Combining Early Grade Assessments to Study Literacy Skills: Addressing the Variability in Tests Taken across Schools and Students

Elaine Allensworth
University of Chicago Consortium on School Research

Alex Gordon
University of Chicago Consortium on School Research

Christopher Young
University of Chicago Consortium on School Research

There is considerable variability in the literacy assessments taken in Kindergarten through second grade, across schools and between multilingual learners and other students, and within students over time. This makes it difficult to study changes in students' acquisition of ELA skills in these formative years, or to evaluate policies and practices meant to support literacy development. Here we examine several popular early grade assessments—the MAP, ACCESS, DIBELS, TRC English & Spanish versions, and apply a novel approach to combining information to develop latent scores of students' literacy development. We find each assessment provides information that is predictive of students' development towards third grade literacy outcomes (ELA grades and state assessment scores), with different strengths and weaknesses, and considerable overlap among them. We further provide evidence of strong predictive validity for the combined scale, even in post-COVID-19 years, suggesting that we could leverage existing assessment information to produce metrics for studying school, district, and state policies and practices around literacy development.

VERSION: January 2025

Suggested citation: Allensworth, Elaine, Alex Gordon, and Christopher Young. (2025). Combining Early Grade Assessments to Study Literacy Skills: Addressing the Variability in Tests Taken across Schools and Students. (EdWorkingPaper: 25 -1133). Retrieved from Annenberg Institute at Brown University: https://doi.org/10.26300/4s34-7648

Combining Early Grade Assessments to Study Literacy Skills: Addressing the Variability in Tests Taken across Schools and Students

Elaine Allensworth, Alex Gordon, and Christopher Young

University of Chicago Consortium on School Research

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through R305X220033 to the University of Chicago. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education. We greatly appreciate the partnership of the Chicago Public Schools in this study, and are particularly grateful to the Director of Literacy, Jane Fleming, for her insights and guidance on this analysis and the broader study.

Combining Early Grade Assessments to Study Literacy Skills: Addressing the Variability in Tests Taken across Schools and Students

ABSTRACT

There is considerable variability in the literacy assessments taken in Kindergarten through second grade, across schools and between multilingual learners and other students, and within students over time. This makes it difficult to study changes in students' acquisition of ELA skills in these formative years, or to evaluate policies and practices meant to support literacy development. Here we examine several popular early grade assessments—the MAP, ACCESS, DIBELS, TRC English & Spanish versions, and apply a novel approach to combining information to develop latent scores of students' literacy development. We find each assessment provides information that is predictive of students' development towards third grade literacy outcomes (ELA grades and state assessment scores), with different strengths and weaknesses, and considerable overlap among them. We further provide evidence of strong predictive validity for the combined scale, even in post-COVID-19 years, suggesting that we could leverage existing assessment information to produce metrics for studying school, district, and state policies and practices around literacy development.

Combining Early Grade Assessments to Study Literacy Skills: Addressing the Variability in Tests Taken across Schools and Students

Federal, state, and district-mandated assessments often begin at grade three, resulting in statewide uniformity. There is much less consistency in the assessments used in the earlier grades. This makes it difficult to determine the progress of learning gains in the early grades in districts or states that do not use consistent assessments. An additional complication for assessing literacy growth in the early grades is that students designated as multilingual learners (MLLs) take different assessments than other students until they reach a specific English Language proficiency level, or they reach a grade level which mandates the same assessment for all students. They then take the same assessments as other students, introducing new groups of students into grade-level averages on tests, and no longer take the tests used for MLLs. This makes it complex to assess either aggregate or individual student growth across grade levels. It also introduces selection bias into the interpretation of trend and subgroup data since decisions about whether students continue to take tests designed for MLLs versus other literacy assessments are often based on whether students meet a proficiency level on the test itself.

In this study, we examine the degree to which the different types of assessments taken in grades K-2 provide similar and different information in terms of the acquisition of early literacy skills and can be used to develop a common metric. We use data from students in kindergarten through grade two in the Chicago Public Schools from school year 2013-14 through 2021-22, examining relationships with later outcomes before and after the COVID-19 pandemic. We show relationships among the different assessments and present a novel approach to linking scores using test records available in administrative data to create latent scores of literacy achievement. As discussed below, this is different from traditional test equating, yet it has advantages in the current context in which there are many different K-2 assessments being used in schools. We include assessments of literacy skills for students for whom English is and is not the primary language spoken at home, since all early grade students are learning foundational skills for reading and reading comprehension in English, even if they are simultaneously developing language skills in a different language. In Chicago, about a third of students were considered MLLs at some point in school. This makes it critical to include their literacy growth in district-wide trends.

This study provides a structure and potential methods for researchers in district or state data offices who are interested in combining scores when students take different tests in different grades or different schools. It provides information that should be of interest to early-grade educators on how various popular assessments align with each other, and how each are related to ELA achievement in third grade achievement, including scores on the state assessment (the PARCC/IAR), NWEA's MAP, and students' ELA grades. It provides insight into how various K-2 assessments used in schools are related to each other, and in what ways they differ, including tests for students classified as multilingual learners.

The Context

The youngest learners saw the largest declines in school-based instructional time with the COVID-19 pandemic. In Chicago, preschool enrollment rates declined from over 29,000 students to just under 12,000 students, and school absence rates in Kindergarten through second grade increased from 5 to 6 percent in pre-pandemic years to 9 to 11 percent in the 2020-21 and 2021-22 school years (Gwynne, Allensworth & Liang, 2022). Thus, the youngest learners had considerable interrupted learning. Information on the impact of the pandemic on the youngest learners is crucial for studying recovery, to identify which schools showed the biggest losses, and which strategies are helping students get back to expected literacy levels. However, schools use an array of different assessments in Kindergarten, first, and second grade, making it difficult to compare post-pandemic performance to pre-pandemic years, or pandemic-era trends across schools.

In Chicago, in the years prior to the pandemic (before Spring 2020), all schools used the MAP assessment in spring of grade two and many schools also administered the MAP at earlier grade levels, beginning in Kindergarten. However, other schools used DIBELS and TRC to assess students' literacy skills in those grade levels. Multilingual learners took the ACCESS every year from fall of Kindergarten until their scores reached the level considered English proficient. All of these assessments are used to measure attainment of early skills in English-language arts, including early literacy skills, language acquisition, and comprehension, but they have different designs. They are not directly comparable to each other, and are used for different purposes within the classroom:

- **NWEA's Measures of Academic Progress (MAP)** assess general knowledge in reading including foundational skills (phonics, phonological awareness), writing, language, reading literature and informational text, speaking and listening (comprehension, vocabulary). It is computer-adaptive so that students answer questions that match their skill level, and it is designed to measure growth (NWEA, 2019). Most schools using the MAP in Chicago administered tests in the fall, winter, and spring. The district required the MAP at all schools in the spring for second to eighth grade, until spring of 2020.

- **WIDA's ACCESS for ELLs English Language Proficiency Test** assesses the developing English language proficiency of MLLs in four language domains: Listening, Reading, Writing, and Speaking. The test content is informed by standards in communication around academic, social and instructional purposes in language arts, mathematics, science, and social science. The WIDA English language development standards correspond to the academic language used in state academic content standards with six levels indicating developing English language proficiency. Proficiency levels are defined based on performance at the word/phrase level, the language forms and conventions level, and linguistic complexity and the discourse level. Administration is through a multistage adaptive design that tailors questions based on students' success with earlier questions and is primarily used to measure progress towards English proficiency (WIDA, 2022). Students take the ACCESS in the winter each year. If their score meets the standard for proficiency, they are no longer considered MLLs and do not take the ACCESS in subsequent years.
- **The Text Reading and Comprehension (TRC) assessment** measures early literacy skills including reading fluency, accuracy, and comprehension. Teachers make running records of student reading and assess student comprehension using a series of leveled books. It is often administered along with DIBELS and together they form the mCLASS®: Reading 3D™ assessment system. TRC is available in both English and Spanish. The

Spanish version is intended for MLLs whose primary language of instruction is Spanish. Amplify revised the Spanish texts used in the assessment in 2019-20 (National Center on Intensive Intervention at AIR (n.d.); Amplify, 2019). Most schools using the TRC in Chicago administered tests in the fall, winter, and spring.

- **Dynamic Indicators of Basic Early Literacy Skills (DIBELS)** assesses early literacy skills through a number of subtests that differ across grades. CPS used DIBELS NEXT which assesses phonemic awareness, phonics, accuracy and fluency in reading, vocabulary and language skills, and reading comprehension. Some of the DIBELS subtests are very specific to letter and word sounds in English. The tests are not designed to directly measure growth over time across grades; benchmarks are aligned across grade levels but average scores are not (Good & Kaminski, 2010). Most schools using the DIBELS in Chicago administered tests in the fall, winter, and spring.

There are both substantive and technical questions to consider before we can incorporate data from these different assessments into a common scale. Substantively, we must ask whether the assessments can form a coherent scale, given a wide array of skills on which students might be assessed in the area of literacy. Technically, we must consider how to develop a common scale across the many different types of assessments.

The potentially-varying content of early literacy assessments. Literacy growth requires the development of component skills (e.g., letter recognition, phonemic awareness, phonics, automaticity of word reading), along with the ability to construct meaning from language based on vocabulary and knowledge of the world, listening and text comprehension, and thinking and reasoning skills (Foorman et al., 2016; National Reading Panel, 2000; Snow, 2006). The development of foundational literacy skills and the development of language are interrelated; for example, phonological awareness may be facilitated by greater vocabulary (allowing for a comparison of sounds from a greater number of words), and reading fluency depends upon knowledge of the subject being read. Thus, literacy development includes instruction in foundational component skills, along with language-rich discussion on wide-ranging topics through listening, drawing, writing, and play (Goswami, 2001; Snow, 2006). The same foundational components and language skills are required for multilingual learners' English literacy development as for students whose home language is English, but with a need for heightened attention to their oral proficiency in English. Stronger development of oral proficiency and foundational component skills in students' home language can also support literacy development in English (August & Shanahan, 2017).

The assessments used to measure early literacy development are not designed to capture the exact same skills, but they each capture elements of literacy development necessary for becoming capable readers, with substantial overlap among them. In kindergarten and first grade, DIBELS emphasizes foundational skill development – letter naming, phonemic awareness, and word reading fluency, with oral reading fluency captured in first grade, and reading comprehension in second grade (University of Oregon, 2023). The greatest contrast from DIBELS may be with ACCESS, which emphasizes English comprehension skills–vocabulary and language development, and the ability to make meaning from oral and written text (WIDA, 2022). TRC emphasizes reading fluency, accuracy, and comprehension, as well as early print concepts and reading behaviors in pre-readers, whether in English or in Spanish, depending on the version (Chicago Public Schools, 2019). NWEA's MAP assesses a wide range of skills, including vocabulary and word structure, phonological awareness, phonics, concepts of print, and oral and written comprehension (NWEA, 2011 & 2019). Because development of literacy component skills and language skills are potentially interrelated and mutually reinforcing, we would expect that students with stronger skills in some areas would also have stronger skills in other areas, or would be more likely to develop those skills in later years than students with weaker skills, regardless of the emphasis of the assessment. One of our first tasks is to understand the degree to which scores on these different assessments show evidence that they measure a similar latent construct of general literacy skills (e.g., show inter-correlations), or that they each provide information of the development of skills that lead to the same literacy outcomes (e.g., third grade ELA grades and test scores).

