they are linked to specific False Pass and False Fail rates so that decision makers can easily see the likely consequences of changing the cut score.

One important open question is whether students who do not meet benchmark for proficiency on Acadience tests are receiving some kind of remedial intervention. For this study, we had no information regarding the type, frequency, or effectiveness of reading interventions in the state if a student does not meet proficiency on the Acadience reading test. Knowing more about how schools, teachers, parents, and administrators use Acadience reading scores to trigger support for students would help clarify several important questions from this work. For example, False Fails include both students who would have failed but didn't because they received support, and students who were simply misclassified due to the inherent imprecision in making predictions of future performance. If more information were available about the nature and frequency of supports being triggered when a student performs below proficiency on the Acadience reading test, it would be possible to estimate what percentage of False Fails are success stories of supports helping a student go from below reading proficiency, to reading proficiency. Moreover, given the state's goal of having 70% of students reading on grade level by 2027, research into the supports being used and their effectiveness in helping students become proficient in reading will clarify which interventions best support progress toward this goal.

References

Assessments. (2024). Utah State Board of Education. https://schools.utah.gov/assessment/assessments

Balhinez, R., & Shaul, S. (2019). The Relationship Between Reading Fluency and Arithmetic Fact Fluency and Their Shared Cognitive Skills: A Developmental Perspective. *Frontiers in Psychology*, *10*, 1281. https://doi.org/10.3389/fpsyg.2019.01281

Barton, K. (2024). *MuMIn: Multi-Model Inference*. https://CRAN.R-project.org/package=MuMIn

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Efron, B., & Tibshirani, R. (1997). Improvements on Cross-Validation: The 632+ Bootstrap Method. *Journal of the American Statistical Association*, *92*(438), 548–560. https://doi.org/10.1080/01621459.1997.10474007

Good III, R. H., Kaminski, R. A., Cummings, K. D., Dufour-Martel, C., Petersen, K., Powell-Smith, K. A., Stollar, S., & Wallin, J. (2011). *Acadience Reading K-6 Assessment Manual*.

Johnson, P. C. D. (2014). Extension of Nakagawa & Schielzeth's R^2_{GLMM} to random slopes models. *Methods in Ecology and Evolution*, *5*(9), 944–946. https://doi.org/10.1111/2041-210X.12225

Knowles, J. E., & Frederick, C. (2024). *merTools: Tools for Analyzing Mixed Effect Regression Models*. https://CRAN.R-project.org/package=merTools

Lüdecke, D., Ben-Shachar, M. S., Patil, I., & Makowski, D. (2020). Extracting, Computing and Exploring the Parameters of Statistical Models using R. *Journal of Open Source Software*, *5*(53), 2445. https://doi.org/10.21105/joss.02445

Mastandrea, M. D., Field, C. B., Stocker, T. F., Edenhofer, O., Ebi, K. L., Frame, D. J., Held, H., Kriegler, E., Mach, K. J., Matschoss, P. R., Plattner, G.-K., Yohe, G. W., & Zwiers, F. W. (2010). *Guidance Note* *for Lead Authors of the IPCC Fifth Assessment Report on Consistent Treatment of Uncertainties*. Intergovernmental Panel on Climste Change (IPCC).

Nakagawa, S., Johnson, P. C. D., & Schielzeth, H. (2017). The coefficient of determination R2 and intra-class correlation coefficient from generalized linear mixed-effects models revisited and expanded. *Journal of the Royal Society Interface*, *14*.

Nakagawa, S., & Schielzeth, H. (2013). A general and simple method for obtaining R^2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution*, *4*(2), 133–142. https://doi.org/10.1111/j.2041-210x.2012.00261.x

R Core Team. (2024). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.r-project.org/

Shin, J., & McMaster, K. (2019). Relations between CBM (oral reading and maze) and reading comprehension on state achievement tests: A meta-analysis. *Journal of School Psychology*, *73*, 131–149. https://doi.org/10.1016/j.jsp.2019.03.005

Taylor, W. L. (1953). “Cloze Procedure”: A New Tool for Measuring Readability. *Journalism Quarterly*, *30*(4), 415–433. https://doi.org/10.1177/107769905303000401

