Placing Multiple Early Literacy Screening Instruments on a Common Risk Metric: A Machine Learning Approach

Sarah Quesen, Aaron Soo Ping Chow, Mariann Lemke, and Sean Tanner

WestEd

Correspondence: Sarah Quesen, WestEd. Email: squesen@wested.org

Keywords: early literacy screening, predictive modeling, random forest

Abstract

Many states now require early literacy screening in the primary grades but let districts choose from a list of approved instruments. Each instrument has its own scale, benchmarks, and definition of risk, so screening results are difficult to compare across instruments, districts, or time, and a state that collects these results centrally has no common metric of risk. This article presents a machine learning approach to address this. A random forest model is trained across all instruments at once and returns, for each student, the predicted probability of scoring in the lowest performance level (Not Meeting) on the grade 3 reading summative assessment, given whatever screening data the student has. This probability serves as a common metric: it is anchored to the same outcome and calibrated against the same observed rates regardless of which instrument or combination of instruments a student took. From these student-level predictions, we derive concordance tables that map each instrument's scores to predicted risk and can be checked against observed outcomes. Using data from one state, spanning 19 instruments and 84,245 students across five cohorts, the model is well calibrated overall, with an Integrated Calibration Index (ICI) of .016 or less at every grade and time point. A prospective evaluation on the most recent cohort preserved discrimination but underpredicted risk as the cohort's outcome rate rose, pointing to a need for periodic recalibration as the screened population changes. Preliminary analyses suggest that calibration may not be uniform across student groups; a differential prediction question is identified here and recommended for future study. The approach is built for aggregate research and reporting rather than individual placement, and its architecture is designed to transfer to states with similar screening policies.

Keywords: early literacy screening, predictive modeling, random forest

Placing Multiple Early Literacy Screening Instruments on a Common Risk Metric: A Machine Learning Approach

Universal early literacy screening has become common policy across the United States. Most states now require schools to assess foundational literacy skills several times a year in the early grades (Gearin et al., 2022; Neuman et al., 2023), and many carry out the requirement through local choice: each district selects a screening instrument from a list the state has approved. Local choice has practical appeal—it lets a district keep an instrument already in use, align screening with its curriculum and intervention systems, and weigh cost and testing time against its own constraints. It also creates a measurement problem that grows with the length of the approved list.

Approved instruments differ in their scales—in the criterion and method used to set benchmarks and in how they define risk (Lemke et al., 2023). One instrument may set a benchmark by predicting performance on a later assessment, while another uses a normative percentile or judgement-based standard setting (January & Klingbeil, 2020). Composite scores from different instruments use different metrics and carry different meanings. This creates challenges at multiple levels. At the individual student level, risk designations can vary. Students moving between districts or in districts using multiple assessments may be identified as at potential risk for reading difficulty on one assessment but not another (Klingbeil et al., 2017). Depending on the time of year and screening assessments, data from one state show rates of agreement ranging from 47 to 85 percent (Lemke et al., 2025). At the school or district levels, differences in risk benchmarks make it challenging to assess progress if assessments change. At the state level, districts that report screening results are not reporting comparable numbers, and a state that collects results from every district has no straightforward way to aggregate or analyze

them. This linking problem is important to how a state makes sense of the data to consider whether and how its screening mandate is producing intended results. We focus here on addressing how to link screening assessments efficiently and meaningfully for aggregate analysis and reporting purposes.

Several familiar approaches can and are already being used to address this challenge. Some states have simply adopted a single assessment. Others allow use of multiple assessments but use a rule such as flagging every student below the 25th national percentile to impose a common benchmark. This approach assumes the instruments are interchangeable at that percentile. A related option standardizes scores within each instrument, placing students on a z-score scale, as program evaluations may do when pooling outcomes across instruments (e.g., Nickow et al., 2024). But districts choose their instruments, so a z-score describes relative standing within nonequivalent groups, and it says nothing about the outcome a screener exists to predict. Another approach is equipercentile linking. Equipercentile linking can place screener scores onto the scale of a common state assessment by matching percentile ranks, which yields concordance tables and a defensible common scale (Kolen & Brennan, 2014). It describes where a current screening score falls within the state assessment distribution, one instrument at a time, and a state could select a risk threshold on the state assessment and apply it to screening scores. Because it links each instrument separately, it does not place all instruments, or the several scores a student accumulates across grades and instruments, on a single metric. Prediction has long been recognized as a distinct form of linking, suited to tests built to different specifications that cannot be equated but can each be related to a common criterion (Dorans, 2004; Linn, 1993). Screening itself rests on prediction: these instruments estimate the risk of later reading difficulty,

so expressing that risk against the state's own definition of a low reading outcome is a natural extension.

This article describes a machine learning approach that supplies such an estimate, using a state's own assessment and defined performance levels as the external criteria. A single model, hereafter the student-level model, is trained across all instruments at the same time, using students for whom both screening scores and a later state assessment outcome are known. The model is a random forest, chosen for well-calibrated probabilities, simple tuning, and native handling of mixed and partial screening histories; the model comparison behind that choice is reported in the Method section. For each student, the model returns the predicted probability of scoring in the lowest performance level (Not Meeting) on the grade 3 reading assessment, given whatever screening data the student has. That probability is the common metric, comparable across instruments and complete or partial screening histories. Additionally, the estimate is anchored to the state's already-defined performance level, which identifies students who need coordinated academic assistance or additional instruction—a level determined through the state's own standard-setting process.

We derive a second product from the student-level model's predictions, concordance tables, which translate each screening instrument's scores into predicted risk that can be validated against observed outcomes; their derivation is described in the Method section. We demonstrate the approach with data from one state, covering 19 screening instruments and a large set of matched records: 84,245 students across five cohorts from kindergarten through grade 3. We report how the student-level model performs, how the concordance tables map screener scores to predicted risk, and how the model's accuracy varies across student groups.

