Predicting Math Proficiency Using K-3 Diagnostic Assessments

Zeyu Xu
American Institutes for Research
1400 Crystal Drive, 10th floor
Arlington, VA 22202-3289
zxu@air.org
202-834-7329
ORCID: 0000-0003-4894-0141

April 14, 2024

Acknowledgments: This work is supported by the National Science Foundation under Grant No. 2000483. The author has no conflict of interest to declare. The author would like to thank research partners Karen Dodd, Erin Chavez, Aaron Butler, and Hannah Poquette from the Kentucky Department of Education; Barrett Ross from the Kentucky Center for Statistics; Kelly DeLong from the Kentucky Center for Mathematics; and Mary Lee Glore from Northern Kentucky University. The author would also like to thank Kyle Pinder for research assistance. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the National Science Foundation, the Kentucky Department of Education, the Kentucky Center for Statistics, the Kentucky Center for Mathematics, or Northern Kentucky University.

Predicting Math Proficiency Using Grade 1–3 Universal Screeners

Abstract: Children's math performance is strongly correlated with later life outcomes, but early gaps in math skills are stubbornly difficult to close. It is therefore important to identify student math needs early. Using Grade 1–3 student records from Kentucky public schools, the study finds that typically recommended cut scores for widely used early grade math screeners severely under-identify student needs in math. Using optimal cut scores estimated by the Classification and Regression Tree (CART) analysis, the likelihood of under identification of at-risk students decreases by an average of 16 percentage points and sensitivity improves by 28 percentage points. These improvements can be achieved without having to collect new data or administer new assessments.

Keywords: Mathematics performance, Early elementary grades, Prediction

Introduction

Large deficits in math skills emerge early, and children from disadvantaged families often begin school with much less math knowledge than their peers (Jordan et al., 2006). Early gaps in math skills are stubbornly difficult to close; without timely interventions, they tend to widen over the course of schooling (Geary et al., 2013; Loeb & Bassok, 2007; Morgan et al., 2011). How to diagnose deficiencies in math skills early, therefore, is critically important.

However, the use of universal screeners to identify early-grade math needs are limited because state standardized testing is not typically required before the 3rd grade. Schools and districts can use assessments like MAP and iReady to measure learning progress and screen for students who may not be on track to meet grade-level benchmarks.[1] Yet the degree to which early grade assessment results can accurately reflect and predict student skills development, as

measured by state standardized tests, is an open question. A study on interim reading screeners administered to K-2 students in North Carolina, for example, found that these screeners did not adequately identify students who were at risk of scoring below proficient on the state reading assessment at the end of Grade 3 (Koon et al., 2020). There are no similar studies on early-grade math screeners,[2] but some teachers express concerns that early grade math screeners may identify deficiency in language skills instead of math skills (Xu et al., 2023). There is a need for more research so that schools and districts can make informed use of early-grade math assessment results.

Using student-level longitudinal administrative data from five cohorts of K-3 students in Kentucky, this study examines the concordance between student performance levels on math screeners and the likelihood that they score at and above grade level on 3rd grade state standardized tests in math. The study uses a machine learning technique to estimate the optimal cutoffs for these screeners that improve the predictive accuracy for 3rd grade math proficiency.

The study is not an evaluation of the reliability and validity of the screeners (Petscher et al., 2019). Rather, it is intended to gauge how well early grade screeners and publisher recommended cut scores align with criteria that states may deem important. As such, the optimal cutoffs estimated in this study should not be simplistically applied to other states or even to schools and time periods in Kentucky that are not included in the current analysis. The study calls attention to the need for each school system to evaluate screener assessments from time to time and to adjust cutoffs so that, in combination with other student information, can better identify student needs in math early. More broadly, this study is related to the literature on identifying students in need of additional resources and support (e.g., Fazlul et al., 2022) as well

as research on early warning systems established for dropout prevention (U.S. Department of Education, 2016).

In what follows, the study first describes the data and sample used in this study. This is followed by an introduction to measures of classification accuracy. The performance of three most frequently used early grade screeners is evaluated using these measures. In the next section, the study explores various ways in which classification accuracy may be improved using the Classification and Regression Tree (CART) analysis. The study concludes with a discussion of findings and some recommendations for more informative use of early grade math screeners.

Data and Samples

This study uses universal math screener scores collected by the Kentucky Center for Mathematics from 146 elementary schools in Kentucky among K-3 students between 2015–2016 and 2019–2020. Table 1 compares the characteristics of these elementary schools and other elementary schools in Kentucky during the same time period. Summary statistics show that schools that administered universal math screeners tend to have a lower percentage of Black students (by 5.6 percentage points) and a higher percentage of White students (by 6.3 percentage points) than other elementary schools. Compared to elementary schools not in this analysis, elementary schools that administered universal math screeners are more likely to be located in towns and less likely to be in urban areas.

Math screener scores are merged with administrative records from the Kentucky Department of Education that contain state standardized test scores (formerly known as K-Prep) between 2015–2016 and 2021–2022 as well as school enrollment data and student characteristics such as race/ethnicity, free/reduced price lunch status, and gender. For students who took the

screener multiple times per year, only the first score (usually in the fall) is used in the main analysis. The number of unique students with both a valid screener score and K-Prep scores is reported in Table 2 by screener type and the grade. The type of screener administered appears to be a school-level choice, as it does not vary within school. The most frequently used screeners are MAP, iReady, and Star. The study focuses on these three screeners due to sample size considerations. In addition, because the number of students who took a math screener in kindergarten is small (except for MAP), the study focuses on screeners administered to students in Grades 1 through 3.