Thiele, C., & Hirschfeld, G. (2021). cutpointr: Improved Estimation and Validation of Optimal Cutpoints in R. *Journal of Statistical Software*, *98*(11), 1–27. https://doi.org/10.18637/jss.v098.i11

Appendix A: Acadience Subscore's Predictive Validity Technical Details

All analyses were conducted using R (version 4.4.0), the statistical programming language (R Core Team, 2024). For each of our multilevel models, we used the lme4 package (Bates et al., 2015). For each grade (i.e., first, second, and third) we ran six multilevel models. Each model consisted of one Acadience subscore as the predictor from the same grade (i.e., Fluency, Retell, and Accuracy), and one third grade RISE reading subscore as the outcome (i.e., Reading Informational Text and Reading Literature). To account for clustering, two random effects were entered into each model: (1) A random intercept for the student's school during that grade, and (2) a random intercept for the student's school in third grade. In addition to these six models, in third grade we analyzed MAZE as a predictor using the same dependent variables and random effects structure as the other models. Finally, again in only third grade, we constructed models using MAZE and Retell as predictors (with no interaction terms), and the same dependent variables (i.e., Reading Informational Text and Reading Literature), and another set of models using MAZE and Fluency as predictors. In total, we fit 24 models. All models were fit using maximum-likelihood estimation using the Nelder-Mead method.

To construct 66% prediction intervals, we used the merTools package (Knowles & Frederick, 2024). The procedure involves supplying new data and the model of interest, extracting the fixed and random coefficients from the model, taking 100 draws from the multivariate normal distribution of the fixed and random coefficients, calculating the linear predictor based on these draws, and incorporating residual variation. Because prediction intervals differ by cluster (i.e., school when student took the Acadience reading test, and the school when the student took the RISE), there is no single prediction interval width as there would be for a regular linear regression. Therefore, to give a single prediction interval width, we constructed a data frame that consisted of every combination of (1) RISE school, Acadience school, and the mean for the predictor at that grade and obtained a prediction interval for every single one of these combinations. The result of this procedure was a data frame with each row having the (1) predicted RISE score, (2) lower bound of the 66% prediction interval, and (3) upper bound of the 66% prediction interval, with each row representing a unique combination of Acadience school, RISE/third grade school, and the mean value of the predictor. Then, we took the difference between the upper bound and lower bound values of each row, and calculated the average of these differences to achieve the 66% prediction interval. Therefore, the prediction interval widths we reported are the average prediction interval widths for the mean score on the predictor of interest.

To calculate R^2, we used the MuMIn package (Barton, 2024). This function calculates a marginal R^2, which represents the variance explained by fixed effects, and a conditional R^2, which represents the variance explained by the entire model (i.e., fixed and random effects). The marginal R^2 is defined as:

$$\frac{\sigma_f^2}{\sigma_f^2 + \sigma_\alpha^2 + \sigma_\epsilon^2}$$

In this equation, σ_f^2 represents the variance of the fixed effects, σ_α^2 represents the variance of the random effects, and σ_ϵ^2 represents the observation-level variance. The approach is derived from the methods outlined in Nakagawa and Schielzeth (2013), Johnson (2014), and Nakagawa and colleagues (2017).

Appendix B: Acadience Subscores' Predictive Validity by Demographics Technical Details

All analyses were conducted using R (version 4.4.0), the statistical programming language (R Core Team, 2024). For each of our multilevel models, we used the lme4 package (Bates et al., 2015). For each grade, we had three predictors (i.e., Accuracy, Fluency, and Retell), two outcomes (i.e., Reading Literature and Reading Informational Text), and five demographic variables (i.e., Gender, Student Economic Status, Student English Learner Status, Student Disability Status, and Race/Ethnicity). Each model had one outcome, one predictor, and one demographic variable. The multilevel models consisted of a main effect of the predictor, a main effect of the demographic variable, and an interaction term between the predictor and demographic variable. To account for clustering, two random effects were entered into each model: (1) A random intercept for the student's school during that grade, and (2) a random intercept for the student's school in third grade. In third grade, in addition to Accuracy, Retell, and Fluency, we analyzed MAZE's interaction with each of the demographic variables. To determine whether the interaction term was practically significant, we obtained standardized coefficients using the parameters package (Lüdecke et al., 2020). This package calculates the standardized parameters by refitting the model using the same data with the variables standardized (i.e., subtracting the mean and dividing by the standard deviation for each observation). We used 0.10 standard deviations as the threshold for practical significance based on our experience in education research.