Methods for developing a common score. Often, when working with different assessments, researchers will standardize data at each grade, and then compare students' scores relative to other students in the grade. However, this method only works if there are similar distributions of student grade levels and scores on the different assessments. If the distribution of scores is very different on one assessment than another, the meaning of a standardized score on one assessment may differ from one assessment to another. Standardizing by grade level also prevents researchers from examining growth from one grade to the next, since standardizing by grade removes any vertical scaling–the mean becomes zero at each grade.

A more traditional approach to the problem of two tests with different scales would be to equate them. This requires that test samples be equivalent, or that there are items included in both tests that can be used to anchor the scores across the two tests. There are different methods for equating ranging from simpler (e.g., equipercentile matching, where the scores that correspond to the same relative abilities are made equivalent for conversion purposes, Angoff, 1971) to complex (smoothing raw data, estimating score probabilities, making continuous function to describe discrete probabilities, and then equipercentile equating and error estimation, von Davier, et al., 2004). The difficulty with any of these methods is that they require various equating assumptions to be met, require connections between each pair of tests, and importantly, they do not provide a theory about how to integrate information from more than one pair of assessments. We want to incorporate many different tests together.

We present an approach to developing a common scale using information from all the assessments together to produce a measure of latent literacy skill development. Prior to the pandemic, in many schools in Chicago, students took multiple assessments in the same quarter. There are thousands of instances in which students took different assessments in the same quarter, allowing for a comparison of their scores. Students also potentially took different types of assessments as they moved across grade levels. For example, students may have taken DIBELS and TRC in Kindergarten and first grade and then taken MAP in second grade. The district required all students to take the MAP in the spring of second grade prior to the pandemic, so almost all students took the MAP in the Spring of second grade, regardless of which assessments they took previously. This allows for a comparison of the score a student received on one assessment in an earlier grade to the score they received on a different assessment in a later grade. We use the overlap in the assessments taken by groups of students in different years to create a latent measure of students' literacy skills at each grade.

Research Questions

We begin by examining properties of each of the K-2 assessments to determine how they differ, and whether the scores capture information about students' literacy skills that is related between assessments. If they are very dissimilar in the construct(s) they measure or the ways in which they measure it, they may not form a coherent latent scale of literacy achievement. There are a number of basic requirements needed to establish that single assessments are measuring constructs that form a coherent underlying scale. Minimum correlations for components of scales are generally about 0.15 (Clark & Watson, 1995), though ideally much higher, e.g. 0.5. The assessments could also differ in the shape of their score distributions, with some assessments spreading out scores at lower- or higher-levels while others have floors or ceilings, requiring some transformation. Some tests are designed to measure growth over time (e.g., MAP) while others are not (e.g., DIBELS). Understanding how scores change as students move through the school year and from one grade to the next helps to guide decisions about the potential form of our model, for example, deciding whether to use non-parametric indicator variables versus a linear growth model. We ask:

1) In what ways are the scores from different K-2 ELA assessments similar and different?

a. How correlated are the assessments with each other?

b. How similar are the assessments in terms of standard deviations and skew?

c. How do average scores on the assessment change from fall of Kindergarten to spring of second grade on each of the assessments?

We then develop a model to combine the scores across assessments and create a latent score for each student in each quarter at each grade level. We assess the predictive validity of the scores produced by the latent model, relative to the predictiveness of the original assessments, using students' third grade ELA achievement: their ELA grades, scores on the fall MAP in third grade, and scores on the spring state assessment (the PARCC or IAR). If they form a coherent scale of literacy development we would expect the combined latent score to be as predictive or more predictive than the individual assessments since the combined scale should have less measurement error than each individual assessment. However, if one or more of the assessments captures skills that are different from the others and not as strongly associated with third grade ELA achievement, the combined latent scores might be less predictive than the individual assessments. Because we expect the COVID-19 pandemic to have caused a disruption in students' learning trajectories, we examine the predictive validity of the K-2 assessments separately for students who reached third grade before the pandemic, and those who would have experienced the pandemic in at least one year in grades K-2. We ask:

2) How predictive are the various assessment scores, and the combined latent scores, of students' third grade ELA achievement at each grade level, in pre-pandemic and post-pandemic years?

Data

We use all observed scores in the district in Kindergarten through second grade from Fall of the 2013-14 school year through Spring 2021-22 on the MAP, DIBELS, TRC-English, TRC-Spanish, or ACCESS to answer RQ1. The ACCESS was given once a year, while the other assessments could have been given up to three times a year. For RQ2, we also analyze third grade data on three indicators of ELA skills: 1) the fall-quarter MAP; 2) the spring state assessment, which was either the PARCC or the Illinois-modified version of the PARCC called the Illinois Assessment of Readiness (IAR); and 3) students' teacher-assigned grades on the literacy standards in reading and writing, given on a 5-point (A-F) scale, and averaged together.

Number of test observations on each assessment. Each test has thousands of observations in each quarter; Table 1 shows the number of observations of each test in each quarter in each grade in pre-pandemic years, and each post-pandemic year. The sum of testing observations in a given year is much higher than the number of students with any test observation (shown in the last column) because students often took more than one type of assessment in a given quarter. Table 2 shows the combinations of assessments taken by the same student in the same grade and quarter, sorted in order of the number of times that particular combination occurs in our data. It was most common for students to either take the combination of DIBELS and TRC-English (with 460,805 instances) or only the MAP (with 368,510 instances) in a given quarter. But there were 91,976 instances where a student took the MAP and TRC-English and DIBELS in the same quarter (and no other assessments), and 7,097 instances of all three plus ACCESS in the same quarter. There were also many different combinations of assessments that students took and as they moved from one grade to another. The large overlap in test-taking in a given quarter, as well as overlap in the types of tests students took as they moved from Kindergarten to second grade, allow us to see how the same student performed on different assessments, and calculate the relationships between the assessments at different grade levels and as students' skills grow over time.

There are fewer assessment observations in more recent years than earlier ones. District enrollment has been declining for over a decade, with the largest declines in the youngest grades. Also, testing was halted when the pandemic hit in Spring 2020, and the district stopped mandating the MAP. During the remote/hybrid year (2020-21), there was no ACCESS testing. Thus, there are fewer test observations per student in post-pandemic years.

Number of students used in the analysis. Table 3 shows the total number of students who enrolled for at least a portion of the year in grades K-3 from 2013-14 to 2021-22 in the district, and the proportion of those students with assessment records in each year. In pre-pandemic years, 82-86 percent of students in Kindergarten and first grade had test records on one of the K-2 assessments included in this study. Students could be missing data because they were not enrolled in CPS long enough to take an assessment (e.g., left before spring testing in a school that did not administer fall tests, or arrived after fall testing and left before winter testing), their school used a different literacy assessment for which central district records were not maintained (e.g., Benchmark Assessment System, Fountas & Pinnell, 2010), or their school did not use a standardized assessment in that grade level. We also removed observations during the data cleaning process, as discussed below. In grades two and three, over 96 percent of students had assessment records in pre-pandemic years because the MAP was required.

Testing rates declined dramatically with the COVID-19 pandemic, beginning in spring 2020. Most students in grades K-3 had data in the 2019-20 school year because they were tested in the fall or winter, before the pandemic hit, but few students had spring 2020 data records. In the 2020-21 school year, schooling was remote in the fall and winter and hybrid in the spring. About half of students in grades K-2 had assessment data (47to 51 percent of students in grades K-2). In the 2021-22 school year, when schools returned to in-person instruction, testing rates were at 75 to77 percent in each grade, which was five to seven percentage points lower in grades K and 1 than in pre-pandemic years, and about 20 percentage points lower in second grade. In 2021-22, 92 percent of third graders took the state assessment (the IAR) which we use to assess predictive validity of our latent achievement scores for grades K-2, but they no longer took the MAP. There were very modest differences in backgrounds of students with test data in post-pandemic years relative to pre-pandemic years, despite lower testing rates; Appendix Table A1 provides a missing data analysis.

Overall, there were 2,422,829 test record observations for 300,887 unique students in the study years included in the linking study (RQ1). A subset of these students who also had third grade data, and who were not enrolled in charter schools, were included for the analyses examining the relationships of the K-2 assessments and the latent scores with third grade outcomes (RQ2). Charter school students were not included for RQ2 analyses because the district does not maintain course grade data for charter schools. Table 4 shows the demographic composition of both samples, which were similar in terms of race, ethnicity, and economic status: about a third of the students were Black, about 45 percent were Latinx, 12 percent were white, and 5 percent were Asian; 30 percent were MLLs; and about 81 percent qualified for free or reduced price lunch. In the validation sample, the average third grade GPA was 2.82 (with a standard deviation of 0.87), and the average IAR/PARCC score was 728.7 (with a standard deviation of 43.10).

Data checking and transformation of assessment scores. For each assessment, we examined distributions of the scores across grade levels and quarters, looking for outliers and unexpected patterns. These informed the specific ways in which each assessment was incorporated into the analysis, and decisions about whether to transform the data. Details on decisions around specific assessments are available in the Appendix, while general properties across the assessments are discussed with RQ1. We did not identify any need for transforming MAP or ACCESS scores. These assessments use Rasch analysis to produce scores on an interval scale that are vertically aligned across grade levels (NWEA, 2019; WIDA, 2022). TRC had to be recoded into a numeric scale; the recoded scale had a strong linear relationship with the MAP (r=0.80), as described in the Appendix. We used the DIBELS composite score, and trimmed extreme positive outliers.

Standardizing the Data. We standardized the scores on each assessment before combining them in the model to put them on roughly the same scale using the full range of observed scores for that assessment from grades K through 3. By standardizing across grade levels we could identify the degree to which students' scores change as they move through the year and across grades, which would not be possible to discern if data were standardized by grade level and quarter. Because students were more likely to take some assessments at higher or lower grade levels than others, we weighted the data when standardizing so that the means and standard deviations were calculated as if the percentage of students in each grade level was the same in all four grades. There were at least several hundred observations on each assessment in each grade level K-3.

Standardizing each assessment did not make them directly comparable since it did not account for any differences in the skill levels of students who took one type of assessment versus the others, nor for differences in the conceptual constructs assessed. The statistical model described below was required for linking the data across assessments. However, it did allow us to make general comparisons across the assessments in terms of the structure of the data and the ways in which scores changed as students progressed through the early grades.

Methods

For RQ1 we used descriptive statistics and correlations to examine the relationships of the assessments with each other, the statistical properties of the assessments, and the ways in which the scores change across quarters and grade levels. To link the assessments and answer RQ2, we developed a model predicting a latent score based on all available assessment data. Our method of linking scores between different assessments is best described as a prediction model where the same student was administered multiple distinct tests at the same time, and across years. It differs from traditional equating studies in several ways.