Classification Accuracy and Benchmarking

Test publishers suggest performance thresholds that can be used to provide guidance on whether a student performs at a certain level. For example, MAP scores are normed against a nationally representative sample of students every year and are grouped into 5 performance levels: low (<21 percentile), low average (21-40 percentile), average (41-60 percentile), high average (61-80 percentile), and high (>80 percentile).[3] Students performing above the 40th percentile on MAP are often considered to be at or above grade level. iReady publishes scale score placement tables that define the expected score range for each subject and grade (and within each grade, for each fall, winter, and spring administration).[4] First grade students who score at or above 402 on the spring administration of math, for example, would be considered as performing at or above grade level. iReady's cut score roughly corresponds to the 40th percentile of each grade, year, and administration. Finally, Star also considers students who score at or above the 40th percentile to be at or above grade level.[5]

States also publish descriptors of performance levels and cut scores for each level. In Kentucky, student performance is categorized as novice, apprentice, proficient, and distinguished based on state standardized tests (Kentucky Department of Education, n.d.). Students achieving proficient and distinguished are considered as performing at or above grade level. Hence, one way to examine the classification accuracy of screeners is to examine the concordance of performance classification between the screener and the state standardized test. Specifically, how well a screener and its cut score can predict the likelihood that a student will perform at or above grade level on 3rd grade K-Prep math can be characterized by a 2X2 contingency table (Table 3).

True positive (TP) represents students who score below grade level based on both the screener and K-Prep. True negative (TN) represents students who score at or above grade level on both the screener and K-Prep. False positive (FP) represent students who are predicted by the screener to be at risk of falling below grade level but who later meet grade-level expectations on K-Prep. Finally, false negative (FN) represent students who are predicted to be not at risk but who later fail to meet grade-level expectations on K-Prep.

Predictions are not expected to be 100% accurate. What would be an acceptable level of classification error? Benchmarks for classification accuracy indices vary. For sensitivity and specificity (see definitions in Table 3), researchers have proposed at least .80 or .90 (Compton, Fuchs, Fuchs, & Bryant, 2006; Jenkins, 2003). Jenkins (2003) proposes a target negative predictive power (definition in Table 3) of .90-.95. In practice, it is important to keep in mind that, although FP and FN both represent prediction errors, the severity of their consequences could vary. In an educational setting, FN could mean that students who need help fail to receive supplemental support in time. This is often more consequential than giving students extra help when they do not need it (FP). For this reason, education policymakers often focus on

maximizing the negative predictive power (i.e., minimizing false negatives—the percentage of students predicted as not at risk who later perform below grade level). When assessing the predictive power of an interim reading assessment administered in grades 1 and 2, for example, the Florida Department of Education focused on the negative predictive power and set a minimum criterion of .85 (Koon et al., 2014).[6]

How well can performance levels based on math screeners predict a student's performance level on 3rd grade K-Prep math? Contingency tables similar to Table 3 can be found in the Appendix (Table A1a-A1c) for detailed breakout of the estimated TP, FP, FN, and TN by screener type and grade. These estimates are used to calculate three measures of classification accuracy: positive predictive power, negative predictive power, and sensitivity. Positive (white bars) and negative (black bars) predictive powers are presented in Figure 1 for the three most widely used math screeners in our data (MAP, iReady and Star) by the grade in which the screener was administered. These measures reflect the degree to which we should trust predictions made using publisher-suggested cutoffs (i.e., the 40th percentile). The figure shows, for example, if a student tested below grade level on MAP in the 2nd grade, there is an 80% chance that the student would not reach grade level expectations for math in the 3rd grade. In general, these math screeners have high positive predictive power, suggesting that schools and teachers should take an at-risk prediction seriously.

By comparison, negative predictive power (black bars in Figure 1) is low across screeners and grades, suggesting that a no-risk prediction should be trusted less. For example, if a student tested at or above grade level based on 2nd grade MAP, there is a 34% chance (1-66%) that the student would not meet grade level expectations for math in the 3rd grade. The dashed line in Figure 1 represents the target negative predictive power (85%) that the Florida Department of

Education set for its early grade reading assessments (Koon et al., 2014). It is clear that the performance of the math screeners and their associated cutoffs is far below that benchmark. As discussed earlier, low negative predictive power could mean many missed opportunities to help students who need support, and so it is a serious concern.

Another measure of classification accuracy is sensitivity. It is the percentage of actual at-risk students (defined as performing below grade level on 3rd grade K-Prep math) who have been correctly predicted by screeners. Figure 2 shows that screeners—including those administered in the fall of 3rd grade)—can identify no more than 65% of students who later failed to meet 3rd grade expectations in math using published cut scores. In other words, screener results and published cutoffs are not sensitive enough to identify math needs among at least 1 in 3 students who later struggle with math.

Classification and Regression Tree (CART) Analysis

Findings from the analysis so far suggest that publisher suggested screener cutoffs might be too low for this particular sample of students and schools in Kentucky. To improve classification accuracy, the study uses the Classification and Regression Tree (CART) method (Breiman et al., 1984) to estimate the optimal cutoffs that minimize classification errors. CART analysis is also used to explore the extent to which additional information about a student could further improve classification accuracy.

CART is a machine learning technique that has been used in education research occasionally (See, for example, Therneau, & Atkinson, 2013; Koon et al., 2014). It is a nonparametric technique used to predict a class Y from inputs X. In our case, the goal is to predict if a student would perform below grade level on 3rd grade K-Prep math (Y) using

screener scores (X). CART involves recursive binary data partitioning. It starts with a test or question about X (e.g., is a MAP score>40th percentile?) that maximizes the information we can get (or reduction in uncertainty) about Y. Students in the original sample (the root node) will be partitioned (branched) into two groups (the internal nodes) depending on whether they pass or fail the test. This process can be repeated by applying a sequence of tests about X, which could be either new cutoff values of the same input variable or new input variables. Data partitioning will eventually reach a terminal node where a class prediction about Y is made.

This study uses the R package, *rpart*, to carry out the CART analysis. By default, *rpart* implements 10-fold cross validation that splits the original analytic sample into training and testing datasets. Information (or uncertainty) is measured using the misclassification rate, and it is possible to assign different costs to FNs and FPs. As discussed earlier, because FNs are of higher policy importance in the current setting, more weights can be assigned to FNs to improve the negative predictive power. *rpart* reports the proportion of classification error in the root node that remains after a split (the relative error), which is analogous to $1\text{-}R^2$ in a regression analysis. The tradeoff between the number of splits (or a complexity parameter) and the reduction in relative error is evaluated to determine a stop criterion. The conventional recommendation is to choose the smallest number of splits that results in a cross-validation relative error less than the minimum cross-validation relative error plus 1 standard error (Therneau et al., 2013).