To obtain the R^2 for each level within a demographic variable (e.g., for the gender variable, separate R^2 for boys and girls), we filtered the data to only include that group (e.g., boys) and reran the model without the interaction term or the main effect of the demographic variable. Then, we used the MuMIn package (Barton, 2024) to calculate R^2. This function calculates a marginal R^2, which represents the variance explained by fixed effects, and a conditional R^2, which represents the variance explained by the entire model (i.e., fixed and random effects). The marginal R^2 is defined as:

$$\frac{\sigma_f^2}{\sigma_f^2 + \sigma_\alpha^2 + \sigma_\epsilon^2}$$

In this equation, σ_f^2 represents the variance of the fixed effects, σ_α^2 represents the variance of the random effects, and σ_ϵ^2 represents the observation-level variance. The approach is derived from the methods outlined in Nakagawa and Schielzeth (2013), Johnson (2014), and Nakagawa and colleagues (2017).

For our combination models (i.e., MAZE and Retell and MAZE and Fluency), we did not run any interaction analyses because there would be no single interaction that would demonstrate a difference in the predictive ability of the model between groups. Instead, for these models, we only evaluated the separate R^2 for each group.

To calculate residuals for individual groups, for each of our original single predictor models (described in Appendix A), we obtained the residuals from the model. Then, using only the residuals that corresponded to that group (e.g., Male students), we calculated the mean residual. We converted this to units of standard deviation by dividing by the standard deviation of the dependent variable from the model. So if the dependent variable was RISE Reading Informational Text in first grade, the mean residual was divided by the standard deviation of Reading Informational Text scores in first grade for all students. To calculate the R^2 for each group using the models with all students included (not separate), we manually calculated the sum of squares residuals, and the total sums of squares separately by group using the residuals derived from the models including all students. We then divided the sum of squares residuals by the total sums of squares to get the R^2.

Appendix C: Full Demographic Interaction Results Table

Grade	Predictor	Moderator	Outcome	β	*p*-value
First	Fluency	Gender (Female - Male)	Literature	-0.01	0.035
			Informational Text	< 0.01	0.471
		Low Income (No - Yes)	Literature	0.06	< 0.001
			Informational Text	< 0.01	0.681
		ELL Status (No - Yes)	Literature	-0.01	0.508
			Informational Text	-0.10	< 0.001
		Disability Status (No - Yes)	Literature	0.12	< 0.001
			Informational Text	0.07	< 0.001
		Hispanic	Literature	0.05	< 0.001
		Black	Literature	0.01	0.782
		Asian	Literature	0.01	0.519
		Native American	Literature	0.04	0.271
		Multiracial	Literature	< 0.01	0.903
		Pacific Islander	Literature	0.01	0.587
		Hispanic	Informational Text	-0.02	0.017
		Black	Informational Text	-0.08	0.012
		Asian	Informational Text	-0.03	0.216
		Native American	Informational Text	< 0.01	0.945
		Multiracial	Informational Text	-0.01	0.507
		Pacific Islander	Informational Text	-0.08	0.002
	Accuracy	Gender (Female - Male)	Literature	-0.07	< 0.001
			Informational Text	-0.05	< 0.001
		Low Income (No - Yes)	Literature	-0.14	< 0.001
			Informational Text	-0.18	< 0.001
		ELL Status (No - Yes)	Literature	-0.20	< 0.001
			Informational Text	-0.22	< 0.001
		Disability Status (No - Yes)	Literature	-0.15	< 0.001
			Informational Text	-0.16	< 0.001
		Hispanic	Literature	-0.15	< 0.001
		Black	Literature	-0.16	< 0.001
		Asian	Literature	-0.05	0.048
		Native American	Literature	-0.16	< 0.001
		Multiracial	Literature	-0.03	0.089
		Pacific Islander	Literature	-0.17	< 0.001
		Hispanic	Informational Text	-0.17	< 0.001
		Black	Informational Text	-0.21	< 0.001
		Asian	Informational Text	-0.07	0.005
		Native American	Informational Text	-0.18	< 0.001
		Multiracial	Informational Text	-0.05	0.004
		Pacific Islander	Informational Text	-0.20	< 0.001
	Retell	Gender (Female - Male)	Literature	-0.01	0.072
			Informational Text	< 0.01	0.962