The predicted probability is computed for each student, but for reporting it is aggregated over many students, whose screening histories differ, to give an expected rate for a school, district, or state. Uses of that kind include tracking the level and trend of predicted risk statewide, comparing risk across districts that screen with different instruments, sizing the expected grade 3 need of a cohort screened this year, and asking whether the screening mandate is producing intended results. It is intended for research and aggregate reporting of this kind rather than individual student decisions. Districts already make individual decisions with their chosen instrument's own benchmarks and validity evidence; the common metric adds comparability across them. We return to this boundary in the Discussion section. The approach could be applied to any state. Its ingredients are a common state assessment with defined performance standards, longitudinal screening data from several instruments, and enough matched records to train and check a model. Because local instrument choice is a common screening policy nationally, the comparability problem, and this response to it, generalizes beyond the case examined here.

Method

Data

Screening data. Student-level screening records covered five school years (2021–2025). Each record identifies the instrument administered, the grade, the testing window (beginning, middle, or end of year), and the composite score for a single administration. The analysis included the 19 instruments that appeared in the data over the study period with at least 1,000 records each (see Table 1). The state's approved list has changed over time, so the set analyzed is broader than the list approved in any single year. The instruments varied widely in data volume. Two of them, mCLASS and DIBELS 8th Edition, are the same assessment delivered through

different platforms, and together they accounted for more than half of all records, followed by i-Ready, Star Early Literacy, and Star Reading. Several instruments contributed few records at any given time point, which limits the precision of any estimate specific to them. One instrument, MAP Reading Fluency, is delivered as two different assessments. Students taking the Foundational Skills assessment, generally in the earlier grades, receive component scores but no single composite, so a composite was derived from those components to place the instrument on a comparable footing for research purposes. Students taking the adaptive Oral Reading assessment receive an oral reading fluency score, in words correct per minute, which was used directly. The two appear as separate inputs, and the derivation of the Foundational Skills composite is described in the Appendix.

Table 1

Screening Instruments and Records, Kindergarten Through Grade 3

Screening instrument	Records, kindergarten through grade 3
mCLASS	388,479
DIBELS 8th Edition	220,344
i-Ready	147,990
Star Early Literacy	89,321
Star Reading	85,021
MAP Reading Fluency (Foundational Skills)	40,373
MAP Reading Fluency (Oral Reading)	24,230
Acadience Reading	14,439
FastBridge earlyReading	12,908
Lexia RAPID	11,797
FastBridge aReading	10,265
Star Early Literacy Spanish	6,446
EarlyBird	3,602
Star Reading Spanish	2,654

Screening instrument	Records, kindergarten through grade 3
FastBridge CBMreading	2,382
aimswebPlus	1,951
MAP Growth	1,908
ISIP ER	1,902
mCLASS Lectura	1,536

Note. Records are student by grade by testing window observations. Instruments with fewer than 1,000 records were excluded. mCLASS and DIBELS 8th Edition are the same assessment delivered through different platforms. MAP Reading Fluency contributes a derived Foundational Skills composite and a separate oral reading fluency measure; the derivation is described in the Appendix.

Outcome. The outcome was performance on the grade 3 statewide reading assessment, specifically whether a student's scale score fell below the boundary between the lowest and the second performance levels. On this state's English language arts assessment, this is the score that separates Not Meeting Expectations from Partially Meeting Expectations. About 19 percent of students in the analytic sample scored below this point. The outcome was modeled as binary, Not Meeting versus all higher levels. A model that predicts the scale score itself, which would let any performance threshold be applied after the fact, is a natural extension of this research and is discussed as a future direction below. Throughout, we use "risk" to mean the predicted probability of scoring Not Meeting on the grade 3 reading assessment, the state's lowest performance level.

Analytic sample. The analytic sample comprised 84,245 students from five cohorts, defined as students who had at least one screening composite score at any point from kindergarten through grade 3 and who had a grade 3 ELA state summative test outcome. Students were not required to have data at every grade or window, or from any particular instrument. A student screened once in grade 2 enters the sample alongside a student with a full record from kindergarten through grade 3. Table 2 shows the sample by cohort. The screening program expanded over the period, from a small set of districts that adopted screening early to a

much larger and more representative group by the final cohort. The rising share of students Not Meeting across cohorts primarily reflects this broadening of the sample; neither the state assessment nor the performance standard changed over the period. The period does, however, span cohorts whose early schooling was disrupted by the COVID-19 pandemic, so some portion of the rise may reflect real differences between cohorts rather than sample composition alone. Because the earliest cohort was small and unrepresentative, it is not examined on its own in the cohort-level analyses reported below, although it remains in the pooled data, including the training data for the prospective evaluation.

Table 2

Analytic Sample by Cohort

Grade 3 assessment year	Students	% Not Meeting
2021	833	7.1
2022	5,805	14.1
2023	13,784	16.6
2024	30,472	19.1
2025	33,351	22.3
Total	84,245	19.4

Note. Students are unique test takers with a grade 3 state English language arts outcome and at least one screening composite score from kindergarten through grade 3. The rising share Not Meeting across cohorts primarily reflects the broadening of the screening sample as districts joined; the assessment and the standard did not change over the period.

Screener and outcome relationships. As context for the model and the concordance tables, we examined the bivariate correlation between each instrument's composite score and the grade 3 reading scale score. Most correlations fell between .55 and .80. The major English language instruments were weakest in kindergarten, with correlations between roughly .35 to .58, and strongest in grade 3 (range of about .67 to .82), consistent with the expectation that screening administered closer in time and learner development to the outcome is more predictive. These

correlations reflect both the instruments and the populations that use them, not instrument quality. They also set a rough ceiling on what a model built on a single instrument can achieve, a ceiling that a model combining several instruments and time points may exceed.