Improvements in Classification Accuracy

Changing cutoffs

In Figure 3, publisher-suggested cutoffs (black dots) are compared with unweighted (open circles) and weighted (blue dots) CART-optimized cutoffs. The unweighted cutoff assigns

equal weights to FNs and FPs, and the weighted cutoff assigns twice as much weight to FNs as to FPs (blue dots) to further improve the negative predictive power. As predicted earlier, CART analysis suggests that the optimal cutoffs are much higher than the 40th percentile typically recommended by publishers. The unweighted cutoff for 3rd grade MAP, for example, is the 58th percentile. Interestingly, a linking study conducted by the MAP publisher for Grades 3-8 assessments arrives at a similar conclusion, suggesting that a student is likely to meet Kentucky's 3rd grade math expectations if the student scores at the 55th percentile or higher on the Fall administration of MAP in Grade 3 (Northwest Evaluation Association, 2016). The consistency in findings between different methods—the NWEA study uses probit regression analysis—is reassuring.

Figure 4 demonstrates that these new cutoffs are substantially more informative. This figure consists of three panels, each representing a measure of classification accuracy. The performance of publisher suggested cutoffs, unweighted CART cutoffs, and weighted CART cutoffs is represented by white, shaded, and black columns, respectively. Panel A shows the negative predictive power, a measure of how much one may trust a not at-risk prediction. By raising the cutoff on 2nd grade MAP from the 40th to the 60th percentile (unweighted), for example, the negative predictive power improves from 66% to 79%. This means that, among students who are initially predicted to be not at-risk, the proportion of those who later fail to meet 3rd grade math expectations is reduced from one-third to one-fifth. The new cutoff also improves the sensitivity from 62% to 85% (Panel B), suggesting a large drop in the percentage of actual at-risk students who are initially misdiagnosed as not at-risk. The new cutoff can substantially boost schools and teachers' confidence in a no-risk prediction.

CART analysis that assigns more weight to FNs raises the cutoffs even higher (Figure 3). The weighted cutoff for 2nd grade MAP, for example, is the 66th percentile. The corresponding negative predictive power improves to 83% (Panel A in Figure 4) and sensitivity improves to 90% (Panel B in Figure 4). These represent additional improvement in classification accuracy (relative to the unweighted cutoff) of 4 and 5 percentage points. By assigning more weights to FNs, the negative predictive power approaches the 85% benchmark that policymakers set for Florida's early grade reading assessment.

Pushing for higher cutoffs makes not at-risk (i.e., negative) predictions more credible, but it comes at the cost of making at-risk (i.e., positive) predictions less accurate. At the extreme when the cutoff is set at the 100th percentile (and therefore all students are classified as at-risk), a positive classification becomes uninformative. Panel C in Figure 4 shows that the unweighted cutoffs lead to an overall gain in classification accuracy, improving the overall proportion correct by 3-21 percentage points. However, the weighted cutoffs result in little or slightly negative changes in overall proportion correct.[7]

5.2. Adding new variables

The CART analysis helps maximize the amount of information that can be extracted from a single screener score. Relying on a single screener score clearly has its limit, and further improvement in classification accuracy requires additional information. This study explores two options. First, many students in the analytic sample took the screener more than once per grade (Appendix Table A3). For these students, using a second screener score may provide additional insights into a student's growth trajectory and improve classification accuracy. As an example, Figure 5 displays a classification tree that utilizes two 2nd grade iReady math scores

(TestPercentile1 and TestPercentile2) to predict whether a student would perform above or below grade level on 3rd grade math. Three pieces of information are presented in each node.

- The label on the top is the **predicted** performance category on 3rd grade K-Prep math.
- The decimal number in the middle is the sample mean of the **observed** value of the response variable, Y. Y=1 when a student performs below grade level and 0 otherwise. Thus, this statistic represents the proportion of students not meeting state standards on 3rd grade math.
- The percentage at the bottom is the percent of total sample that is partitioned into that node.

Students classified into the bottom left leaf node, for example, are *predicted* to meet 3rd grade expectations for math. However, 21% of these students *actually* failed to meet 3rd grade expectations. Students in this node account for 20% of all students.

The branches are determined using an input variable and a cutoff. CART estimates that the first and most informative split is the second iReady score at the 58th percentile. Students who scored at or above it are sorted down to the left, and they are predicted to meet 3rd grade expectations for math. Otherwise, students are sorted down the right branch. Here, CART determines that additional information from the first iReady score is needed to make a better prediction. It finds that among students who scored below the 58th percentile on the second iReady, those who scored at or above the 53rd percentile on the first iReady are likely to meet 3rd grade math standards. These students are sorted down to the middle leaf node. Otherwise, students are sorted down the right branch, arriving at the leaf node that predicts below grade level performance on math by grade 3.

Figure 6 reports the extent to which classification accuracy can be gained by using two screener scores. For brevity, results are shown for analyses that assign equal weights to FNs and FPs. It is constructed similarly to Figure 4, with white bars representing results using the first screener only and black bars representing results using two screener scores. Overall, Figure 6 shows that a second screener score adds very little to classification accuracy in most cases. In some cases, classification accuracy is lowered relative to when only the first screener score is used, suggesting that a second score may add more noise than signal. To continue the example of 2nd grade iReady, the classification tree presented in Figure 5 improves classification accuracy only marginally (by about 1 percentage point) relative to when only the first iReady score is used to predict 3rd grade math performance.