Grade	Predictor	Moderator	Outcome	β	*p*-value
		Low Income (No - Yes)	Literature	0.07	< 0.001
			Informational Text	0.02	0.002
		ELL Status (No - Yes)	Literature	0.02	0.199
			Informational Text	-0.07	< 0.001
		Disability Status (No - Yes)	Literature	0.19	< 0.001
			Informational Text	0.14	< 0.001
		Hispanic	Literature	0.07	< 0.001
		Black	Literature	0.03	0.306
		Asian	Literature	0.01	0.608
		Native American	Literature	0.04	0.319
		Multiracial	Literature	0.01	0.460
		Pacific Islander	Literature	0.03	0.301
		Hispanic	Informational Text	0.01	0.320
		Black	Informational Text	-0.02	0.485
		Asian	Informational Text	< 0.01	0.899
		Native American	Informational Text	0.02	0.672
		Multiracial	Informational Text	-0.01	0.713
		Pacific Islander	Informational Text	-0.06	0.046
Second	Fluency	Gender (Female - Male)	Literature	-0.01	0.008
			Informational Text	< 0.01	0.822
		Low Income (No - Yes)	Literature	< 0.01	0.355
			Informational Text	-0.05	< 0.001
		ELL Status (No - Yes)	Informational Text	-0.15	< 0.001
			Literature	-0.07	< 0.001
		Disability Status (No - Yes)	Informational Text	-0.03	< 0.001
			Literature	0.03	< 0.001
		Hispanic	Literature	-0.02	0.006
		Black	Literature	-0.03	0.109
		Asian	Literature	0.03	0.094
		Native American	Literature	-0.04	0.118
		Multiracial	Literature	< 0.01	0.782
		Pacific Islander	Literature	-0.01	0.544
		Hispanic	Informational Text	-0.08	< 0.001
		Black	Informational Text	-0.11	< 0.001
		Asian	Informational Text	0.01	0.488
		Native American	Informational Text	-0.10	< 0.001
		Multiracial	Informational Text	-0.04	0.003
		Pacific Islander	Informational Text	-0.12	< 0.001
	Accuracy	Gender (Female - Male)	Literature	-0.08	< 0.001
			Informational Text	-0.05	< 0.001
		Low Income (No - Yes)	Literature	-0.21	< 0.001
			Informational Text	-0.20	< 0.001
		ELL Status (No - Yes)	Literature	-0.21	< 0.001
			Informational Text	-0.21	< 0.001