Modeling Approach

The student-level model is a random forest classifier (Breiman, 2001), estimated with ranger, an implementation of random forests in R (Wright & Ziegler, 2017). It takes as inputs all of a student's available screening composite scores and returns the estimated probability of Not Meeting at grade 3. A defining choice was to train one model across all 19 instruments rather than a separate model for each. To make this choice, we compared the two approaches. For any single instrument, a logistic regression fit to that instrument alone discriminated about as well as the random forest, so the single model is not justified by sharper predictions for a typical student. The pooled model's advantages are structural. One model produces estimates on one probability scale for every instrument. A student screened in different years, for example, one instrument in grade 1 and another in grade 2, receives a prediction that uses both scores. An instrument with a modest number of records contributes to and benefits from a model fit to the full sample rather than requiring a separate model of its own, which is useful given that several instruments would not support a stable standalone fit. Importantly, when a state revises its approved list, a student's earlier scores on a discontinued instrument still inform the prediction, whereas a per-instrument linking approach would lose that history once the instrument left the list.

Because screening is typically carried out multiple times during the year (most frequently at the beginning, middle, and end of a school year), a separate model was trained at each of 12 time points, from the beginning of kindergarten through the end of grade 3. The models are cumulative (e.g., the model at the end of grade 2 uses every screening score from the beginning

of kindergarten through the end of grade 2). The number of input features therefore grows from 9 at the start of kindergarten to 143 at the end of grade 3. This is far below the 228 features that 19 instruments across 12 windows could in principle produce, because each instrument appears only in the grades and windows where it is actually administered, not in all of them. The number of features added at each time point reflects how many instruments contribute a score at that point. When a student has no score for a given instrument and window, the missing value is filled by median imputation, and the model's output is driven by the scores the student actually has. A sensitivity analysis at the end of grade 2 compared four approaches to missing data: median imputation, missingness indicators (binary flags appended for each missing value), ranger's native handling of missing values through surrogate splits, and complete-case analysis restricted to students with scores at all time points. The first three produced concordance tables whose predicted Not Meeting rates agreed within one percentage point across score bins. Complete-case analysis diverged by up to five to six percentage points at low scores, predicting smaller proportions Not Meeting compared to the other methods. Restricting to students with a score at every time point selects a non-representative subset, so this approach was not used. Because the three approaches that retain all students produced functionally equivalent results, median imputation was chosen for simplicity.

Concordance tables. The student-level model is accompanied by a second, derived product. For each instrument at each time point, students are grouped into bins on that instrument's score scale, and the mean model-predicted probability of Not Meeting within each bin is reported alongside the observed share of students in the bin who scored Not Meeting, with a bootstrap confidence interval on the observed rate. The result is a concordance table: a one-to-one mapping from a score range on one instrument at one time point to a predicted risk, with the

observed rate beside it as a check. The two products answer different questions. The student-level model conditions on a student's full history, so two students with the same score can carry different predicted risks; a concordance table conditions on one score at one time point and averages over the varied histories in each bin. Results below are labeled by which product they evaluate.

Model selection. Three algorithms were evaluated on the same data: penalized (ridge) logistic regression, random forest, and gradient boosted trees. Each was tuned by randomized search with fivefold cross validation. Random forest and gradient boosting performed similarly on discrimination, and the random forest was chosen for slightly better calibrated probabilities, simpler tuning, and feature importance measures that are easier to interpret, which mattered for the subgroup analyses described later. The random forest also avoids a logistic regression assumption that the relationship between a screener score and the log odds of the outcome is linear. A formal test (Box & Tidwell, 1962) rejected that assumption for 11 of the 14 instrument by time point combinations that had enough data to test. At sample sizes this large, the test detects even modest departures from linearity, so the results are treated as supporting evidence. The practical concern is that where the assumption fails, logistic regression can misstate risk at the ends of the score scale, understating it at one extreme and overstating it at the other depending on the direction of the departure, and score interpretations at the tail of a scale may carry the most consequence (Van Calster et al., 2019). The random forest makes no such assumption and lets the data set the shape of the relationship.

Evaluation. To evaluate the model, 15 percent of students were held out before training in a stratified random sample that preserved the proportion Not Meeting. All performance figures reported below come from this holdout and are restricted to students who had at least one real

score (not imputed) at the time point in question. Discrimination is reported with the area under the receiver operating characteristic curve (AUC), the probability that a randomly chosen student who later scored Not Meeting received a higher predicted risk than a randomly chosen student who did not where .50 is chance-level ranking and 1.0 is perfect separation. Calibration is reported with the Brier score (Brier, 1950), the ICI, and the calibration slope. The Brier score is the mean squared difference between predicted probabilities and observed outcomes. The ICI is the average absolute difference between predicted risk and loess-smoothed observed risk, taken across the distribution of predicted probabilities and estimated here with a local linear smoother with tricube weights and a span of two-thirds (Austin & Steyerberg, 2019). For the Brier score and the ICI, lower values indicate better calibrated probabilities; an ICI near zero means that a predicted 30 percent behaves like an observed 30 percent. The calibration slope's ideal value is 1.0. For reporting, students are also classified at a probability threshold, and sensitivity, specificity, and predictive values are summarized. In addition to the random holdout, a prospective evaluation trained the models on the four earlier cohorts only and evaluated them on the most recent cohort, restricted in the same way to students with a real score at each window; it covers the nine time points from grade 1 through grade 3. The distinction between discrimination and calibration is central to this application. We return to it in the Discussion section (Steyerberg et al., 2010).