A second option to improve classification accuracy is to borrow strength from student characteristics that are typically available in administrative records. These include student gender, subsidized meals status, special education status, and race/ethnicity. In most cases, CART analysis shows that these variables add little information. The only exceptions are when 1st grade iReady and Star screener scores are used to predict 3rd grade K-Prep math performance (Figure 7). Among students who scored below the 83rd percentile (but at or above the 48th percentile) on 1st grade iReady (Figure 7, panel A), those who were eligible for subsidized meals are predicted to perform below grade level on 3rd grade K-Prep math and those who were ineligible for subsidized meals are predicted to meet 3rd grade math expectations. Among students who scored between the 62nd and 83rd percentile on 1st grade Star, those who were eligible for subsidized meals are predicted to perform below grade level on 3rd grade K-Prep math; among those who were not eligible for subsidized meals, students who received special

education are predicted to perform below 3rd grade math expectations whereas students who did not receive special education are predicted to perform at or above 3rd grade math expectations.

The added classification nuances in these two cases, however, do not lead to pronounced improvement in overall classification accuracy. Figure 8 (and Table A4 in the appendix) shows that the overall proportion correct increases by 1-2 percentage points only. Although negative predictive power and sensitivity both improve markedly (by 13 and 17 percentage points, respectively) when subsidized meals status is used with 1st grade iReady scores to predict 3rd grade math performance, student characteristics appear to add more noise than information to 1st grade Start performance.

Discussion

Children's math performance is strongly correlated with later life outcomes. Accurately predicting how they will perform on standardized tests is therefore important. There are ongoing debates about the role of standardized tests in today's education systems. For example, there are legitimate concerns about the use of test scores for accountability purposes, and whether test scores have a *causal* relationship with educational attainment and labor market outcomes remains an open question (Goldhaber & Özek, 2018). However, a vast empirical literature demonstrates strong *correlations* between standardized test scores and later outcomes. Reviews of evidence from the United States (e.g., Altonji & Pierret, 2001; Hanushek & Woessmann, 2008; Murnane et al., 2000) conclude that one standard deviation increase in math test scores at the end of high school translates into 12% higher annual earnings. Adolescent math performance, in turn, is strongly predicted by students' early-grade math skills (Goldhaber et al., 2021; Watts et al., 2014). Such association is twice as strong in math as in reading (Duncan et al., 2007), and

math performance at age 7 is shown to be a stronger predictor of socioeconomic outcomes at age 42 than family backgrounds (Ritchie & Bates, 2013).

This study finds that publisher suggested performance cutoffs tend to be too low in the analytic sample. This leads to significant under identification of student needs in math. For example, about one-third of students deemed at or above grade level based on screener cutoffs later scored below proficiency on 3rd grade state assessment in math. Publisher cutoffs are also not sufficiently sensitive, identifying only about one-third of students who performed below 3rd grade expectations. Under identification of student math needs could result in missed opportunities to offer help to students who need it.

The study shows that by optimizing the cutoffs, classification accuracy can be improved substantially. The overall classification accuracy improves by an average of 8 percentage points (with a range of 3-21 percentage points). Importantly, the under identification of at-risk students is reduced by an average of about 16 percentage points (with a range of 6-32 percentage points), and the sensitivity improves by 28 percentage points (with a range of 11-61 percentage points). These improvements are achieved without having to collect new data or administer new assessments, suggesting that tailoring cutoffs to local context could be a highly cost-efficient way for schools and districts to minimize the risk of under identifying students with potential deficiency in math skills.

Depending on state and local policy priorities, the accuracy of no-risk predictions can be further improved, but at the cost of potentially over subscribing supplemental support to students who do not need it. Although missing opportunities to provide support is likely more consequential than providing extra support that is unnecessary, policymakers need to evaluate the tradeoffs based on available resources and opportunity costs of those resources.

Notes

1. These assessments have also become important tools to measure Covid-19 learning loss and recovery (e.g., Lewis et .al, 2021; Tirado, 2021).

2. There are linking studies for Grades 3-8 that examine, for example, the correspondence between cut scores on MAP and the benchmarks on state standardized tests (Northwest Evaluation Association, 2016).

3. Test reports often include many additional metrics, such as growth scores, to help students, teachers, and schools understand performance on specific domains of skills and performance trends. See more details at: https://teach.mapnwea.org/impl/maphelp/Content/Data/SampleReports/StudentProgressReport.htm.

4. See https://www.esboces.org/cms/lib/NY01914091/Centricity/Domain/533/iready-placement-tables-2018-2019.pdf.

5. See https://www.renaissance.com/2016/05/12/giving-meaning-to-test-scores/.

6. Another reason for focusing on FNs is that students predicted to be not at-risk are unlikely to receive supplemental support provided by schools. By contrast, students predicted to be at-risk are more likely to receive interventions. If such interventions were effective, students who are initially low-performing could meet grade-level expectations later. These students would look like FPs when in fact they reflect the success of interventions.

7. The optimal cutoffs estimated by CART, with and without equal weights being assigned to FNs and FPs, are reported in Appendix Table A2 along with measures of classification accuracy.

References

Altonji, J. G., & Pierret, C. R. (2001). Employer Learning and Statistical Discrimination. *Quarterly Journal of Economics, 116*(1): 313–50.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees*. Belmont, CA: Wadsworth International Group.

Compton, D. L., Fuchs, D., Fuchs, L. S., & Bryant, J. D. (2006). Selecting at-risk readers in first grade for early intervention: A two-year longitudinal study of decision rules and procedures. *Journal of Educational Psychology*, *98*(2), 394–409.

Duncan, G. J., Dowsett, C. J., Claessens, A., Magnuson, K., Huston, A. C., Klebanov, P., ... & Sexton, H. (2007). School readiness and later achievement. *Developmental psychology*, *43*(6), 1428.

Fazlul, I., Koedel, C., & Parsons, E. (2022). *A New Framework for Identifying At-Risk Students in Public Schools.* Working Paper No. 261-0122. National Center for Analysis of Longitudinal Data in Education Research (CALDER).

Geary, D.C., Hoard, M. K., Nugent, L., & Bailey, H. D. (2013). Adolescents' functional numeracy is predicted by their school entry number system knowledge. *PLoS ONE, 8*(1): e54651.

Goldhaber, D., & Özek, U. (2019). How much should we rely on student test achievement as a measure of success? *Educational Researcher*, *48*(7), 479–483.