Grade	Predictor	Moderator	Outcome	β	*p*-value
		Disability Status (No - Yes)	Literature	-0.26	< 0.001
			Informational Text	-0.25	< 0.001
		Hispanic	Literature	-0.18	< 0.001
		Black	Literature	-0.19	< 0.001
		Asian	Literature	-0.05	0.023
		Native American	Literature	-0.21	< 0.001
		Multiracial	Literature	-0.04	0.017
		Pacific Islander	Literature	-0.19	< 0.001
		Hispanic	Informational Text	-0.18	< 0.001
		Black	Informational Text	-0.21	< 0.001
		Asian	Informational Text	-0.05	0.029
		Native American	Informational Text	-0.20	< 0.001
		Multiracial	Informational Text	-0.07	< 0.001
		Pacific Islander	Informational Text	-0.21	< 0.001
	Retell	Gender (Female - Male)	Literature	-0.01	0.168
			Informational Text	< 0.01	0.583
		Low Income (No - Yes)	Literature	0.04	< 0.001
			Informational Text	< 0.01	0.962
		ELL Status (No - Yes)	Literature	-0.02	0.014
			Informational Text	-0.08	< 0.001
		Disability Status (No - Yes)	Literature	0.13	< 0.001
			Informational Text	0.08	< 0.001
		Hispanic	Literature	0.01	0.046
		Black	Literature	0.01	0.584
		Asian	Literature	0.05	0.018
		Native American	Literature	-0.03	0.388
		Multiracial	Literature	-0.01	0.704
		Pacific Islander	Literature	-0.04	0.078
		Hispanic	Informational Text	-0.02	0.001
		Black	Informational Text	-0.04	0.103
		Asian	Informational Text	< 0.01	0.989
		Native American	Informational Text	-0.03	0.312
		Multiracial	Informational Text	-0.05	0.002
		Pacific Islander	Informational Text	-0.12	< 0.001
Third	Fluency	Gender (Female - Male)	Literature	-0.02	< 0.001
			Informational Text	-0.01	0.062
		Low Income (No - Yes)	Literature	-0.01	0.050
			Informational Text	-0.06	< 0.001
		ELL Status (No - Yes)	Literature	-0.08	< 0.001
			Informational Text	-0.16	< 0.001
		Disability Status (No - Yes)	Literature	-0.01	0.006
			Informational Text	-0.06	< 0.001
		Hispanic	Literature	-0.02	< 0.001
		Black	Literature	-0.06	0.001

Grade	Predictor	Moderator	Outcome	β	*p*-value
		Asian	Literature	0.04	0.006
		Native American	Literature	-0.05	0.030
		Multiracial	Literature	-0.01	0.606
		Pacific Islander	Literature	-0.03	0.080
		Hispanic	Informational Text	-0.09	< 0.001
		Black	Informational Text	-0.13	< 0.001
		Asian	Informational Text	0.01	0.697
		Native American	Informational Text	-0.09	< 0.001
		Multiracial	Informational Text	-0.04	0.002
		Pacific Islander	Informational Text	-0.13	< 0.001
	Accuracy	Gender (Female - Male)	Literature	-0.10	< 0.001
			Informational Text	-0.06	< 0.001
		Low Income (No - Yes)	Literature	-0.23	< 0.001
			Informational Text	-0.23	< 0.001
		ELL Status (No - Yes)	Literature	-0.21	< 0.001
			Informational Text	-0.21	< 0.001
		Disability Status (No - Yes)	Literature	-0.29	< 0.001
			Informational Text	-0.27	< 0.001
		Hispanic	Literature	-0.19	< 0.001
		Black	Literature	-0.22	< 0.001
		Asian	Literature	-0.02	0.298
		Native American	Literature	-0.24	< 0.001
		Multiracial	Literature	-0.05	0.001
		Pacific Islander	Literature	-0.19	< 0.001
		Hispanic	Informational Text	-0.19	< 0.001
		Black	Informational Text	-0.22	< 0.001
		Asian	Informational Text	-0.02	0.479
		Native American	Informational Text	-0.21	< 0.001
		Multiracial	Informational Text	-0.06	< 0.001
		Pacific Islander	Informational Text	-0.21	< 0.001
	Retell	Gender (Female - Male)	Literature	-0.01	0.048
			Informational Text	< 0.01	0.806
		Low Income (No - Yes)	Literature	0.05	< 0.001
			Informational Text	< 0.01	0.853
		ELL Status (No - Yes)	Literature	-0.01	0.267
			Informational Text	-0.09	< 0.001
		Disability Status (No - Yes)	Literature	0.12	< 0.001
			Informational Text	0.07	< 0.001
		Hispanic	Literature	0.03	< 0.001
		Black	Literature	0.01	0.717
		Asian	Literature	0.01	0.517
		Native American	Literature	0.01	0.835
		Multiracial	Literature	-0.02	0.079
		Pacific Islander	Literature	-0.02	0.215