Results

The model's ability to separate students who did and did not score Not Meeting improved as screening data accumulated. At the beginning of kindergarten, with little screening information and an outcome almost 4 years away, the AUC was .73 and sensitivity, the share of students who scored Not Meeting that the model correctly flagged, was minimal: the model

could rank students somewhat but could not reliably flag those at risk (i.e., likely to score Not Meeting). By the end of grade 2, the AUC reached .88 and sensitivity rose to .56, so the model correctly identified more than half of the students who ultimately scored Not Meeting. At the end of grade 3, the AUC was .91 and sensitivity reached .61. Table 3 reports discrimination and calibration for the student-level model at every time point, and Figure 1 shows the steady gains across all 12, with AUC rising and the Brier score falling as screening data accumulate.

Table 3

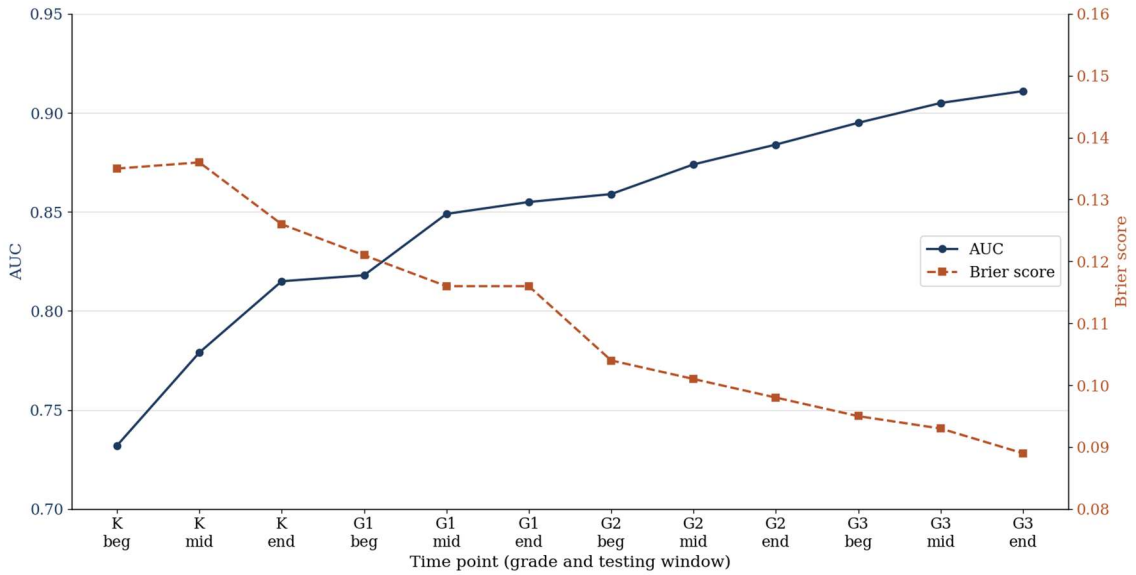
Discrimination and Calibration by Time Point

Time point	Features	*n*	AUC	Brier	ICI	Calibration slope
Kindergarten, beginning	9	454	.732	.135	.013	0.875
Kindergarten, middle	20	746	.779	.136	.015	0.973
Kindergarten, end	29	769	.815	.126	.016	1.080
Grade 1, beginning	39	2,320	.818	.121	.013	0.989
Grade 1, middle	50	2,391	.849	.116	.015	1.078
Grade 1, end	62	2,331	.855	.116	.016	1.069
Grade 2, beginning	75	5,245	.859	.104	.012	0.967
Grade 2, middle	88	5,376	.874	.101	.011	0.963
Grade 2, end	101	5,514	.884	.098	.011	0.970
Grade 3, beginning	115	11,688	.895	.095	.009	1.025
Grade 3, middle	129	11,961	.905	.093	.008	1.038
Grade 3, end	143	11,183	.911	.089	.010	1.051

Note. Features means the number of input variables, and *n* is the number of holdout students with a real score at the time point. AUC = area under the receiver operating characteristic curve; ICI = Integrated Calibration Index (Austin & Steyerberg, 2019). Beginning, middle, and end correspond to the fall, winter, and spring testing windows. All values are computed on the 15% holdout sample, restricted at each time point to students with a real score in that window.

Figure 1

Model Discrimination and Calibration Across Time Points



Note. AUC (left axis) and Brier score (right axis) computed at each of the 12 time points on holdout students with a real score in that window. Discrimination improves and the Brier score declines as screening data accumulate. Beginning, middle, and end correspond to the fall, winter, and spring testing windows.

The probability estimates were well-calibrated throughout. The Brier score fell from .135 at the beginning of kindergarten to .089 at the end of grade 3, and the ICI ranged from .008 to .016 across the 12 time points. In practical terms, when the model assigned a group of students a predicted risk near 30 percent, close to 30 percent of them scored Not Meeting.

Because the sample's composition shifted as the screening program expanded, calibration was also examined within the two most recent cohorts separately. Among holdout students in those cohorts, ICIs at the end of grade 2 and the beginning and end of grade 3 ranged from .009 to .020. Mean predicted risk fell within about two percentage points of the observed rate in each case, indicating that the pooled model's probabilities remain dependable for the population an operational model would currently serve.

The prospective evaluation is a more stringent test of deployment. Models trained only on the four earlier cohorts were applied to the most recent cohort and evaluated, as elsewhere, on students with a real score at each window (window samples of 10,317 to 32,153 students).

Discrimination was nearly as strong as in the random holdout, with AUC rising from .81 at the beginning of grade 1 to .90 at the end of grade 3, running .01 to .035 below the corresponding random-holdout values, with the largest gaps at the grade 2 windows. Calibration did not hold as well. The most recent cohort's Not Meeting rate exceeded the training cohorts' rate (22.3% against 17.6%), and the prospective model underpredicted risk across the scale. ICIs ranged from .023 to .047 compared to .008 to .016 in the random holdout, and the aggregate percent Not Meeting among screened students was understated by about four percentage points at the grade 1 windows, narrowing to between two and three points by grade 3. This drift tracks the base rate shift in size and direction, which suggests that periodic recalibration is needed to maintain model accuracy as a tested population changes.