Goldhaber, D., Wolff, M., & Daly, T. (2021). *Assessing the Accuracy of Elementary School Test Scores as Predictors of Students' High School Outcomes*. Working Paper No. 235-0821-2. National Center for Analysis of Longitudinal Data in Education Research (CALDER).

Hanushek, E. A., & Woessmann, L. (2008). The role of cognitive skills in economic development. *Journal of economic literature*, *46*(3), 607–668.

Jordan, N. C., Kaplan, D., Nabors Olah, L., & Locuniak, M. N. (2006). Number sense growth in kindergarten: A longitudinal investigation of children at risk for mathematics difficulties. *Child development*, *77*(1), 153–175.

Kentucky Department of Education. (n.d.). *2021-2022 student performance level cut scores*. https://education.ky.gov/AA/distsupp/Documents/KSA_AKSA_Cut_Scores_21_22.pdf.

Koon, S., Petscher, Y., & Foorman, B.R. (2014). *Using evidence-based decision trees instead of formulas to identify at-risk readers* (REL 2014–036). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Southeast.

Koon, S., Foorman, B., & Galloway, T. (2020). *Identifying North Carolina Students at Risk of Scoring below Proficient in Reading at the End of Grade 3* (REL 2020-030). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Southeast.

Lewis et al. (July 2021). *Learning during COVID-19: Reading and math achievement in the 2020-2021 school year.* Northwest Evaluation Association (NWEA).

Loeb, S., & Bassok, D. (2007). Early childhood and the achievement gap. *Handbook of research in education finance and policy*, 517–534.

Morgan, P. L., Farkas, G., & Wu, Q. (2011). Kindergarten children's growth trajectories in reading and mathematics: Who falls increasingly behind?. *Journal of learning disabilities, 44*(5), 472–488.

Murnane, R. J., Willett, J. B., Duhaldeborde, Y. & Tyler, J. (2000). How Important Are the Cognitive Skills of Teenagers in Predicting Subsequent Earnings. *Journal of Policy Analysis and Management, 19*(4): 547–68.

Northwest Evaluation Association. (2016). *Linking the Kentucky K-PREP assessments to NWEA MAP tests*. Northwest Evaluation Association (NWEA).

Petscher, Y., Fien, H., Stanley, C., Gearin, B., Gaab, N., Fletcher, J.M., & Johnson, E. (2019). *Screening for Dyslexia*. Washington, DC: U.S. Department of Education, Office of Elementary and Secondary Education, Office of Special Education Programs, National Center on Improving Literacy.

Ritchie, S. J., & Bates, T. C. (2013). Enduring links from childhood mathematics and reading achievement to adult socioeconomic status. *Psychological science*, *24*(7), 1301–1308.

Tirado, Andrea. (July 2021). *Review of Literature: COVID-19 Learning Loss and Strategies for Recovery*. Information Capsule Research Services.

Therneau, T. M. & Atkinson, E. J. (2013). *An introduction to recursive partitioning using the RPART routines*. Technical report, Mayo Foundation.

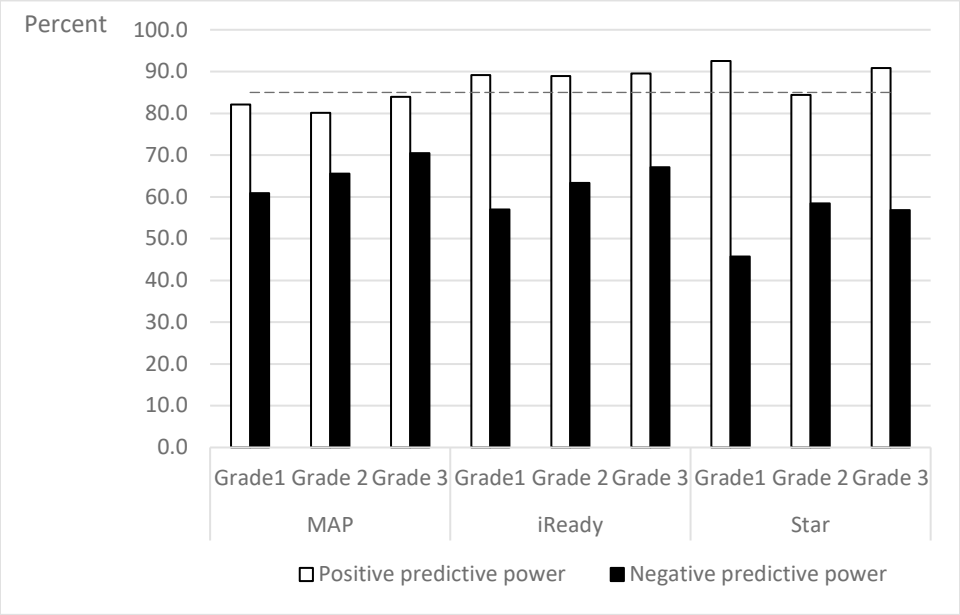
U.S. Department of Education. (2016). *Issue Brief: Early Warning Systems*. U.S. Department of Education, Office of Planning, Evluation and Policy Development, Policy and Program Studies Service.

Watts, T. W., Duncan, G. J., Siegler, R. S., & Davis-Kean, P. E. (2014). What's past is prologue: Relations between early mathematics knowledge and high school achievement. *Educational Researcher, 43*(7), 352–360.

Xu, Z., Ozek, U., Levin, J., & Lee, D. H. (2023). *Effects of Large-Scale Early Math Interventions on Student Outcomes: Evidence from Kentucky's Math Achievement Fund*.

CALDER Working Paper 279-0323. Washington, DC: National Center for Analysis of Longitudinal Data in Education Research.

Figure 1. Negative and positive predictive power, by screener type and grade



Note: The dashed line represents the target predictive power (85%) that researchers and policymakers have suggested. Positive predictive power is the percentage of students who were initially predicted to be at risk based on screener scores who actually scored in the novice or apprentice range on K-Prep 3rd grade math later. Negative predictive power is the percentage of students who were initially predicted to be not at risk based on screener scores who actually scored in the proficient or distinguished range on K-Prep 3rd grade math later.