Grade	Predictor	Moderator	Outcome	β	*p*-value
		Hispanic	Informational Text	-0.03	< 0.001
		Black	Informational Text	-0.04	0.063
		Asian	Informational Text	-0.03	0.091
		Native American	Informational Text	-0.01	0.765
		Multiracial	Informational Text	-0.04	0.004
		Pacific Islander	Informational Text	-0.11	< 0.001
	MAZE	Gender (Female - Male)	Literature	-0.01	0.045
			Informational Text	< 0.01	0.610
		Low Income (No - Yes)	Literature	0.03	< 0.001
			Informational Text	-0.02	< 0.001
		ELL Status (No - Yes)	Literature	-0.04	< 0.001
			Informational Text	-0.13	< 0.001
		Disability Status (No - Yes)	Literature	0.05	< 0.001
			Informational Text	0.01	0.343
		Hispanic	Literature	0.01	0.083
		Black	Literature	< 0.01	0.845
		Asian	Literature	-0.01	0.425
		Native American	Literature	-0.03	0.234
		Multiracial	Literature	-0.02	0.054
		Pacific Islander	Literature	< 0.01	0.794
		Hispanic	Informational Text	-0.05	< 0.001
		Black	Informational Text	-0.07	0.002
		Asian	Informational Text	-0.03	0.090
		Native American	Informational Text	-0.06	0.020
		Multiracial	Informational Text	-0.04	0.001
		Pacific Islander	Informational Text	-0.11	< 0.001

Appendix D: Reading on Grade Level Cut Scores Technical Details

All cut score analyses were done using R (version 4.4.0), the statistical programming language (R Core Team, 2024). To determine the optimal cut point for each grade (first, second, and third), we used the cutpointr package (Thiele & Hirschfeld, 2021). We evaluated three different optimal cut points for each grade: (1) maximize accuracy, (2) minimize cost, and treat false positives as twice as costly as false negatives, and (3) minimize cost, and treat false negatives as twice as costly as false positives. The traditional method to determine optimal cut scores is to construct a data set consisting of all possible cut scores, calculating the performance metric (e.g., accuracy) for each of these cut scores, and then selecting the cut score with the best value of the chosen metric. However, research shows that this approach is prone to be not replicable between samples. To account for this, we used cutpointr's bootstrapping procedure. Specifically, the function takes a sample that is drawn with replacement (i.e., an observation can be drawn more than once) that is the same size as the original data set. On average, 63.2% of all observations of the original data are within a bootstrap sample (Efron & Tibshirani, 1997). On this sample (i.e., called the in-bag sample), the cut point is estimated using the traditional approach. This process is repeated 100 times (i.e., draw sample with replacement, calculate cut point). Then, the optimal cut point is calculated as the average of the 100 cut points from the in-bag samples.

To evaluate whether cut scores differed between pre-pandemic and post-pandemic data, we used a permutation test procedure for every grade. This procedure involves randomly assigning third grade year to the entire data set in the same proportion that exists in the original data. For example, an observation may have attended third grade in 2021, but they were randomly assigned 2022. This is done for every observation, but in the end the same proportion of rows are assigned 2022, 2021, and so on, as existed in the original data. After randomly assigning years, the optimal cut point is calculated using the traditional approach for both pre-pandemic and post-pandemic years (reminder, the years are assigned randomly). Then, we calculated the difference between the two cut scores. This process is then repeated 10,000 times, resulting in a data frame with 10,000 values for the difference between the pre-pandemic and post-pandemic cut scores. This procedure essentially provides a null hypothesis. That is, a distribution of differences in cut scores that we would expect, assuming there are no true differences in the cut scores between pre-pandemic and post-pandemic data. Then, using the real data, we calculated the pre-pandemic and post-pandemic cut scores using the bootstrapping procedure described above, took their difference, and compared the difference to our distribution of differences. If the true difference in cut scores was equal to or below the 2.5 percentile value for the difference or equal to or above the 97.5 percentile value for the difference, then this indicated that the cut score was significant and that the pre-pandemic and post-pandemic cut score truly differ from each other.