Reporting also requires a decision rule. For that purpose, a student is classified as at risk when the predicted probability of Not Meeting is at least .50; that is, when the model estimates the student is more likely than not to score Not Meeting. Because most students have predicted probabilities below .50, the rule flags students with high specificity, above .92 at every time point and high negative predictive value: almost all the students that it does not flag go on to score in a higher level. The cost is sensitivity. At the beginning of kindergarten, sensitivity is around .01, and even at the end of grade 2, where it reaches .56, the rule misses close to half of the students who score Not Meeting. Positive predictive value climbs as screening data accumulate, from low values in the early grades to about .70 by grade 3. Table 4 reports sensitivity, specificity, positive predictive value, and negative predictive value for the student-level model at the .50 threshold at each time point. At the beginning of grade 3, for example, the rule flags 1,871 of 11,688 students, of whom 1,253 scored Not Meeting and 618 did not.

Table 4

Classification Accuracy at a .50 Threshold, by Time Point

Time point	*n*	Sensitivity	Specificity	PPV	NPV
Kindergarten, beginning	454	.012	.992	.250	.820
Kindergarten, middle	746	.180	.973	.628	.825
Kindergarten, end	769	.286	.959	.638	.843
Grade 1, beginning	2,320	.282	.948	.555	.852
Grade 1, middle	2,391	.409	.943	.638	.867
Grade 1, end	2,331	.482	.929	.636	.875
Grade 2, beginning	5,245	.441	.941	.616	.887
Grade 2, middle	5,376	.515	.929	.615	.897
Grade 2, end	5,514	.558	.934	.655	.904
Grade 3, beginning	11,688	.574	.935	.670	.905
Grade 3, middle	11,961	.597	.937	.696	.906
Grade 3, end	11,183	.608	.939	.699	.911

Note. Students with a predicted probability of Not Meeting at or above .50 are classified as at risk. PPV = positive predictive value; NPV = negative predictive value. *n* is the number of holdout students with a real score at the time point.

The central product of the analysis is the predicted probability itself. For every student at every time point, the model returns one number between 0 and 1, the estimated risk of scoring Not Meeting at grade 3, and that number is anchored to the same outcome for all 19 instruments. A predicted risk of .35 is the same statement about a student screened with mCLASS, with i-Ready, or with any other instrument, in the sense that each is calibrated against the same observed rates of Not Meeting at grade 3. This is the common metric that the screening mandate otherwise lacks: one number with one definition, anchored to the state's own assessment, in place of 19 instrument-specific benchmarks. The strength of the claim rests on calibration holding within each instrument's user population, not only in aggregate. Within-instrument calibration was checked for the four highest-volume instruments at the end of grade 2 and the beginning of grade 3. ICIs ranged from .005 to .023 and mean predicted risk fell within 2.3 percentage points of the observed rate in every case. The largest gap, for Star Reading at the end

of grade 2, was a modest overprediction of risk (15.9% predicted against 13.6% observed). The preliminary subgroup analyses reported below examine whether calibration is uniform across student groups.

Concordance tables, the derived product described in the Method section, put the metric in a form that analysts can easily use for research purposes. Table 5, Table 6, and Figure 2 evaluate this derived product rather than the student-level model. For each instrument at each available time point, students are grouped into score bins, and the table reports the average predicted risk in each bin, the observed share of students in that bin who scored Not Meeting, and a confidence interval for the observed rate. Table 5 shows the full table for mCLASS at the beginning of grade 3, and Figure 2 plots predicted against observed risk for the same bins. Predicted and observed values agree closely across most of the scale: a student scoring from 302 to 315 has a predicted risk near .32 against an observed rate of 30 percent, and predicted risk falls to about 1 percent at the top of the scale. The largest gap is at the lowest scores, where the model slightly underestimates risk, about 71 percent predicted against 77 percent observed. In two bins in the middle of the distribution, the divergence runs the other way, with the model assigning slightly more risk than students were observed to carry. The diagonal in Figure 2 makes both patterns visible, and the table translates screening scores into a risk metric, with the observed rate beside it as a check.

Table 5

Concordance Table: mCLASS, Beginning of Grade 3

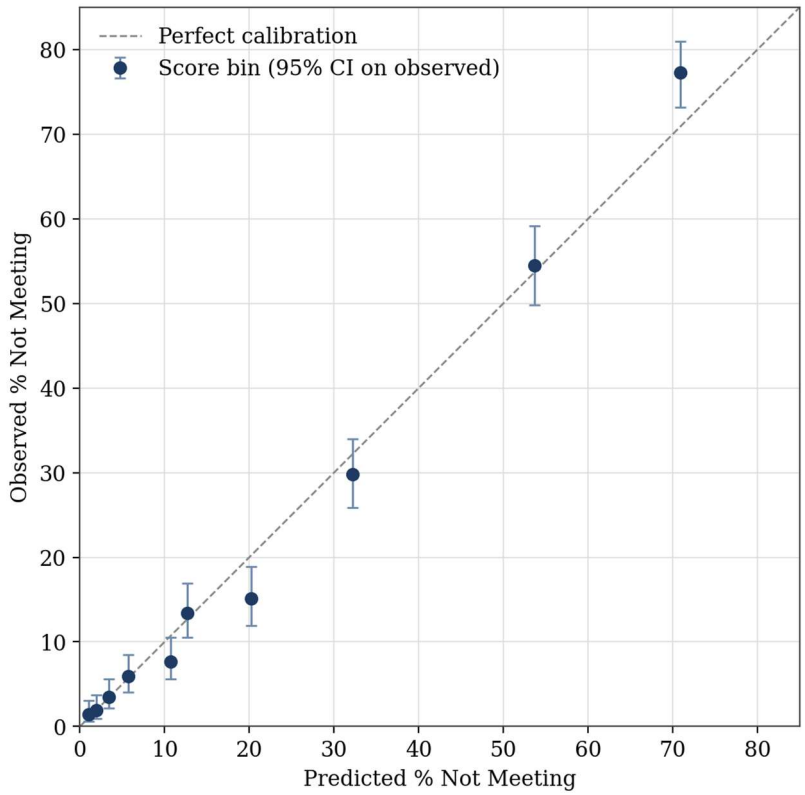
Score range	*n*	Predicted % Not Meeting	Observed % Not Meeting	95% CI (observed)
268 to 288	449	70.9	77.3	[73.2, 80.9]
289 to 302	433	53.7	54.5	[49.8, 59.1]