Figure 2. Sensitivity of screeners, by screener type and grade

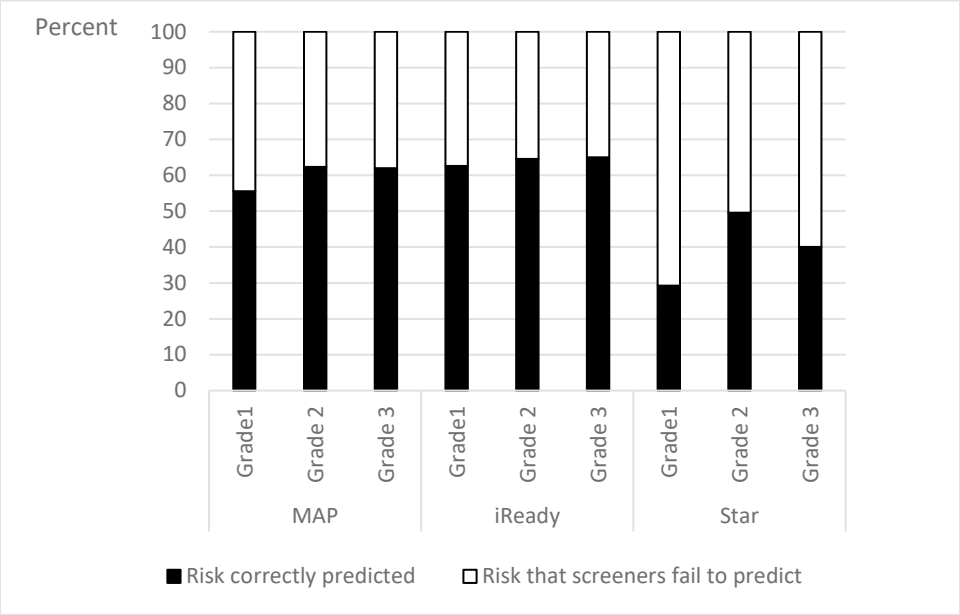
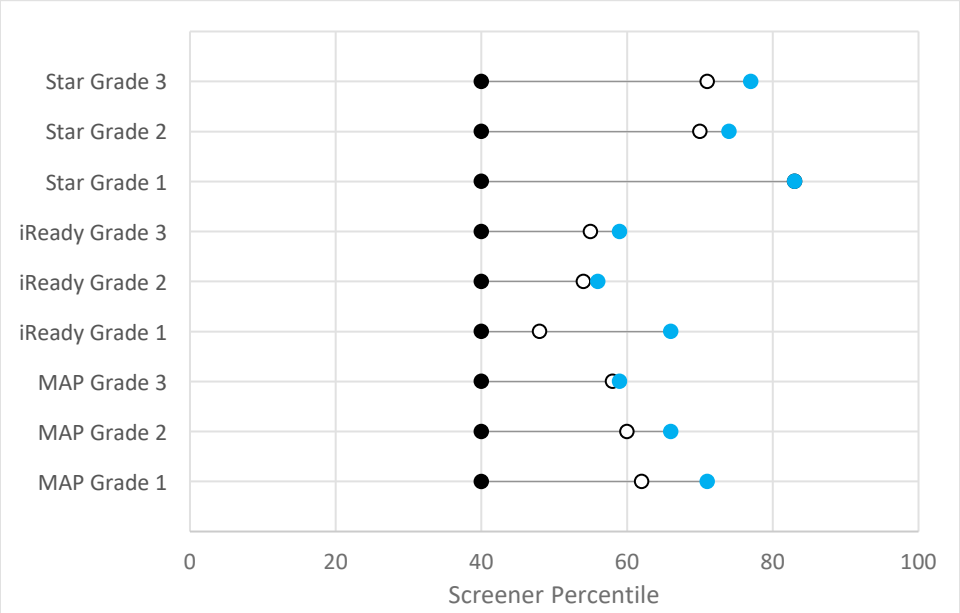
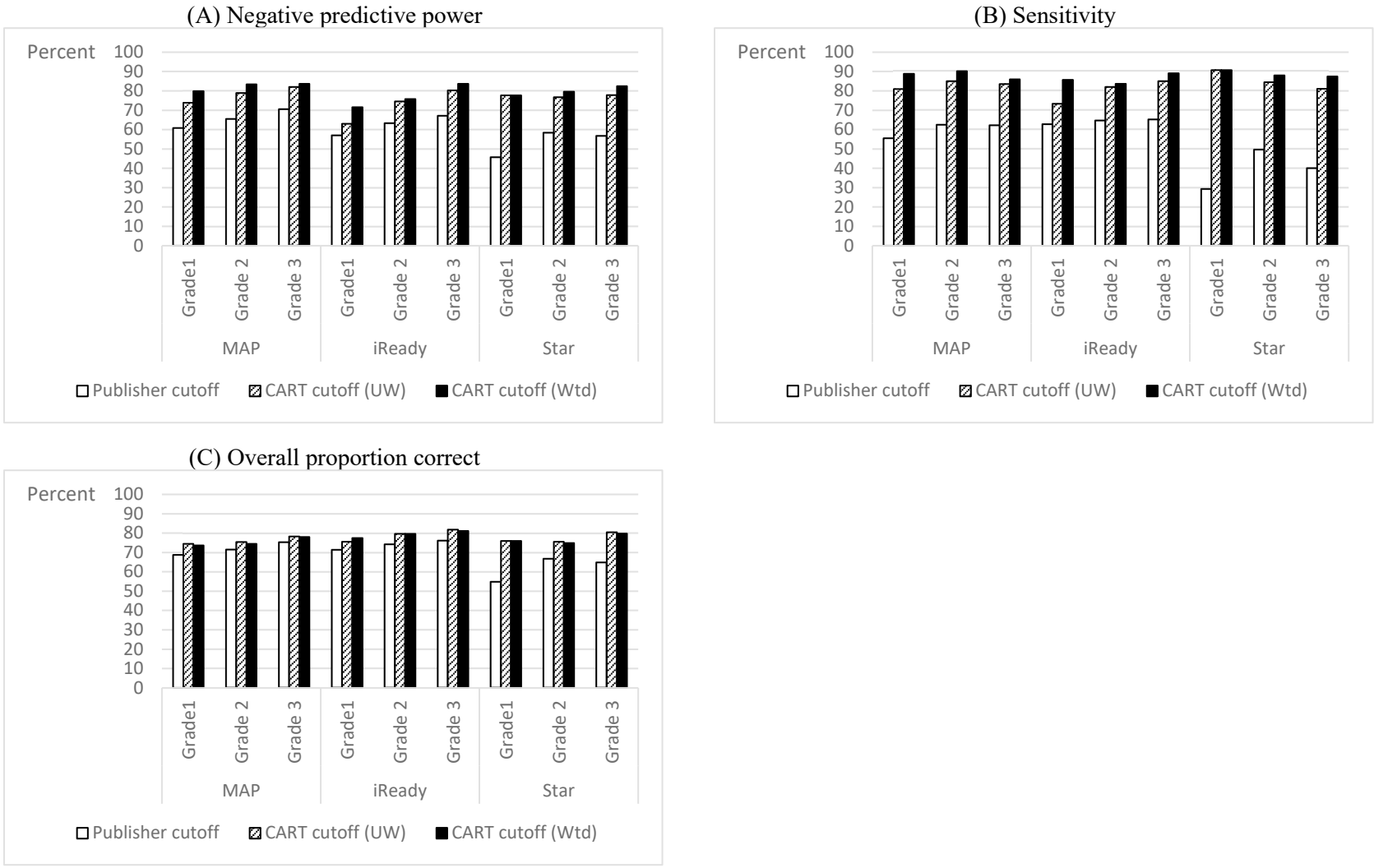