First, it uses pre-existing test records rather than a researcher-designed administration of different assessments to group(s) of students in typical equating (Kolen & Brennan, 2014). This method allows for linking without making students take assessments they otherwise would not. In some districts, there might not be students of all skill levels who take common assessments. In our study, the range of literacy skills among students who took each of the assessments is broad, providing common support for comparing each of the assessments. For example, comparing scores of students who took the MAP and another assessment in the same quarter, scores on each assessment range from below -2 standard deviations on the MAP scale to over 2 standard deviations. The one exception is for observations on TRC-Spanish among students with the highest scores on MAP, TRC or DIBELS; the highest values on these assessments are just under two standard deviations on the standardized MAP scale for students with TRC-Spanish scores. This is discussed later with Figure 1.

Second, we leverage within-student growth across assessments to estimate relationships among the assessments. Students who take the same tests at one grade often take different tests at the next grade. Nearly all students in pre-pandemic years took the MAP in the second grade, allowing for a common assessment across almost all students who reached second grade before 2020.

A final way this method is different from traditional equating is that it produces latent scores for each student using multiple sources of information, rather than a symmetrical conversion between one pair of tests, which is a requirement of equating (Dorans, Moses& Eignor, 2010). Advantages and disadvantages are discussed further below.

Statistical Model. To combine the scores, we used a multilevel model with test observations from grades K-2 treated as repeated measures within individual students, see Equation 1. Analyses were performed with R Statistical Software (v4.2.2; R Core Team 2022), using the lme4 package (Bates, et al., 2015). We confirmed that we obtained similar results using Bryk and Raudenbush's HLM Scientific Software version 8.2 (Raudenbush & Congdon, 2021). All observed assessment scores were regressed on a set of dummy variables indicating which grade (Kindergarten, First or Second grade) and quarter the test was taken (fall, winter, spring). There was no intercept, so coefficients on the variables representing each grade and quarter represent the average score at that time point in pre-pandemic years. We used non-parametric (categorical) terms to capture the specific grade and point in the year that the test was taken rather than linear growth terms because of idiosyncratic growth rates across the assessments: scores on DIBELS are not designed to increase monotonically over time, and many assessments show a pattern of decline between spring of one school year and the following fall.

We included fixed effects for each assessment other than MAP to adjust for the average difference in standardized scores relative to the standardized MAP in each grade and quarter. This was necessary because similar standardized scores could represent different skill levels either because: 1) they were standardized based on different samples (i.e., students taking one assessment in a particular grade may systematically be lower-achieving than students taking another assessment), or 2) growth patterns on the assessments differ, with students showing larger gains at particular points on some assessments than others. We chose the MAP as the excluded group, from which deviations were assessed by dummy variables for each of the other assessments, because we had confidence in its scaling, because it was vertically-aligned across the grade levels, and because it was required for all students in the spring of second grade in pre-pandemic years. We included separate terms for post-pandemic years because of concern that the pandemic would change growth patterns in scores during those years.

In the full model, test observations at time (t) are nested in students (i): [1]

$$Score_{ti} = \pi_{1i}(GKFall)_{ti} + \pi_{2i}(GK\ Winter)_{ti} + \pi_{3i}(GKSpring)_{ti} + \pi_{4i}(G1Fall)_{ti} + \pi_{5i}(G1Winter)_{ti} + \pi_{6i}(G1Spring)_{ti} + \pi_{7i}(G2Fall)_{ti} + \pi_{8i}(G2Winter)_{ti} + \pi_{9i}(G2Spring)_{ti}$$

$$+ \sum_{k=10}^{16} \pi_{ki} \left(PostQuarter_{q,ti}\right) + \sum_{l=17}^{22} \pi_{li} \left(Grade_{g,ti}\ x\ PostYear_{y,ti}\right) +$$

$$\sum_{l=23}^{52} \pi_{li} \left(Grade_{g,ti}\ x\ Quarter_{q,ti}\ x\ Assessment_{a,ti}\right)$$

$$+ r_{1g,i} + r_{2g,i} + r_{3g,i} + e_{ti}$$

$$\ldots All\ other\ coefficients\ fixed\ across\ students$$

- π_1 through π_9 are the average standardized assessment scores in spring at each grade level in each quarter in pre-pandemic years. There is no intercept for the model.
- π_{10} through π_{16} are the difference in scores relative to pre-pandemic years for each post-pandemic quarter represented separately: Spring 2020; Fall, Winter, and Spring SY2021; Fall, Winter, and Spring SY2022.
- π_{17} through π_{22} are the interaction of grade (first or second grade) with post-pandemic year (2021, 2022), and Spring 2020. These coefficients show post-pandemic year differences for first and second graders relative to kindergarteners.
- π_{23} through π_{52} capture the difference in average scores on a particular assessment at a particular grade and quarter relative to the NWEA-MAP. Each coefficient represents a particular Grade (K, 1, 2) x Quarter (fall, winter, and spring) x Assessment (all assessments other than the NWEA-MAP) 3-way interaction. Their inclusion removes consistent differences observed between each assessment and the MAP from the residuals.
- $r_{1gi}, r_{2gi}, \text{and } r_{3gi}$ are random effects (EB residuals) for each student at each grade. They indicate the degree to which the student's scores at that grade level were different from those of other students, net of differences due to assessment time point (year or quarter) and type of assessment. The random effects, plus the fixed effects $\pi_1 - \pi_{22}$ are used to produce the latent scores.

This method does not translate scores on one assessment into MAP scores, but instead identifies an overall shift in scoring for a particular assessment for a particular quarter and grade based on the average differences in scores at a particular grade and quarter relative to the MAP, and to other assessments which are simultaneously also being compared to each other. This has disadvantages and advantages to a one-to-one equating. To the extent that there are differences across assessments in the meaning of a one standard deviation change at a particular grade level, this model does not adjust for those differences and put them on an existing scale. Instead, it shrinks differences towards the most reliable estimate based on the relationships of all the assessments with each other and students' past and future performance, creating a new scale. As discussed further below under RQ1, each of the assessments has some weaknesses for particular groups of students. For those assessments that provide more differentiation at low or high score values than other assessments, this method has an advantage in that it does not constrain that variation to match the assessment with less differentiation, which would occur with traditional equating. Instead, it averages out the score as represented on the different assessments taken by the student. This method also utilizes information from multiple assessments to come up with a most-informed latent score, rather than sequentially comparing just two assessments at a time. It is not clear how traditional equating would handle multiple assessments.

We produced a random effect for each student at each grade. The three random effects were allowed to covary to leverage relationships between assessments administered to students across grade levels in the linking process. Note that in R you must specify that the random terms co-vary, while it is the default in Scientific Software's HLM. Besides leveraging cross-grade information to further support the linking process, this also produces imputed latent scores for students who were not tested in a given quarter or grade, based on all of the other assessment data available for the student.

We ran the model in two ways, first with just data from pre-pandemic years, then with data from all years. We wanted to assess the relationships among assessments using data that were not affected by the disruption in schooling. When we ran the model with pandemic-era data we included dummy variables representing quarters in the pandemic years (Spring 2020 and each quarter of 2021 and 2022), along with interaction terms of post-pandemic years by grade level, to capture changes in learning in those years that might have been different by grade level. These terms also provide estimates of the effects of the pandemic on literacy scores. The coefficients on the different assessments were very similar in the two models. The only coefficient that changed by more than 0.03 points was the coefficient for TRC-Spanish.

Not only did the coefficients for TRC-Spanish change from pre-to post-policy, but we also found that the TRC-Spanish scores from pre-pandemic years were much less predictive of third grade outcomes than TRC-Spanish scores in post-pandemic years or the other assessments. We believe these results could have occurred either because of irregular implementation in pre-pandemic years, with the potential for selection bias in when the tests were used, or because of the change in the TRC-Spanish assessment that occurred between these two time periods. For this reason, and others described below, we decided that the TRC-Spanish scores in pre-pandemic years were not as reliable as the other assessments, and we removed all pre-pandemic TRC-Spanish scores from the final run of the model. Table A2 in the Appendix provides coefficients from the pre-pandemic models with and without TRC-Spanish scores.

Latent scores for each student in each grade were produced by combining a student's grade-specific random effect estimate with the appropriate grade, quarter, and year fixed effects coefficients and interactions. Coefficients representing specific tests were not included as this would re-introduce score differences across the assessments, reproducing students' original scores. The latent scores represent the best estimate of a student's ability on the MAP standardized scale using all information from all assessments a student took. In the case that a student left the district or otherwise had an inactive enrollment status, we did not produce a latent score for that student in that year, though Empirical Bayes estimates of every student's latent scores exist for each grade, regardless of whether the student was active, based on their performance in other grades and the observed correlations between grades.

We then assessed the predictive validity of the latent scores relative to the predictive validity of individual assessments by examining correlations of scores at each grade with students' third grade ELA grades, their fall third grade MAP score, and their spring score on the state ELA assessment (the IAR or the PARCC). We ran correlations separately for students with third grade data in pre-pandemic years from those with data in post-pandemic years in case disruption caused by the pandemic influenced the value of the early grade assessment data as an indicator of progression towards third grade literacy outcomes.

Results

RQ1. In what ways are the assessment scores similar and different from each other*?*

Relationships among the assessments. Table 5 shows correlations among each pair of assessments using observations from tests that were taken by the same students at the same point in time. All of the assessments were positively correlated with each other, surpassing the 0.50 threshold of a moderate relationship, and many are highly correlated, surpassing a 0.70

threshold of a strong relationship. The lowest correlations are between DIBELS and ACCESS (0.54) and DIBELS and TRC-Spanish (0.63). DIBELS tests phonemic awareness specifically in English, thus, it makes conceptual sense that it would have lower correlations with the two assessments that are designed for students living in households where a language other than English is spoken at home, and where students are likely simultaneously learning to read in multiple languages with different letter sounds. The assessment that is least correlated with the others overall is TRC-Spanish, which is correlated at about 0.65 with each of the other assessments except TRC-English, with which it is strongly correlated at 0.77. The MAP shows the strongest correlations with the other assessments, with correlations of 0.78 to 0.83 with each of the other assessments (ACCESS, DIBELS, and TRC-English), except TRC-Spanish, where the correlation is 0.65. In general, correlations of greater than 0.50 suggest that there is considerable overlap between constructs being measured, and the smallest correlations are much larger than the recommended correlations for producing a reliable scale without redundancy, 0.15 - 0.50 (Clark & Watson, 1995).

The relationships between assessments are not necessarily linear, despite being correlated, and some assessments provide more differentiation among particular groups of students at the low and high ends of the scales. Figure 1 provides scatterplots of each assessment with each of the other assessments in standardized units. The diagonal boxes provide the horizontal and vertical axis labels. For example, the box in Row 1, Column 5 shows the relationship between the ACCESS on the horizontal axis and the MAP on the vertical axis. The patterns are described below:

The assessments not designed for MLLs do not differentiate ELA skills for students with very low ACCESS scores (i.e., low English proficiency), but the same tests are strongly aligned with ACCESS scores among students with moderate-to-high ACCESS scores. The scatterplot in Row 1, Column 5 of MAP by ACCESS shows that for students with ACCESS scores below -1 s.d. (which corresponds to an ACCESS score of about 270), the slope between ACCESS and MAP is flat–all students with ACCESS scores below that point get a similar low score on the MAP. However, for students scoring above 270 on the ACCESS, there is a very strong relationship between the two assessments such that knowing a student's score on one test would provide a very good prediction of their score on the other test. The same general pattern can be seen with ACCESS and DIBELS (row 2, column 5) and ACCESS and TRC-English (row 3, column 5). TRC-Spanish (row 4, column 5) shows a similar pattern, but the transition point is lower, closer to -1.5 s.d. (200 on the ACCESS).