Score range	*n*	Predicted % Not Meeting	Observed % Not Meeting	95% CI (observed)
303 to 315	483	32.2	29.8	[25.9, 34.0]
316 to 325	405	20.2	15.1	[11.9, 18.9]
326 to 335	440	12.7	13.4	[10.5, 16.9]
336 to 346	454	10.7	7.7	[5.6, 10.5]
347 to 359	443	5.7	5.9	[4.1, 8.5]
360 to 368	453	3.4	3.5	[2.2, 5.6]
369 to 376	425	1.9	1.9	[1.0, 3.7]
377 to 467	420	1.0	1.4	[0.6, 3.0]

Note. mCLASS scores at the beginning of grade 3 (n = 4,405). Predicted % Not Meeting is the model's average predicted probability for students in each score range; observed % Not Meeting is the share who scored Not Meeting at grade 3. The 95% confidence interval is a Wilson interval for the observed proportion; the model's prediction is consistent with the observed rate in a bin when the predicted value falls within this interval. Score ranges are contiguous, and mCLASS composite scores are integers, so each score falls in exactly one range.

Figure 2

Calibration of Predicted Risk: mCLASS, Beginning of Grade 3



Note. Graph shows observed against predicted percentage Not Meeting for the score bins in Table 5 (mCLASS, beginning of grade 3). The dashed line marks perfect calibration, and the error bars are Wilson 95% intervals on the observed rate. Points near

the line indicate close agreement. The lowest score bin sits above the line, where the model slightly underestimates risk, and two bins in the middle of the scale sit slightly below it, where the model assigns somewhat more risk than was observed.

Because every instrument is expressed in the same risk metric, the tables also make instruments comparable at a chosen level of risk. At the end of grade 2, a predicted risk near .50, that is, roughly even odds of scoring Not Meeting, corresponds to about 404 on mCLASS, 403 on DIBELS 8th Edition, 442 on i-Ready, and 827 on Star Reading (see Table 6). These are not the publishers' benchmarks, which were set against different criteria, but rather the scores at which students in this state's data have historically had about an even chance of scoring Not Meeting at grade 3. States may find it informative to compare these outcome-anchored scores with the publishers' own benchmarks.

Table 6

Screening Scores Corresponding to a Common Level of Predicted Risk

Screening instrument	Score at predicted risk near .50
mCLASS	404
DIBELS 8th Edition	403
i-Ready	442
Star Reading	827

Note. Values are end of grade 2 scores at which the model's predicted probability of Not Meeting at grade 3 is about .50, that is, roughly even odds. They are derived from the model and the screening data and are not the publishers' benchmarks. These four are the highest-volume instruments. The score corresponding to any chosen level of risk can be read for any instrument from its own concordance table, where the instrument's observed score range spans that level of risk.

Calibration Across Student Groups. As a preliminary check, discrimination and calibration were also examined separately by student group, including gender, race and ethnicity, English Learner status, economic disadvantage, and special education status. Overall, discrimination was similar across groups, with the notable exception of students receiving special education services. Calibration did not appear uniform: for English Learner students, students from economically disadvantaged backgrounds, and Hispanic students, the model's

predicted probabilities diverged more from actual outcomes, and students in these groups who were flagged as not at risk failed to score above the Not Meeting threshold at grade 3 at higher rates than their peers.

If this pattern holds under closer study, it is primarily a property of the data rather than of the algorithm. Students in these groups who earn the same screening score as their peers would be, on average, more likely to fall short on the grade 3 assessment, and a model built only on screening scores cannot reflect that difference unless group membership enters as a predictor. The practical consequence would be that a single probability threshold applied across all students identifies a smaller share of the students who are actually at risk in these groups. A full treatment of these differential prediction questions, including the magnitudes involved, is deferred to future research.

Discussion

The contribution of this work is a way to place results from different screening instruments on a single probability scale. The predicted probability of scoring Not Meeting translates each instrument's scores into a shared risk metric, which lets a state compare risk across students, regardless of which instrument each took, and aggregate results across districts in a consistent, outcome-anchored way.

A common metric is not a claim that the instruments measure the same thing. The *Standards for Educational and Psychological Testing* caution that scores from different tests should not be treated as interchangeable without evidence of comparability (American Educational Research Association et al., 2014), and the variation in screener and outcome correlations confirms that these instruments differ in construct coverage, scale properties, and the populations they serve. What the model supplies is not interchangeability but a common external

anchor: each instrument's scores are interpreted against the same state summative assessment of grade 3 reading. That is a weaker but more defensible claim than direct comparability, and we judge it sufficient for research and aggregate reporting.

The introduction contrasted equipercentile linking with prediction as forms of linking; the results allow a sharper point. The proper analogue of an equipercentile concordance table is not the student-level model but the concordance tables derived from it, since both map a single score on one instrument to one value. The difference is the criterion. An equipercentile table matches percentile ranks in a concurrent sample, so its entries describe where a score stands in the outcome distribution (Dorans, 2004; Kolen & Brennan, 2014). The tables here report an outcome-anchored predicted risk instead, and each entry carries the observed Not Meeting rate beside it, so the mapping can be checked against the criterion it claims to estimate. Equipercentile linking suits a single instrument, a strong concurrent sample, and a question about current standing; the predictive tables suit a system of many instruments and a question about risk of a defined future outcome. The two approaches are complementary rather than competing.