Figure 3. Comparisons of publisher suggested cutoffs for grade-level performance and CART-optimized cutoffs, by screener type and grade.



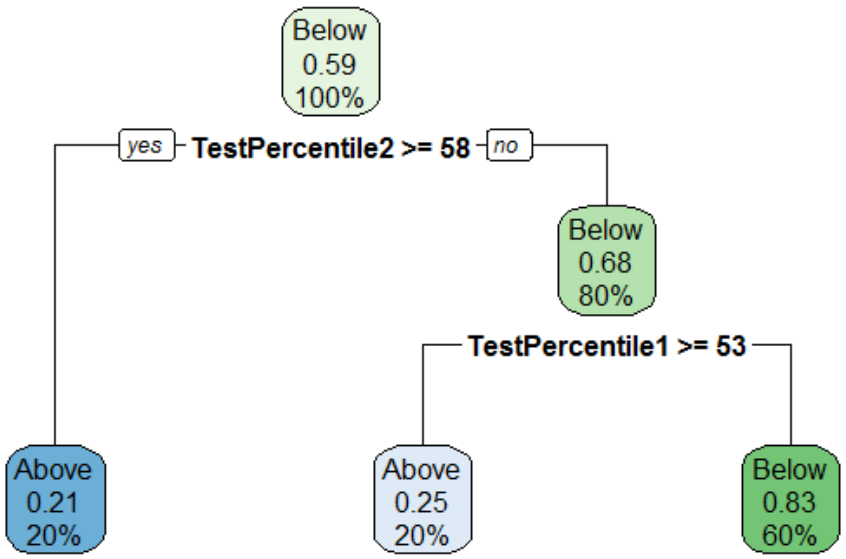
Note: Black dots represent publisher suggested cutoff. Open circles represent CART estimated cutoff that gives equal weights to false positives and false negatives. Blue dots represent CART estimated cutoff that gives twice as much weight to false negatives as to false positives.

Figure 4. Measures of classification accuracy, by screener type, grade, and cutoff



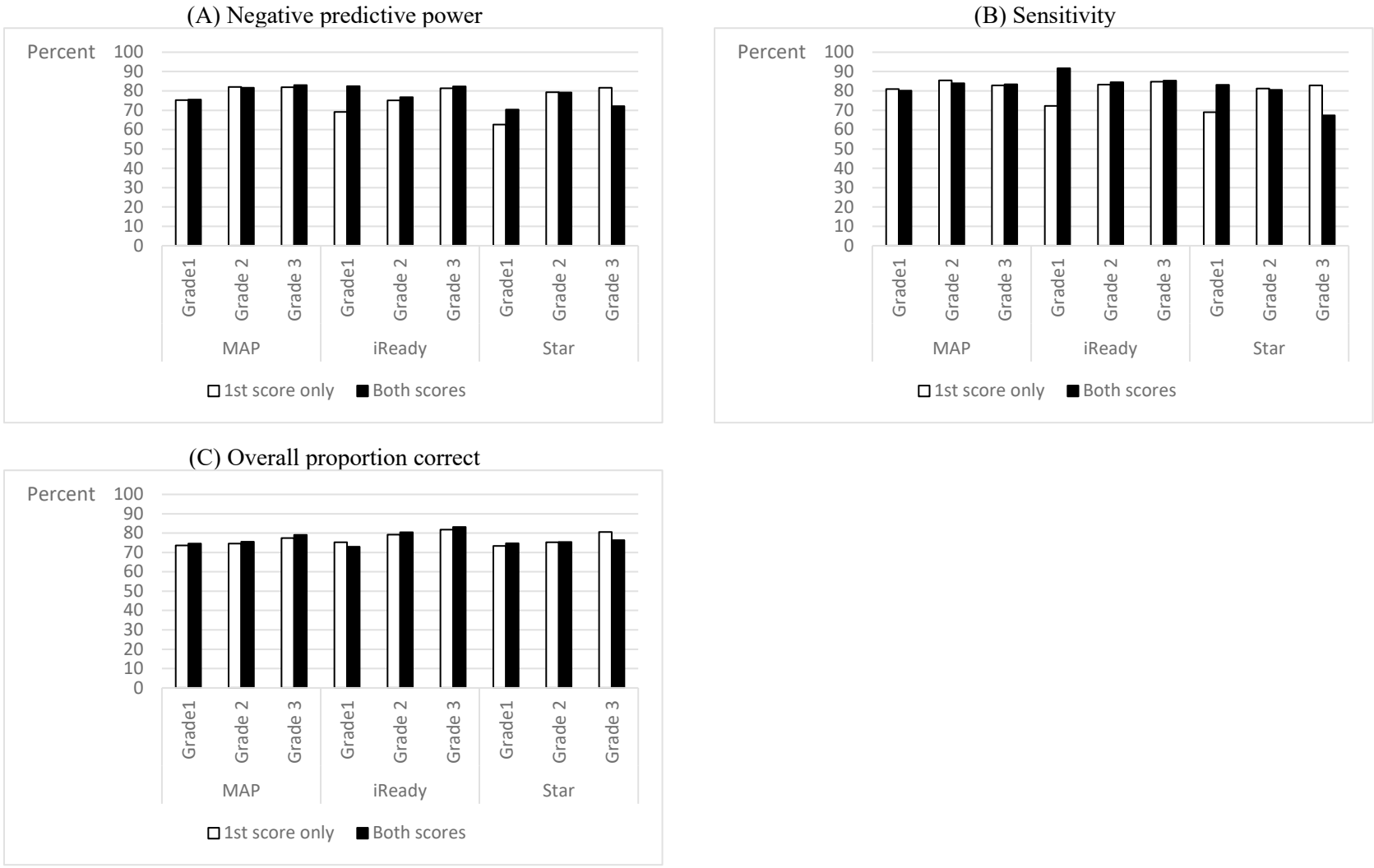
Note: "CART cutoff (UW)" is based on CART analysis that assigns equal weights to false negatives and false positives. "CART cutoff (Wtd)" is based on CART analysis that assigns twice as much weight to false negatives as to false positives.