TRC differentiates students with very high achievement more than DIBELS or ACCESS, and to some extent MAP, on the skills assessed by the TRC. Column 3 shows the relationship of TRC-English scores (on the horizontal axis) with each of the other assessments (on the vertical axis). Each of the scatterplots (except for the one showing its relationship with TRC-Spanish) tends to flatten out on the right side of the figure; this means that as students get higher scores on TRC-English they do not necessarily get higher on the other assessment. This suggests TRC-English might capture specific high-level skills that are distinct from the other assessments.

Students who score at the bottom end of the DIBELS scale do not necessarily have very low scores on other assessments. Column 2 shows the relationship of DIBELS scores with each of the other assessments, with DIBELS scores represented on the horizontal axis. On the far left side of each figure are the dots representing students with very low DIBELS scores. These students have a wide range of scores on the other assessments, with the exception of TRC-English. This suggests other assessments capture skills that are not measured on DIBELS.

Properties of the assessments. Summary statistics on each assessment at each grade level and quarter are provided in Table 6. The unstandardized means and standard deviations are included in the table, but it is easiest to compare the assessments by using the statistics based on standardized scores, which are on the right side of the table. The standardized means show a fairly similar range across the grades from about a standard deviation below the mean in fall of Kindergarten to about a standard deviation above the mean in spring of second grade. Note that because the scores are not yet linked, differences across the assessments could exist either because of differences in assessment scoring, or because of differences in the sample of students taking a particular assessment in a particular quarter and grade.

Figure 2 graphs the average scores on each assessment, making it easier to see how the scores differ within and across each grade level. Scores on the MAP increase linearly by quarter within each grade, with setbacks from the spring of one grade to the fall of the next. TRC-English scores show a similar pattern, but scores increase at a lower rate from Kindergarten to first grade on the TRC-English assessment than first to second or second to third grade. In contrast, MAP and ACCESS scores increase more from Kindergarten to first grade than from first to second grade. What looks like a lower change in scores on ACCESS from first to second grade is at least partly an artifact of how the test is given–students who reach the benchmark score at a particular grade no longer take the ACCESS in the following year. TRC-Spanish scores show a similar pattern as TRC-English scores, but with smaller increases from the winter to spring quarters within a given grade level; this may be a selection artifact since it is given less frequently in the spring. Average DIBELS scores increase considerably from fall to winter in Kindergarten, then are fairly similar at each of the next five administration points, until rising again in winter and spring of second grade; DIBELS is not designed to measure growth. Differences in the ways that average scores change across the grade levels, especially DIBELS, led us to decide to use non-parametric (categorical) terms in the predictive model representing each grade level and quarter.

The standard deviations increase at older grade levels (see Table 6), with the exception of the ACCESS, with standard deviations of about 1.0 across the different assessments in second grade. The standard deviation of ACCESS is smaller at older grades, but students with high scores are not observed on the ACCESS once they reach proficiency. Only the TRC has a substantial skew; it is largest in fall of Kindergarten and dissipates at later quarters and grade levels. In the very early grades, many students receive very low TRC scores, but scores improve over time, forming a fairly distribution by winter of first grade.

RQ2. How predictive are the assessment scores, and the combined latent scores, of students' third grade ELA achievement?

The predictive model produces estimates of students' latent literacy achievement at each grade level, K-2. Coefficients and variance components from the final model that includes variables for post-pandemic years are shown in Table 7. Coefficients from models that only include pre-pandemic data are available in the Appendix. Each main effect grade term (Kindergarten, 1st Grade, 2nd Grade by quarter) represents the average standardized score for students in each grade and quarter in pre-pandemic years. These are followed by coefficients for each post-pandemic quarter that show the deviation for kindergarten scores in each pandemic quarter relative to pre-pandemic years. Following those are interaction terms of post-pandemic quarters times first or second grade which show the deviation of first and second grade scores from the deviation in kindergarten scores in each post-pandemic quarter. The interactions should be added to the main effect coefficients for each post-pandemic year to get the total deviation for first or second graders in a specific post-pandemic quarter.

Coefficients on the right of the table indicate the degree to which standardized scores on a specific assessment test in that quarter are higher or lower relative to MAP standardized scores in that quarter and year for students with the same latent literacy skills. These terms adjust for differences in the assessment scales, and differences in the samples of students who took the assessments on which the scores were standardized. For example, scale differences can be seen in the Kindergarten DIBELS coefficient; DIBELS scores tend to be much higher in kindergarten relative to the total scale of DIBELS scores than kindergarten MAP scores are to the total scale of MAP scores (see Figure 2). Thus, the coefficients on kindergarten DIBELS scores are large and positive in those quarters, indicating that students with the same latent score would have a higher standardized DIBELS score than a standardized MAP score in those quarters. Differences based on the sample of students that take each assessment can be seen in the large positive coefficients on TRC-Spanish. Average scores on TRC-Spanish relative to the range of TRC-Spanish scores show fairly similar patterns of growth by grade and quarters as the MAP, using the non-linked standardized scores (see Figure 2). But the positive coefficients at all grade levels and quarters suggest that the entire observed distribution of standardized scores on TRC-Spanish is at a lower ELA skill level in terms of latent achievement than corresponding standardized scores on the MAP. The same standardized score represents a higher latent value on the MAP than on TRC-Spanish.

Variance and covariance of latent scores at each grade. The variance components at the bottom of Table 7 provide the standard deviation of the latent scores at each grade level: 0.428 in kindergarten, 0.603 in first grade, and 0.751 in second grade. As observed with the raw assessment scores in Table 6, the standard deviations are larger among students in older grades than in younger grades. In general, the standard deviations of the latent scores are similar or smaller than the standard deviations of the individual assessments (Table 6). Combining data across assessments likely adjusts for random error on any one assessment, resulting in less variation overall, but potentially more accurate and reliable scores than with a single assessment. The latent scores are highly correlated from one grade to the next, with correlations of about 0.9 between sequential grades (Kindergarten and First Grade r = 0.90 and First Grade and Second Grade r=0.93), and 0.77 between kindergarten and second grade.

Predictions of third grade assessments in pre-pandemic years. Latent scores are highly predictive of third grade assessment scores in pre-pandemic years (see the first row of each grade level on the left side of Table 8), with correlations ranging from 0.67 (kindergarten latent score predicting third grade PARCC/IAR score) to 0.82 (second grade latent score predicting third grade MAP score) in pre-pandemic years. Almost all the individual assessments are moderately-to-highly predictive of third grade assessment scores, although slightly less predictive than the latent scores. MAP scores are the most predictive, with correlations ranging from 0.66 (kindergarten latent score predicting third grade PARCC/IAR score) to 0.82 (second grade latent score predicting third grade MAP score). DIBELS, TRC-English, and ACCESS have correlations with third grade assessments ranging from 0.52 (Kindergarten DIBELS predicting third grade PARCC/IAR) to 0.74 (second grade TRC-English predicting third grade MAP). Only TRC-Spanish does not show a relationship with third grade assessment scores in pre-pandemic years. There may be selection issues since fewer students took TRC-Spanish in pre-pandemic years. In addition, Amplify revised the Spanish texts used in the assessment in 2019-20, which may have improved its accuracy. For this reason, we did not include pre-pandemic TRC-Spanish observations in the final model.

Predictions of third grade ELA grades in pre-pandemic years. The latent scores in grades K to 2 are also highly correlated with students' ELA grades in third grade, ranging 0.52 in kindergarten to 0.63 in second grade. In general, standardized assessments and course grades tend to be correlated at around 0.50 (Brookhart et. al. 2016), so these correlations are in the higher range of what is typical. Most of the individual assessments also are predictive of students' third grade ELA grades. MAP scores show similar relationships with third grade ELA grades as the latent scores at each grade level. Kindergarten scores on DIBELS, TRC-English, and ACCESS are correlated at between 0.35 and 0.43 with third grade ELA grades, while second grade scores show correlations of 0.45 to 0.53. TRC-Spanish scores are not correlated with students' third grade ELA grades in the pre-pandemic cohorts.

Predictions of third grade outcomes for post-pandemic third graders. Post-pandemic, the correlations between K-2 literacy scores and third grade ELA outcomes are smaller than observed in cohorts of students that reached third grade before the pandemic hit, see the right side of Table 8. All of the post-pandemic third graders experienced the pandemic for at least one year in kindergarten through second grade, which likely disrupted the patterns of growth observed before the pandemic. At the same time, all of the individual assessments and the latent scores show correlations that range from 0.30 to 0.64 with third grade literacy grades, and from 0.47 to 0.81 with third grade IAR scores, with the exception of TRC-Spanish scores in kindergarten. Post-pandemic third graders would have taken kindergarten TRC-Spanish tests prior to the release of the revised texts in 2019-20 (since students would have been in kindergarten and first grade before 2019-20). There are much stronger correlations of TRC-Spanish scores in first and second grade with third grade literacy outcomes in the post-pandemic years. The latent scores continue to show moderately-strong relationships with third grade grades (0.42 to 0.47) and strong relationships with IAR (0.66 to 0.72), but the relationships are smaller, potentially as a result of pandemic disruption, but also potentially because latent scores are no longer based as heavily on the MAP as in pre-pandemic cohorts.

Overall, there is considerable evidence that this method produces latent scores from the multiple assessments that can be used to discern changes in literacy growth over time, despite the differences that exist. The assessments are all moderately-to-highly correlated with each other. The latent scores are highly predictive of third grade ELA outcomes. Even though correlations of K-2 assessments and latent scores with third grade outcomes are smaller among students who experienced the pandemic during grades K-2 than among students who did not experience the pandemic before third grade, the relationships of the latent scores with third grade outcomes remain moderately-large in size, suggesting they provide reliable information about students' development of skills that matter for third grade literacy outcomes.

Discussion, Limitations, and Future Research

The state of early-grade assessment systems, with schools choosing different assessments in different schools, and MLLs moving into and out of assessment systems as they gain English proficiency, makes it very difficult to study school, district or statewide trends in literacy development, or to examine the effects of new policies on early learning. While the different assessments each have unique features, and may focus on different components of literacy development, the skills that they measure are interrelated and necessary for all learners. Thus, while each assessment has some strengths and weaknesses for specific subgroups of students, there is also considerable overlap and commonality in the information they provide as a whole. Combining scores across the assessments seems feasible, with resulting scores providing at least as good of an indication of students' literacy progression as any of the individual assessments. This can be useful for the purposes of research. It can also be useful for the purpose of assessing progress in a school, district, or state in the early grades.