For aggregate reporting, calibration matters more than discrimination. Discrimination concerns whether the model ranks students appropriately, and AUC values reaching .88 by the end of grade 2 and .91 by the end of grade 3 indicate that it does. Calibration concerns whether the predicted values can be interpreted as probabilities so that a predicted 30 percent corresponds to an observed 30 percent. A model can rank students well yet attach distorted probabilities, producing concordance tables that mislead despite a strong AUC (Steyerberg et al., 2010). Calibration was a design priority here, and the model meets it: with an ICI of .016 or less at every time point, the predicted probabilities can be read at face value, which is what the concordance tables require. The discrimination gradient also has a practical reading. At an AUC

of .73, at the start of kindergarten, the model ranks students only coarsely, so estimates there support only broad descriptions of need. At .91, at the end of grade 3, the ranking supports finer work, such as comparing risk across instruments or districts at specific score levels. What changes is not the meaning of a calibrated probability but how finely results can be cut.

Because the model places every student on one calibrated scale, it is possible to ask whether the scale is equally accurate for every group. A preliminary investigation found that English Learner students, students from low-income backgrounds, and Hispanic students who were flagged as not at risk scored Not Meeting at grade 3 at higher rates than similarly flagged peers. A single threshold may therefore flag a smaller share of the students who are actually at risk in these groups. These initial findings merit a dedicated study, and we return to them under Future Research below.

The predicted probability is best interpreted as a statement about a group, not an individual. A predicted risk of 40 percent means that among students with similar screening profiles, about 40 percent have historically scored Not Meeting. It does not assign a 40 percent chance to a particular student. Any use that treated it that way, such as placing a particular student into intervention based on this probability rather than on the screener's own benchmarks would require its own validity argument, with individual-level accuracy evidence, attention to the consequences of misclassification, and sensitivity beyond what Table 4 shows for the early grades (Kane, 2013). The model also cannot anticipate instruction. A student flagged as at risk in grade 1 who then receives effective intervention may score in a higher level by grade 3. The model's predictions reflect whatever mix of intervention or lack of intervention in the historical data. Predictions describe what has typically happened for students with a given profile, not what will happen for any student.

No element of the modeling architecture is specific to the instruments or assessment used here. The approach requires a state outcome assessment with defined performance levels, longitudinal screening data from several instruments linked to that outcome, and enough matched records to train and check a model. State agencies already run predictive models on longitudinal administrative data at scale; for example, in early warning systems for high school dropouts (Bowers et al., 2013; Knowles, 2015). The present approach extends that practice to early literacy. Because local instrument choice is common, many states could apply this same approach, although performance in other states would depend on its data volume, instrument mix, base rate, and the quality of the linkage between screening records and the outcome. The cumulative feature design, which uses a student's full history rather than a single score, is especially suited to programs that screen across several grades and testing windows.

Limitations. Several limitations qualify these results. The screening program in the state whose data is used here is still evolving, so the analytic sample does not represent every student in the state, and its composition has shifted over time. The model should be retrained as the sample stabilizes, and an approach in which a stable model is held for several years before it is estimated again may suit a mature screening program. Because the model pools five cohorts whose composition changed over that time period, with the share Not Meeting rising from 7 percent to 22 percent, its calibration could be affected if the relationship between screening scores and the outcome differs across cohorts rather than only the mix of students who were screened. The within-cohort checks reported in the Results section address this concern for the two most recent cohorts, and the prospective evaluation quantifies the calibration drift that a deployed model would face as the population shifts. The model is least reliable in kindergarten, where early scores carry limited information about an outcome that is years away and the

holdout samples are small. The calibration slope at the start of kindergarten is 0.875, below the ideal of 1.0, which indicates that the probability estimates there are somewhat compressed, so kindergarten results warrant caution. Instruments with few records yield concordance tables with wider intervals that should be used carefully. Finally, every result is anchored to the current state assessment scale and the current performance standards; rescaling the assessment or moving the performance boundary would require the model and all concordance tables to be regenerated.

Conclusion

States that require early literacy screening while allowing local instrument choice have a comparability problem at the aggregate level since risk may have different definitions and thresholds across different screeners. A single predictive model trained across all instruments offers a possible solution. It returns a common, externally anchored, calibrated risk metric that spans every instrument, accommodates mixed and partial screening histories, and produces concordance tables that can be checked against observed outcomes. The same common scale that makes aggregate reporting possible also makes questions about group-level accuracy visible and checkable. A state that mandates screening without a common metric can report how many students were screened, but not, in any comparable way, what the screening found. The approach here turns a patchwork of incompatible benchmarks into one number that a state can track over time, compare across districts and instruments, and check against the outcome it claims to estimate.

Future Research. To support state implementation, a validity argument for the intended use of the common metric should be carefully constructed—that states explicitly how the scores will be used and by whom, including the level of aggregation at which results are reported (Kane, 2013). The aggregate uses emphasized here, such as an expected percentage of students

predicted to score Not Meeting at the district or state level, are the natural starting point for that argument. The design choices examined in this article, from the outcome definition to the reporting threshold, are interdependent.

A full study of differential prediction across student groups and across instruments is still needed. The preliminary subgroup results reported here suggest that the model's risk classifications are less reliable for English Learner students, students from low-income backgrounds, and Hispanic students, identifying a smaller share of students who subsequently perform in the Not Meeting category than for similarly performing peers not in those groups. A dedicated study should estimate the size of those calibration gaps by group, instrument, and time point and weigh the available responses: holding one threshold and accepting uneven identification, setting thresholds separately by group, or adding demographic predictors.