These results also suggest that districts could potentially combine information of multilingual learners and non-MLLs in the early years to get a comprehensive view of literacy growth, rather than one that is biased by students moving from one assessment system to another. The ACCESS provides better information on ELA progression for MLLs who have the lowest English proficiency skills (e.g., levels 1 and 2 on the ACCESS) than the assessments intended for students whose home language is English. Yet, there is a surprisingly strong relationship between ACCESS and the other assessments among most students taking the ACCESS, even though they have not reached grade-level proficiency. The ACCESS provides a perspective on students' ELA growth that is particularly useful for students whose home language is not English, capturing oral and verbal fluency, as well as academic skills, which are foundations for literacy for all students. Incorporating the progression of multilingual learners in school or district trends should be possible.

Limitations. This method does not produce scores that provide the specific information that may be available in individual assessment reports, cannot tell practitioners exactly which literacy skills students have, or whether there are specific areas that show stronger or weaker growth than others. It is a scale that measures general progression towards third grade literacy goals, utilizing whatever information is available in each of the assessments. It also depends on a sufficiently large and diverse base of students who took both tests. Not all districts would have overlap in tests taken at the same time, or among students with both high and low skills. Thus, they might lack common support to be able to develop a valid combined score for all students.

Future research. We plan to use latent scores to understand trends during school years affected by the pandemic and pandemic recovery, and the influence of different school practices on student academic recovery from pandemic-era setbacks. We further intend to investigate adjustments to the model which generates latent scores in the hopes of reducing its complexity without affecting its predictive capacity - in other words, testing other specifications which might produce a simpler model that produces scores with the same external validity in predicting third grade outcomes. Finally, we are trying different methods of incorporating a new assessment that was only given in post-pandemic years (iReady) into the creation of latent scores to study changes in scores over a longer post-pandemic period.

Tables and Figures

Table 1. Number of Students Taking Each Assessment Given in Grades K-2 in Each Year

School Year	Grade	Quarter	Access	DIBELS	MAP	TRC-English	TRC-Spanish	Any Test
2013-2014 through 2018-19	Grade K	Fall	-	67,203	28,172	74,907	15,109	117,540
	Grade K	Winter	6,298	64,074	29,119	77,182	15,559	119,803
	Grade K	Spring	-	57,574	27,458	63,739	12,720	129,692
	Grade 1	Fall	-	60,409	29,401	81,170	16,498	124,326
	Grade 1	Winter	49,269	54,946	33,957	81,598	16,185	128,436
	Grade 1	Spring	-	49,058	37,433	67,269	12,964	138,791
	Grade 2	Fall	-	50,927	70,490	83,121	14,789	142,400
	Grade 2	Winter	1,148	46,644	129,576	82,890	14,272	155,367
	Grade 2	Spring	-	42,877	157,543	65,940	10,874	166,967
2019-2020	Grade K	Fall	-	9,617	3,650	11,009	1,646	16,362
	Grade K	Winter	6,063	9,272	4,173	10,698	1,808	16,361
	Grade K	Spring	-	-	-	-	-	6,063
	Grade 1	Fall	-	8,073	3,729	11,416	1,897	16,779
	Grade 1	Winter	6,054	7,434	5,397	10,902	1,777	17,303
	Grade 1	Spring	-	-	-	-	-	6,054
	Grade 2	Fall	-	6,123	13,060	10,875	1,717	20,703
	Grade 2	Winter	6,716	5,704	18,576	10,362	1,583	21,720
	Grade 2	Spring	-	-	-	-	-	6,716
2020-2021	Grade K	Fall	-	6,969	-	4,050	435	7,803
	Grade K	Winter	-	6,320	-	7,674	1,427	9,758
	Grade K	Spring	-	6,451	254	8,587	1,668	10,701
	Grade 1	Fall	-	5,041	-	7,063	1,010	8,670
	Grade 1	Winter	-	5,903	-	9,318	1,560	10,994
	Grade 1	Spring	-	5,415	276	9,715	1,684	11,611
	Grade 2	Fall	-	3,291	-	6,670	999	7,792
	Grade 2	Winter	-	4,287	-	8,726	1,208	9,864
	Grade 2	Spring	-	4,335	268	8,896	1,330	10,333
2021-2022	Grade K	Fall	-	7,955	2,760	5,516	-	11,493
	Grade K	Winter	5,444	8,051	1,950	7,992	1,668	12,562
	Grade K	Spring	-	7,401	3,138	9,139	1,718	16,197
	Grade 1	Fall	-	8,105	3,478	9,028	-	13,365
	Grade 1	Winter	6,657	7,178	2,238	9,395	1,784	13,595
	Grade 1	Spring	-	6,085	3,595	9,747	1,879	17,144
	Grade 2	Fall	-	6,035	3,891	10,032	-	13,813
	Grade 2	Winter	5,625	5,834	2,393	9,846	1,561	13,482
	Grade 2	Spring		5,565	4,016	10,117	1,617	17,116

Table 2. Combinations of Assessments Taken in the Same quarter in Grades K-2

Test combination in a specific quarter	Number of Observations with the specific combination
Dibels TRC-English	460,805
MAP only	368,510
TRC-English only	210,498
TRC-Spanish only	101,564
MAP Dibels TRC-English	91,976
MAP TRC-English	74,580
Access only	61,385
Dibels only	39,233
MAP Access	33,909
Access TRC-Spanish	28,225
Dibels Access TRC-English	20,098
MAP Dibels	15,464
MAP TRC-Spanish	10,123
Access TRC-English	8,851
MAP Dibels Access TRC-English	7,097
MAP Access TRC-English	6,262
MAP Access TRC-Spanish	6,125
Dibels Access	5,559
Dibels TRC-English TRC-Spanish	4,800
TRC-English TRC-Spanish	4,345
MAP Dibels Access	2,503
MAP TRC-English TRC-Spanish	1,262
Access TRC-English TRC-Spanish	972
MAP Access TRC-English TRC-Spanish	909
Dibels Access TRC-English TRC-Spanish	893
MAP Dibels TRC-English TRC-Spanish	823
MAP Dibels Access TRC-English TRC-Spanish	418
Dibels TRC-Spanish	405
Other combinations	82

Note: This shows the number of times a particular test combination was taken by a student in the same quarter. Students are represented once for each quarter they participated in testing.

Table 3. Percentage of students with assessment data in each grade and year

Number of active students	2013-14	2014-15	2015-16	2016-17	2017-18	2018-19	2019-20	2020-21	2021-22
Kindergarten	31,762	30,257	28,911	27,355	26,284	25,395	25,805	23,071	22,713
1st grade	32,611	31,701	30,514	28,610	27,336	26,020	25,407	24,542	23,639
2nd grade	30,885	31,651	30,921	29,354	28,011	26,644	25,622	24,281	23,785
3rd grade	31,492	31,795	32,424	31,149	29,861	28,240	26,909	24,735	23,429
Percent of active students with any data in each grade on these assessments: MAP, DIBELS, TRC, ACCESS									
Kindergarten	84%	85%	85%	83%	82%	82%	78%	50%	75%
1st Grade	86%	86%	85%	83%	82%	83%	80%	51%	77%
2nd Grade	96%	98%	97%	97%	97%	97%	94%	47%	76%
3rd Grade	96%	98%	97%	97%	97%	97%	91%	0%	24%
3rd Grade % with PARCC/IAR	N/A	92%	94%	94%	93%	94%	N/A	51%	92%

Note: Any student who was actively enrolled when district enrollment snapshots were taken in September, January or May are included in these numbers, although some students would not have been actively enrolled at the time of testing. Illinois began administering the PARCC in the 2014-15 school year, previously it administered a different assessment (the ISAT). Lower testing rates in post-pandemic years led us to conduct a missing data analysis. Following the methods to test for balance with propensity score matching in Zhang et. al. (2018), and the criteria in Stuart, Lee, and Leacy (2013) we found that students with test data differed little from the population of students in terms of income, race, ethnicity, gender, attendance, and ELA grades in any year. See Appendix Table A1 for details.

Table 4. Summary Statistics for Linking and Validation Samples

	Linking Sample Students with any K-2 test data	Validation Sample Subset of linking sample with 3rd grade data, not including charter school students
No. of unique students	300,887	213,147
	Percent	Percent
% Asian	4.6%	4.6%
% Black	37.0%	35.1%
% Latinx	44.4%	46.5%
% White	12.0%	12.0%
% Other race/ethnicity	2.0%	1.8%
% MLL	30.0%	30.6%
% FRPL	80.8%	81.6%
% Charter or Special Ed	6.6%	0.0%

Note: The linking sample only includes students who were active in CPS long enough to appear in at least one enrollment snapshot, taken by the district once in the fall, winter, and spring, who also have at least one assessment data point in grades K-2. We do not have ELA grades for students at charter schools so they are removed when examining predictive validity.

Table 5. Correlations among K-2 Assessments and number of observations with each combination

	MAP	DIBELS	TRC English	ACCESS	TRC Spanish
MAP		**0.78**	**0.83**	**0.77**	**0.65**
DIBELS	118,295		**0.79**	**0.54**	**0.63**
TRC English	183,310	586,903		**0.75**	**0.77**
ACCESS	57,233	36,634	45,498		**0.65**
TRC Spanish	3,003	4,346	6,351	7,395	

Note: Correlations appear in the upper-right of the table. They are calculated based on students who took the two assessments in the same grade and quarter. The bottom left shows the number of students who took each pair of assessments on whom the correlations are based.

Table 6. Summary Statistics on Each K-2 Assessment in Each quarter

Grade &					Mean	S.D.	Mean	SD	
Quarter	Test	N	Min	Max	original	original	standardized	standardized	Skew
K Fall	Dibels	91,744	1	161	35.53	24.59	-1.007	0.221	0.591
K Fall	NWEA	34,582	100	215	141.35	12.59	-1.338	0.546	0.815
K Fall	TRC English	95,482	2	28	2.69	1.24	-1.170	0.193	4.296
K Fall	TRC Spanish	17,012	2	10	2.24	0.46	-0.901	0.090	2.389
K Winter	Dibels	87,717	1	376	120.16	57.08	-0.248	0.512	0.070
K Winter	NWEA	35,242	102	222	150.51	14.78	-0.941	0.641	0.602
K Winter	TRC English	103,546	2	29	4.14	1.93	-0.943	0.302	2.301
K Winter	TRC Spanish	14,482	2	15	3.12	1.18	-0.729	0.230	2.634
K Spring	Access	57,805	100	333	202.92	62.35	-1.204	0.989	0.051
K Spring	Dibels	71,426	1	333	131.17	54.68	-0.149	0.490	-0.003
K Spring	NWEA	30,850	104	225	160.18	15.87	-0.521	0.689	0.343
K Spring	TRC English	81,465	2	29	6.10	2.97	-0.638	0.463	1.473
K Spring	TRC Spanish	6,062	2	15	3.86	1.88	-0.585	0.368	1.762
1 Fall	Dibels	81,628	1	328	100.46	48.57	-0.424	0.436	0.267
1 Fall	NWEA	36,608	103	229	160.28	16.98	-0.517	0.737	0.318
1 Fall	TRC English	108,677	2	29	5.59	3.05	-0.718	0.475	1.531
1 Fall	TRC Spanish	9,346	2	16	3.84	2.08	-0.588	0.406	2.014
1 Winter	Dibels	75,461	1	528	124.70	93.80	-0.207	0.841	0.859
1 Winter	NWEA	41,592	108	242	169.39	17.59	-0.121	0.763	0.091
1 Winter	TRC English	111,213	2	29	8.17	3.94	-0.316	0.614	0.670
1 Winter	TRC Spanish	8,295	2	19	6.38	3.51	-0.092	0.685	0.620
1 Spring	Access	61,980	100	383	277.01	29.44	-0.029	0.467	-0.414
1 Spring	Dibels	60,558	1	453	138.35	96.63	-0.084	0.867	0.162
1 Spring	NWEA	41,304	100	239	177.94	17.49	0.249	0.759	-0.080
1 Spring	TRC English	86,731	2	29	11.15	4.80	0.148	0.748	0.109
1 Spring	TRC Spanish	4,217	2	28	7.70	4.22	0.166	0.823	0.273
2 Fall	Dibels	66,376	1	437	131.73	90.42	-0.144	0.811	0.116
2 Fall	NWEA	87,441	113	239	171.74	17.72	-0.020	0.769	0.262
2 Fall	TRC English	110,698	2	29	10.47	4.70	0.043	0.733	0.108
2 Fall	TRC Spanish	5,086	2	22	8.19	4.50	0.262	0.879	0.117
2 Winter	Dibels	62,469	1	1000	165.23	122.74	0.157	1.101	0.298
2 Winter	NWEA	150,545	115	247	178.28	17.77	0.264	0.771	-0.016
2 Winter	TRC English	111,824	2	29	12.63	5.08	0.379	0.792	-0.054
2 Winter	TRC Spanish	6,058	2	29	10.36	4.78	0.686	0.933	-0.325
2 Spring	Access	63,489	149	402	306.07	30.58	0.432	0.485	-0.759
2 Spring	Dibels	52,777	1	1000	204.50	133.06	0.509	1.193	0.114
2 Spring	NWEA	161,827	114	250	186.10	17.32	0.603	0.752	-0.267
2 Spring	TRC English	84,953	2	29	15.13	5.64	0.769	0.880	-0.139
2 Spring	TRC Spanish	3,221	2	29	11.10	5.24	0.830	1.023	-0.257

Note: Each assessment was standardized based on the range of values on that assessment across all grade levels, K-3, with data weighted to have the same representation at each grade level.

Table 7. Model to Produce Latent Scores: Coefficients and Variance Components

	Estimate	Std. Error		Estimate	Std. Error
K Fall	-1.470	0.002	DIBELS K Fall	0.419	0.003
K Winter	-1.065	0.002	DIBELS K Winter	0.839	0.003
K Spring	-0.615	0.002	DIBELS K Spring	0.531	0.003
1st Fall	-0.693	0.002	DIBELS 1st Fall	0.402	0.003
1st Winter	-0.286	0.002	DIBELS 1st Winter	0.256	0.002
1st Spring	0.117	0.002	DIBELS 1st Spring	0.000	0.003
2nd Fall	-0.115	0.002	DIBELS 2nd Fall	0.196	0.002
2nd Winter	0.223	0.002	DIBELS 2nd Winter	0.256	0.002
2nd Spring	0.560	0.002	DIBELS 2nd Spring	0.303	0.002
2020 Spring	-0.102	0.006	TRC-Eng K Fall	0.242	0.003
2021 Fall	-0.068	0.004	TRC-Eng K Winter	0.092	0.003
2021 Winter	-0.169	0.004	TRC-Eng K Spring	-0.014	0.003
2021 Spring	-0.301	0.004	TRC-Eng 1st Fall	-0.049	0.002
2022 Fall	-0.100	0.004	TRC-Eng 1st Winter	-0.024	0.002
2022 Winter	-0.161	0.004	TRC-Eng 1st Spring	0.074	0.002
2022 Spring	-0.215	0.004	TRC-Eng 2nd Fall	0.177	0.002
2020 1st grade	-0.009	0.008	TRC-Eng 2nd Winter	0.202	0.001
2020 2nd grade	0.014	0.008	TRC-Eng 2nd Spring	0.294	0.002
2021 1st grade	-0.053	0.005	TRC-Span K Fall	0.820	0.019
2021 2nd grade	-0.061	0.006	TRC-Span K Winter	0.781	0.008
2022 1st grade	-0.110	0.006	TRC-Span K Spring	0.671	0.008
2022 2nd grade	-0.125	0.006	TRC-Span 1st Fall	0.556	0.013
			TRC-Span 1st Winter	0.660	0.008
			TRC-Span 1st Spring	0.795	0.008
			TRC-Span 2nd Fall	0.882	0.014
			TRC-Span 2nd Winter	0.940	0.009
			TRC-Span 2nd Spring	1.097	0.009
			ACCESS K	-0.414	0.003
			ACCESS 1st	0.103	0.003
			ACCESS 2nd	0.166	0.002
Variance components	Variance	Std.Dev.	**Correlations of Variance Components**		
Kindergarten	0.183	0.428		Kindergarten	1st grade
1st Grade	0.363	0.603	1st grade	0.90	
2nd Grade	0.565	0.751	2nd grade	0.77	0.93
Within-student Residual	0.116	0.341			

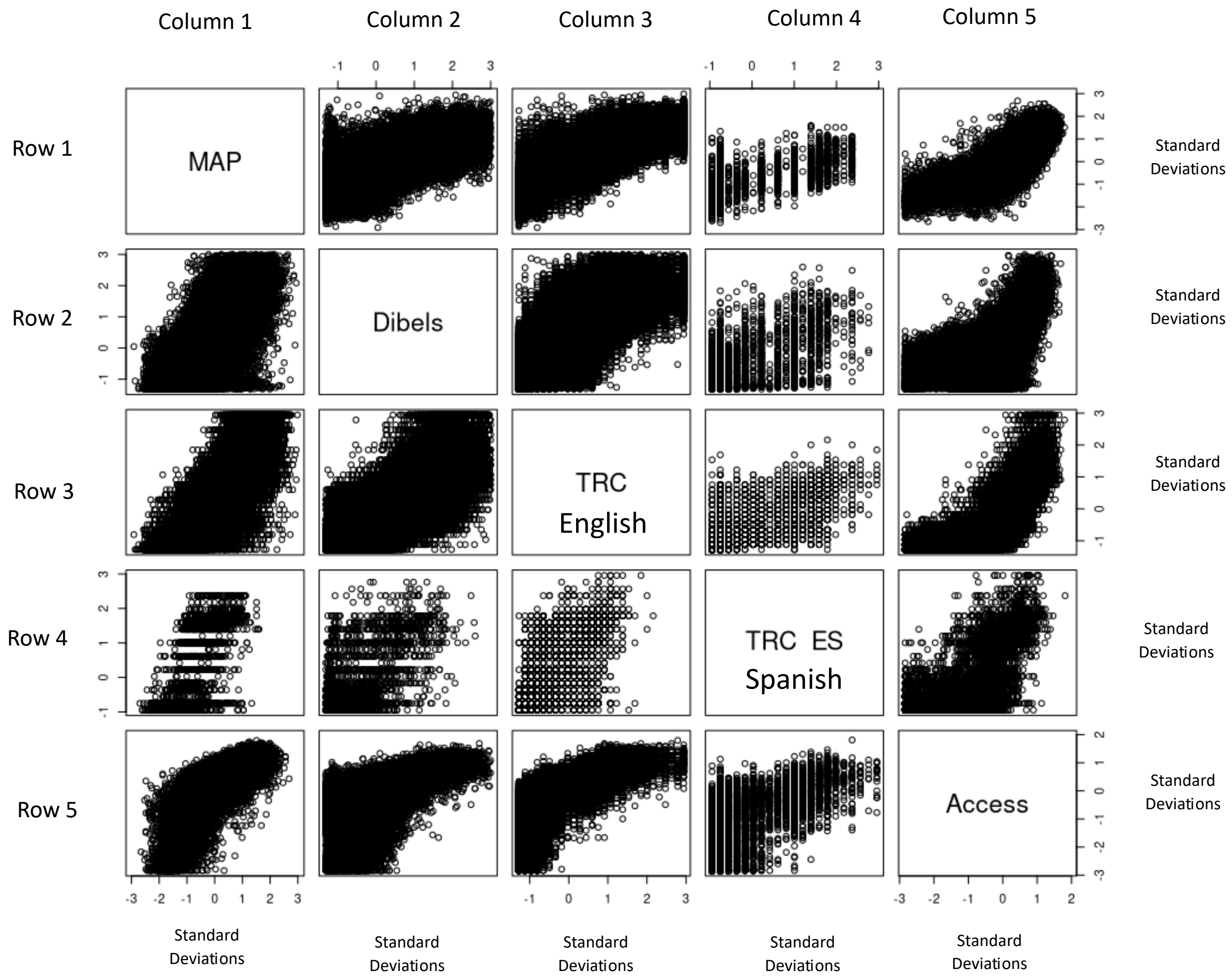
Notes: These coefficients come from a model that did not include TRC-Spanish test observations in pre-pandemic years. Model used 2,422,829 test observations from 300,887 unique students. The coefficients on the left show the average standardized scores in the spring of each grade level (Kindergarten, 1st Grade, and 2nd Grade), and the differences in average scores in pandemic years/quarters. The coefficients on the right show the difference in standardized scores on each assessment at each time point relative to MAP standardized scores at the same time point. All coefficients were included together in the model.