The prospective evaluation showed well-preserved discrimination alongside a calibration drift tied to the rising base rate, which raises the practical question of maintenance: whether periodic recalibration of the probabilities, with full retraining reserved for a less frequent, norming-style refresh, keeps the concordance tables accurate between updates. The minimum record count for adding a new instrument should also be established empirically, by simulating prediction error at different sample sizes with attention to how much of that error reflects differences between districts, before the 1,000-record floor used here is treated as a rule.

Finally, the next planned study is an extension of the outcome. A model fit to the grade 3 scale score itself, such as a quantile random forest, would let any performance boundary be applied after the fact and would make the approach robust to changes in the cut. Its outputs would still be reported only in aggregate; publishing predicted scale scores for individual students would invite exactly the uses that the approach is not built to support. Comparing the

score thresholds implied by such a model with those produced by equipercentile linking for the major instruments is a natural companion analysis.

References

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

Austin, P. C., & Steyerberg, E. W. (2019). The Integrated Calibration Index (ICI) and related metrics for quantifying the calibration of logistic regression models. *Statistics in Medicine, 38*(21), 4051–4065. https://doi.org/10.1002/sim.8281

Bowers, A. J., Sprott, R., & Taff, S. A. (2013). Do we know who will drop out? A review of the predictors of dropping out of high school: Precision, sensitivity, and specificity. *The High School Journal, 96*(2), 77–100. https://doi.org/10.1353/hsj.2013.0000

Box, G. E. P., & Tidwell, P. W. (1962). Transformation of the independent variables. *Technometrics, 4*(4), 531–550.

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review, 78*(1), 1–3.

Dorans, N. J. (2004). Equating, concordance, and expectation. *Applied Psychological Measurement, 28*(4), 227–246. https://doi.org/10.1177/0146621604265031

Gearin, B., Petscher, Y., Stanley, C., Nelson, N. J., & Fien, H. (2022). Document analysis of state dyslexia legislation suggests likely heterogeneous effects on student and school outcomes. *Learning Disability Quarterly, 45*(4), 267–279. https://doi.org/10.1177/0731948721991549

January, S.-A. A., & Klingbeil, D. A. (2020). Universal screening in grades K–2: A systematic review and meta-analysis of early reading curriculum-based measures. *Journal of School Psychology, 82*, 103–122. https://doi.org/10.1016/j.jsp.2020.08.007

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*(1), 1–73.

Klingbeil, D. A., Nelson, P. M., Van Norman, E. R., & Birr, C. (2017). Diagnostic accuracy of multivariate universal screening procedures for reading in upper elementary grades. *Remedial and Special Education, 38*(5), 308–320. https://doi.org/10.1177/0741932517697446

Knowles, J. E. (2015). Of needles and haystacks: Building an accurate statewide dropout early warning system in Wisconsin. *Journal of Educational Data Mining, 7*(3), 18–67.

Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking: Methods and practices* (3rd ed.). Springer.

Lemke, M., Faulkner-Bond, M., LeBeau, B., Soo Ping Chow, A., & Quesen, S. (2025). *Early literacy performance in Massachusetts: Results of ongoing analysis of literacy screening assessments (2023/24 data).* WestEd.

Lemke, M., Murphy, D., Soo Ping Chow, A., Spencer, H., & Zhang, A. (2023). *A first look at early literacy performance in Massachusetts: Results of initial analysis based on state grantee literacy screening assessments.* WestEd.

Linn, R. L. (1993). Linking results of distinct assessments. *Applied Measurement in Education, 6*(1), 83–102. https://doi.org/10.1207/s15324818ame0601_5

Neuman, S. B., Quintero, E., & Reist, K. (2023). *Reading reform across America: A survey of state legislation*. Albert Shanker Institute.

Nickow, A., Oreopoulos, P., & Quan, V. (2024). The promise of tutoring for prekindergarten through 12th-grade learning: A systematic review and meta-analysis of the experimental evidence. *American Educational Research Journal, 61*(1), 74–107. https://doi.org/10.3102/00028312231208687

Steyerberg, E. W., Vickers, A. J., Cook, N. R., Gerds, T., Gonen, M., Obuchowski, N., Pencina, M. J., & Kattan, M. W. (2010). Assessing the performance of prediction models: A framework for traditional and novel measures. *Epidemiology, 21*(1), 128–138.

Van Calster, B., McLernon, D. J., van Smeden, M., Wynants, L., & Steyerberg, E. W. (2019). Calibration: The Achilles heel of predictive analytics. *BMC Medicine, 17*, Article 230. https://doi.org/10.1186/s12916-019-1466-7

Wright, M. N., & Ziegler, A. (2017). ranger: A fast implementation of random forests for high dimensional data in C++ and R. *Journal of Statistical Software, 77*(1), 1–17. https://doi.org/10.18637/jss.v077.i01

Appendix: Derived Composite for MAP Reading Fluency

MAP Reading Fluency does not report a single composite score of the kind the other instruments provide that is used to flag students at risk of reading difficulty. For students taking the foundational assessment, a composite was constructed to place the instrument on a comparable footing with the rest. The composite was derived from the instrument's component scores from the phonological awareness, phonics & word recognition, language comprehension, and sentence reading fluency measures. Students for whom the component scores needed to build the composite were unavailable were treated as missing on this instrument and handled by the same median imputation used for the other instruments. This composite was calculated using the following steps:

1. Use subtest scores to predict significant risk in a logistic regression equation.
2. Create a weighted composite score by using the coefficients from the logistic regression results as weights for their corresponding subtest scores and summing the weighted scores.
3. Rescale the MAP Reading Fluency weighted composite score to a scale with a mean of 1,000 and a standard deviation of 200.

For students taking the adaptive Oral Reading assessment, the oral reading fluency measure, reported in words correct per minute, was used as their composite score as this measure is used to flag students who take this assessment. Both the derived composite and the separate oral reading fluency measure appear as individual rows in Table 1.