Table 8. Correlations of K-2 Latent Scores and Assessments with 3rd grade achievement

	Pre-pandemic Third Graders				**Post-pandemic Third Graders**		
	n cases	**3rd grade IAR/PARCC**	3rd grade MAP	**3rd grade ELA grades**	n cases	**3rd grade IAR/PARCC**	**3rd grade ELA grades**
Kindergarten latent scores	97,885	**0.67**	0.68	**0.52**	89,989	**0.65**	**0.46**
Kindergarten DIBELS	39,403	**0.52**	0.55	**0.41**	31,617	**0.47**	**0.35**
Kindergarten MAP	17,296	**0.66**	0.67	**0.55**	13,312	**0.66**	**0.47**
Kindergarten ACCESS	33,235	**0.53**	0.54	**0.37**	24,256	**0.50**	**0.33**
Kindergarten TRC English	38,743	**0.57**	0.58	**0.46**	42,262	**0.56**	**0.39**
Kindergarten TRC Spanish	1,791	**0.20**	0.26	**0.07**	4,249	**0.12**	**0.00**
Grade 1 latent scores	98,134	**0.76**	0.78	**0.61**	72,035	**0.71**	**0.51**
Grade 1 DIBELS	31,339	**0.65**	0.65	**0.52**	18,330	**0.57**	**0.36**
Grade 1 MAP	24,601	**0.76**	0.78	**0.63**	9,847	**0.76**	**0.51**
Grade 1 ACCESS	32,627	**0.60**	0.60	**0.42**	19,328	**0.60**	**0.38**
Grade 1 TRC-English	54,902	**0.68**	0.70	**0.55**	31,316	**0.65**	**0.47**
Grade 1 TRC-Spanish	541	**0.00**	0.20	**-0.07**	3,646	**0.16**	**0.56**
Grade 2 latent scores	110,591	**0.79**	0.82	**0.63**	53,150	**0.71**	**0.53**
Grade 2 DIBELS	25,657	**0.66**	0.71	**0.51**	9,877	**0.63**	**0.54**
Grade 2 G1 MAP	100,601	**0.77**	0.82	**0.62**	4,280	**0.80**	**0.65**
Grade 2 ACCESS	32,257	**0.67**	0.70	**0.48**	12,315	**0.62**	**0.40**
Grade 2 TRC-English	52,104	**0.69**	0.74	**0.56**	18,960	**0.68**	**0.61**
Grade 2 TRC-Spanish	219	**0.04**	0.07	**-0.02**	2,956	**0.45**	**0.55**

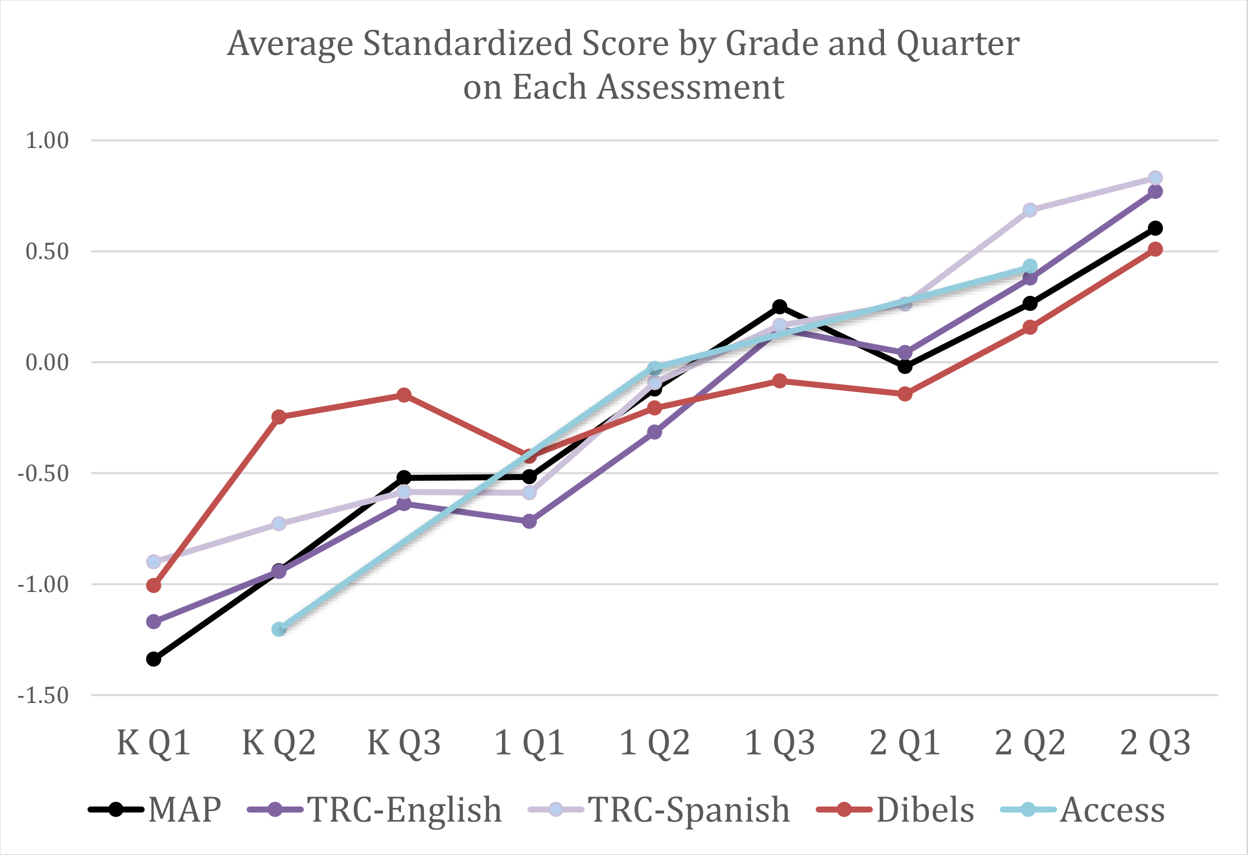
Notes: Correlations are based on students with assessment data on the indicated tests in the spring of each grade, except ACCESS which is taken in the winter. Literacy grades include teacher-assigned grades in reading and writing. Pre-pandemic third graders were in third grade by the 2018-19 school year. Post-pandemic third graders were in third grade in 2020-21 or 2021-22; these students would have had learning disrupted by the pandemic and half would have been in first grade and half in second grade in 2019-20, when there were no spring assessments.

Figure 1. Scatterplots showing relationships between each pair of assessments



Note: Dots display the standardized scores for each student who had scores on two different assessments in the same quarter and grade across all years of the study. A standardized score of -1 s.d. on ACCESS corresponds to about a 270 on the ACCESS scale.

Figure 2.



Note: The horizontal axis identifies each quarter (1=fall, 2=winter, 3=spring) in each grade (kindergarten, 1st, 2nd). The figure shows average scores on each assessment at each grade and quarter after standardizing across grade levels K-3 using just data from that assessment. This shows patterns in the data before linking the assessments through the analytic model. Changes from one quarter to the next are influenced by which students take each assessment at each time point, as well as the particular scoring methods of each assessment.

References

Amplify Education. (2019). *Amplify. Introducing Revised Authentic Spanish Texts*. https://amplify.com/pdf/uploads/2019/11/Titles-of-Revised-Spanish-TRC-Texts-v2-1.pdf.

Angoff, W. H. (1971). *The College Board admissions testing program: A technical report on research and development activities relating to the Scholastic Aptitude Test and Achievement Tests*. College Entrance Examination Board.

August, D., & Shanahan, T. (2017). *Developing literacy in second-language learners: Report of the National Literacy Panel on Language-Minority Children and Youth*. Routledge Press.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. doi:10.18637/jss.v067.i01.

Brookhart, S. M., Guskey, T. R., Bowers, A. J., McMillan, J. H., Smith, J. K., Smith, L. F., Stevens, M.T., & Welsh, M. E. (2016). A century of grading research: Meaning and value in the most common educational measure. *Review of Educational Research*, *86*(4), 803-848.

Chicago Public Schools. (2019). *TRC Factsheet*. https://www.cps.edu/globalassets/cps-pages/academics/student-assessments/trcfactsheet.pdf.

Clark, L. A., & Watson, D. (2016). Constructing validity: Basic issues in objective scale development. In A. E. Kazdin (Ed.), *Methodological issues and strategies in clinical research* (4th ed., pp. 187–203). American Psychological Association.

Dorans, N. J., Moses, T. P., & Eignor, D. R. (2010). Principles and practices of test score equating. *ETS Research Report Series*, *2010*(2), i-41.

Foorman, B., Beyler, N., Borradaile, K., Coyne, M., Denton, C. A., Dimino, J., Furgeson, J., Hayes, L., Henke, J., Justice, L., Keating, B., Lewis, W., Sattar, S., Streke, A., Wagner, R., & Wissel, S. (2016). Foundational skills to support reading for understanding in kindergarten through 3rd grade (NCEE 2016-4008). Washington, DC: National Center for Education Evaluation and Regional Assistance (NCEE), Institute of Education Sciences, U.S. Department of Education. Retrieved from the NCEE website: http://whatworks.ed.gov.

Fountas, I.C., & Pinnell, G.S. (2010). Benchmark Assessment System, BAS. Heineman Publishing.

Good III, R. H., & R.A. Kaminski. (2010). DIBELS® NEXT Assessment Manual. Dynamic Measurement Group, Inc.: https://dys-add.com/resources/DiblesNext/DIBELSNext_AssessmentManual.pdf

Goswami, U. (2001). Early phonological development and the acquisition of literacy. In S. B. Neuman & D. K. Dickinson (Eds.), *Handbook of early literacy research* (pp. 111-125). The Guilford Press.

Gwynne, J. A., Allensworth, E. M., & Liang, Y. A. (2022). Student engagement in learning during COVID-19. University of Chicago Consortium on School Research.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking*. Springer Publishing.

National Center on Intensive Intervention at the American Institutes for Research. (n.d.) https://charts.intensiveintervention.org/screening/tool/?id=29bb78bf24b4bb43.

National Reading Panel (US), National Institute of Child Health, & Human Development (US). (2000). *Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction: Reports of the subgroups*. National Institute of Child Health and Human Development, National Institutes of Health.

NWEA. (2019). MAP® Growth™ technical report.

NWEA. (2011). Technical manual for measures of academic progress (MAP) and measures of academic progress for primary grades (MPG).

Raudenbush, S.W., & Congdon, R.T. (2021). *HLM 8: Hierarchical linear and nonlinear modeling*. Scientific Software International, Inc.

Snow, C. E. (2006). What counts as literacy in early childhood? In K. McCartney & D. Phillips (Eds.), *Blackwell handbook of early childhood development* (pp. 274-294). Blackwell Publishing Ltd.. https://psycnet.apa.org/record/2006-04286-014.

Stuart, E. A., Lee, B. K., & Leacy, F. P. (2013). Prognostic score–based balance measures can be a useful diagnostic for propensity score methods in comparative effectiveness research. *Journal of Clinical Epidemiology*, *66*(8), S84-S90.

University of Oregon (2023). 8th Edition of Dynamic Indicators of Basic Early Literacy Skills (DIBELS®): Administration and Scoring Guide, 2023 Edition. Available: https://dibels.uoregon.edu.

von Davier, A. A., Holland, P. W., & Thayer, D. T. (2004). The chain and post-stratification methods for observed-score equating: Their relationship to population invariance. *Journal of Educational Measurement*, *41*(1), 15-32.

WIDA Center for Applied Linguistics. (2022). Annual Technical Report for ACCESS for ELLs Online English Language Proficiency Test Series 503, 2021–2022 Administration, Annual Technical Report No. 18A., University of Wisconsin.

Zhang, Z., Kim, H. J., Lonjon, G., & Zhu, Y. (2019). Balance diagnostics after propensity score matching. *Annals of Translational Medicine*, *7*(1), https://doi.org/10.21037/atm.2018.12.10.

Appendix

Information on Data Cleaning and Transformations

MAP and ACCESS. We did not identify any need for transforming these scores. The distributions were not highly skewed and did not contain large outlier scores. Average scores increased from quarter to quarter and grade to grade, as expected.

TRC English and Spanish. These scores were not originally numeric. Proficiency levels are based on a series of performance standards that result in 28 possible values. PC, RB, and A reflect the lowest levels of performance, with later letters in the alphabet representing progressively higher levels of proficiency. We converted this overall performance indicator into an integer scale by assigning the lowest level of performance, PC, to an integer equivalent of 1, RB to 2, A to 3, B to 4, and so on with Z corresponding to 28. We then examined a scatterplot of the relationship of scores on this scale to students' subsequent scores on the fall third grade MAP–a required assessment for all students which is on an interval scale with strong vertical equating–to determine whether the rescaled TRC scores corresponded in ways that suggested an interval scale, as well, or would need to undergo further transformation. We found the rescaled TRC had a strong linear relationship with students' MAP scores in third grade, with a correlation of 0.80 across the grade levels. That gave us confidence to use the rescaled TRC as if it were an interval scale.

TRC-Spanish. We noticed that a large percentage of students who took the TRC-Spanish assessment received the lowest score possible. We suspected these scores did not represent complete test records and found that students with the lowest scores on TRC-Spanish did not necessarily have low scores on other assessments. Rates of chronic absenteeism were also higher among students with zeroes on the TRC-Spanish assessment than the general population, which would be consistent with the theory that they represented incomplete records. Therefore, we removed observations with the bottom score from the dataset, and they are not included in any of the tables in the manuscript.

DIBELS. There are many different subtests associated with DIBELS. We considered and examined scores from the DIBELS subtests, and decided to use the composite score in the analyses because it was constructed for all tested time points and showed strong correlations with the subtests (r=0.51 to 0.92), and with the MAP reading score among students who took both assessments at the same time point (r=0.78). DIBELS was also unique in that some data points were extreme positive outliers, so values of greater than 4 standard deviations above the mean were trimmed before entering scores in the model.

Missing Data Analysis

Table A1 compares the characteristics of students with assessment data relative to the population of K-2 students, separately for pre-pandemic years, the 2020-21 school year, and the 2021-22 school year. Mean standardized mean differences (SMD) under 0.10 to 0.25 are generally considered acceptable when checking for balance in propensity score matching studies to consider two groups comparable (e.g., see Stuart, Lee and Leacy, 2013; Zhang et. al., 2018), and the SMD on all variables except is less than 0.10. The largest difference is a slight over-representation of multilingual learners in 2022 (31 percent of tested students compared to 28 percent of all students).

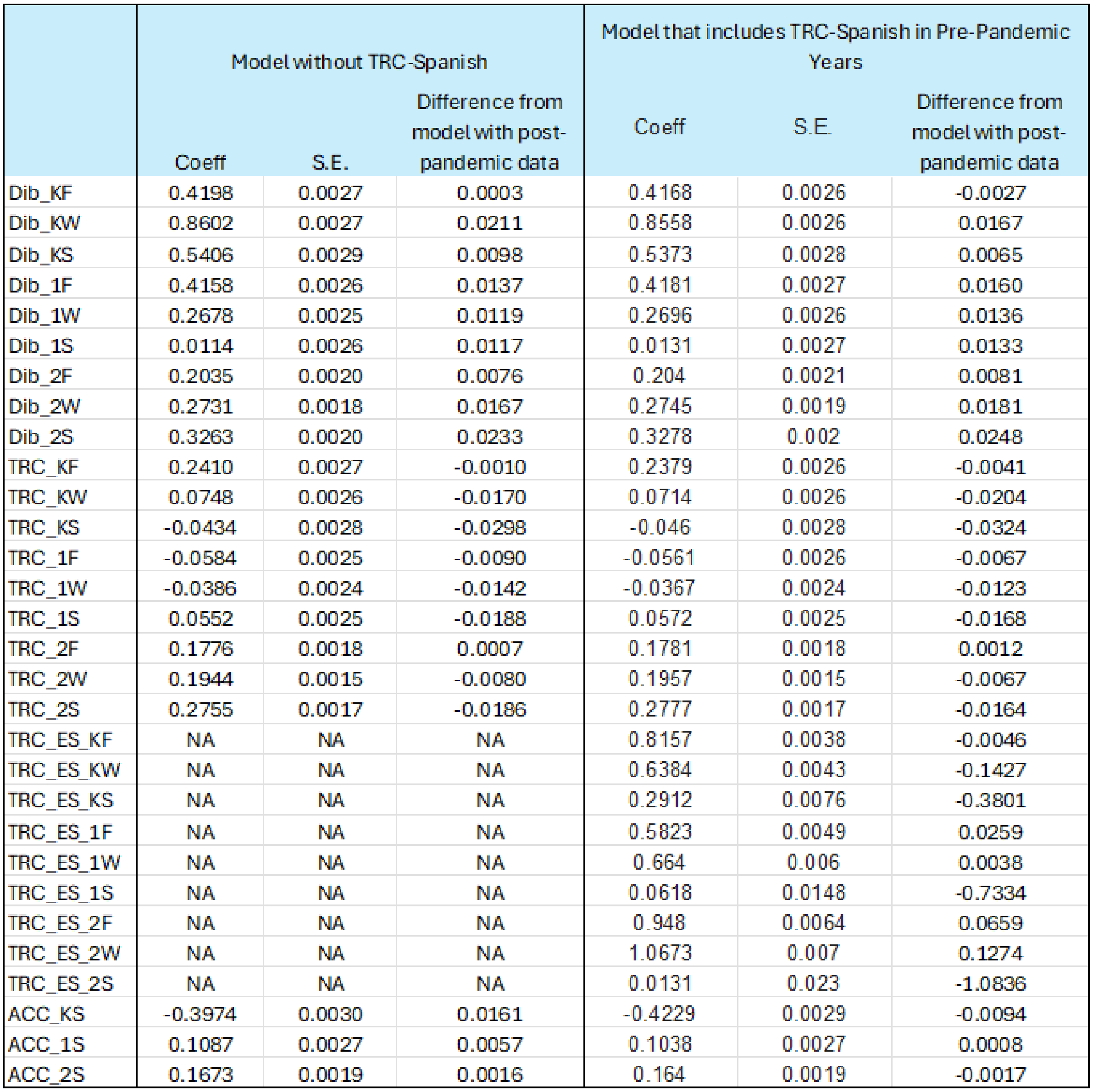
Table A1. Missing Data Analysis
Characteristics of All Active Students in Grades K-2 Versus Students with Test Data

Pre-pandemic Years	All Students		Tested		
Number of students	524,222		463,032		
	Mean	SD	Mean	SD	SMD
%Free/reduced lunch	67%	47%	67%	47%	0.009
%Black	36%	48%	35%	48%	0.011
%Latinx	45%	50%	46%	50%	0.015
% MLL	30%	46%	32%	47%	0.034
Attendance rate	94%	6%	95%	6%	0.020
% Chronically absent	15%	36%	14%	35%	0.011
GPA	2.79	0.96	2.78	0.96	0.001
2020-2021	All Students		Tested		
Number of students	71,894		35,328		
	Mean	SD	Mean	SD	SMD
%Free/reduced lunch	57%	50%	58%	49%	0.017
%Black	34%	47%	36%	48%	0.041
%Latinx	44%	50%	44%	50%	<0.001
% MLL	27%	44%	27%	44%	0.007
Attendance rate	91%	14%	91%	13%	0.015
% Chronically absent	25%	43%	26%	44%	0.011
GPA	2.97	0.98	2.95	0.96	0.025
2021-2022	All Students		Tested		
Number of students	70,137		53,377		
	Mean	SD	Mean	SD	SMD
%Free/reduced lunch	55%	50%	55%	50%	0.008
%Black	34%	47%	33%	47%	0.021
%Latinx	44%	50%	46%	50%	0.040
% MLL	28%	45%	31%	46%	0.083
Attendance rate	88%	10%	89%	10%	0.021
% Chronically absent	43%	49%	43%	49%	0.001
GPA	2.96	0.91	2.94	0.91	0.022

Note: We include an observation for each year for each student since students could have been in kindergarten in pre-pandemic years and first or second grade in 2021 or 2022. The standardized mean difference (SMD) is the difference in means between the two groups divided by the square root of the average within-group variance. It indicates group differences independent of indicator units or sample size, with differences of less than 0.1 generally considered comparable.

Coefficients from models that jut use data from pre-pandemic years

Table A2. Coefficients on Assessment Variables from Model with Just Pre-Pandemic Data



	Model without TRC-Spanish			Model that includes TRC-Spanish in Pre-Pandemic Years		
	Coeff	S.E.	Difference from model with post-pandemic data	Coeff	S.E.	Difference from model with post-pandemic data
Dib_KF	0.4198	0.0027	0.0003	0.4168	0.0026	-0.0027
Dib_KW	0.8602	0.0027	0.0211	0.8558	0.0026	0.0167
Dib_KS	0.5406	0.0029	0.0098	0.5373	0.0028	0.0065
Dib_1F	0.4158	0.0026	0.0137	0.4181	0.0027	0.0160
Dib_1W	0.2678	0.0025	0.0119	0.2696	0.0026	0.0136
Dib_1S	0.0114	0.0026	0.0117	0.0131	0.0027	0.0133
Dib_2F	0.2035	0.0020	0.0076	0.204	0.0021	0.0081
Dib_2W	0.2731	0.0018	0.0167	0.2745	0.0019	0.0181
Dib_2S	0.3263	0.0020	0.0233	0.3278	0.002	0.0248
TRC_KF	0.2410	0.0027	-0.0010	0.2379	0.0026	-0.0041
TRC_KW	0.0748	0.0026	-0.0170	0.0714	0.0026	-0.0204
TRC_KS	-0.0434	0.0028	-0.0298	-0.046	0.0028	-0.0324
TRC_1F	-0.0584	0.0025	-0.0090	-0.0561	0.0026	-0.0067
TRC_1W	-0.0386	0.0024	-0.0142	-0.0367	0.0024	-0.0123
TRC_1S	0.0552	0.0025	-0.0188	0.0572	0.0025	-0.0168
TRC_2F	0.1776	0.0018	0.0007	0.1781	0.0018	0.0012
TRC_2W	0.1944	0.0015	-0.0080	0.1957	0.0015	-0.0067
TRC_2S	0.2755	0.0017	-0.0186	0.2777	0.0017	-0.0164
TRC_ES_KF	NA	NA	NA	0.8157	0.0038	-0.0046
TRC_ES_KW	NA	NA	NA	0.6384	0.0043	-0.1427
TRC_ES_KS	NA	NA	NA	0.2912	0.0076	-0.3801
TRC_ES_1F	NA	NA	NA	0.5823	0.0049	0.0259
TRC_ES_1W	NA	NA	NA	0.664	0.006	0.0038
TRC_ES_1S	NA	NA	NA	0.0618	0.0148	-0.7334
TRC_ES_2F	NA	NA	NA	0.948	0.0064	0.0659
TRC_ES_2W	NA	NA	NA	1.0673	0.007	0.1274
TRC_ES_2S	NA	NA	NA	0.0131	0.023	-1.0836
ACC_KS	-0.3974	0.0030	0.0161	-0.4229	0.0029	-0.0094
ACC_1S	0.1087	0.0027	0.0057	0.1038	0.0027	0.0008
ACC_2S	0.1673	0.0019	0.0016	0.164	0.0019	-0.0017

Note: Coefficients from models with post-pandemic are shown in Table 7. Dib is DIBELS, TRC is TRC-English, TRC_ES is TRC-Spanish, ACC is ACCESS.