Figure 5. Classification tree using 2nd grade iReady screener score to predict 3rd grade math proficiency



Note: In each node, the label on the top is the *predicted* performance category on 3rd grade K-Prep math, the decimal number in the middle is the *observed* proportion of students who perform below grade level on 3rd grade K-Prep math, and the bottom percentage is the percent of total sample that is partitioned into the node. The first split is based on whether a student scored at or above the 58th percentile on the second iReady screener test, and the second split is based on whether a student scored at or above the 53rd percentile on the first iReady screener test. For each split, students who meet the criterion are branched to the left. Otherwise, they are branched to the right.

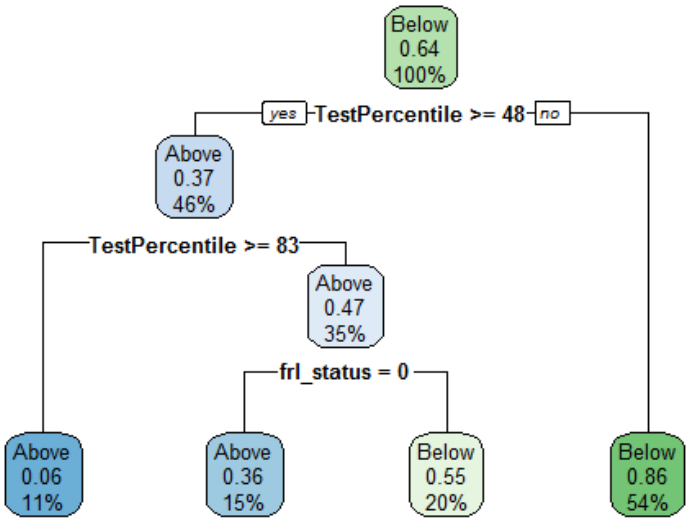
Figure 6. Changes in classification accuracy by adding a second screener score, unweighted, by screener type and grade



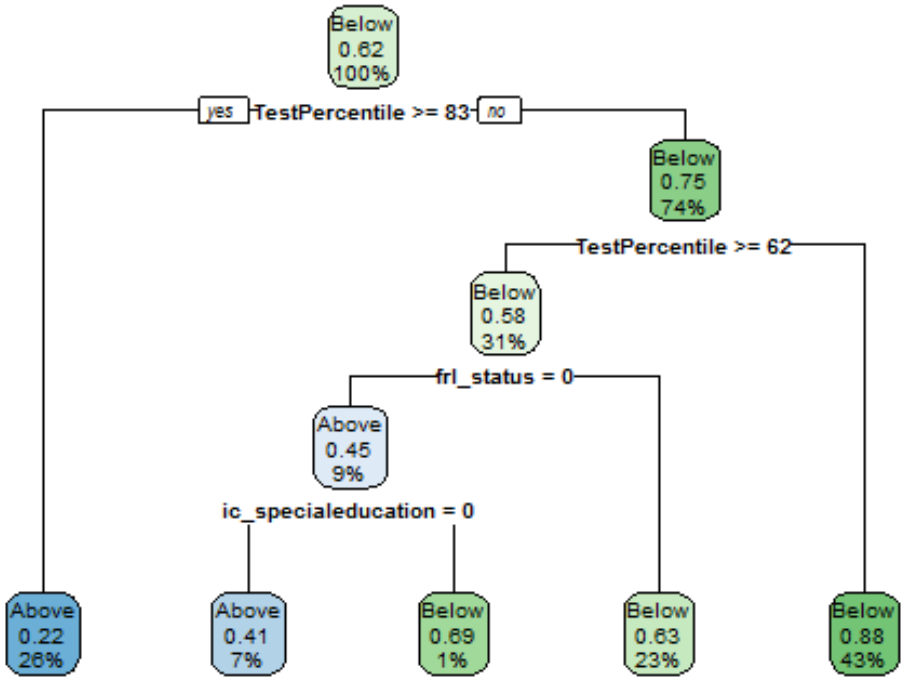
Note: Analyses were conducted using students with 2 or more screener scores in the same grade. Unweighted means that equal weights were assigned to false positives and false negatives.

Figure 7. Classification trees using 1st grade screener score with student characteristics to predict 3rd grade math proficiency

(A) iReady (1st grade)

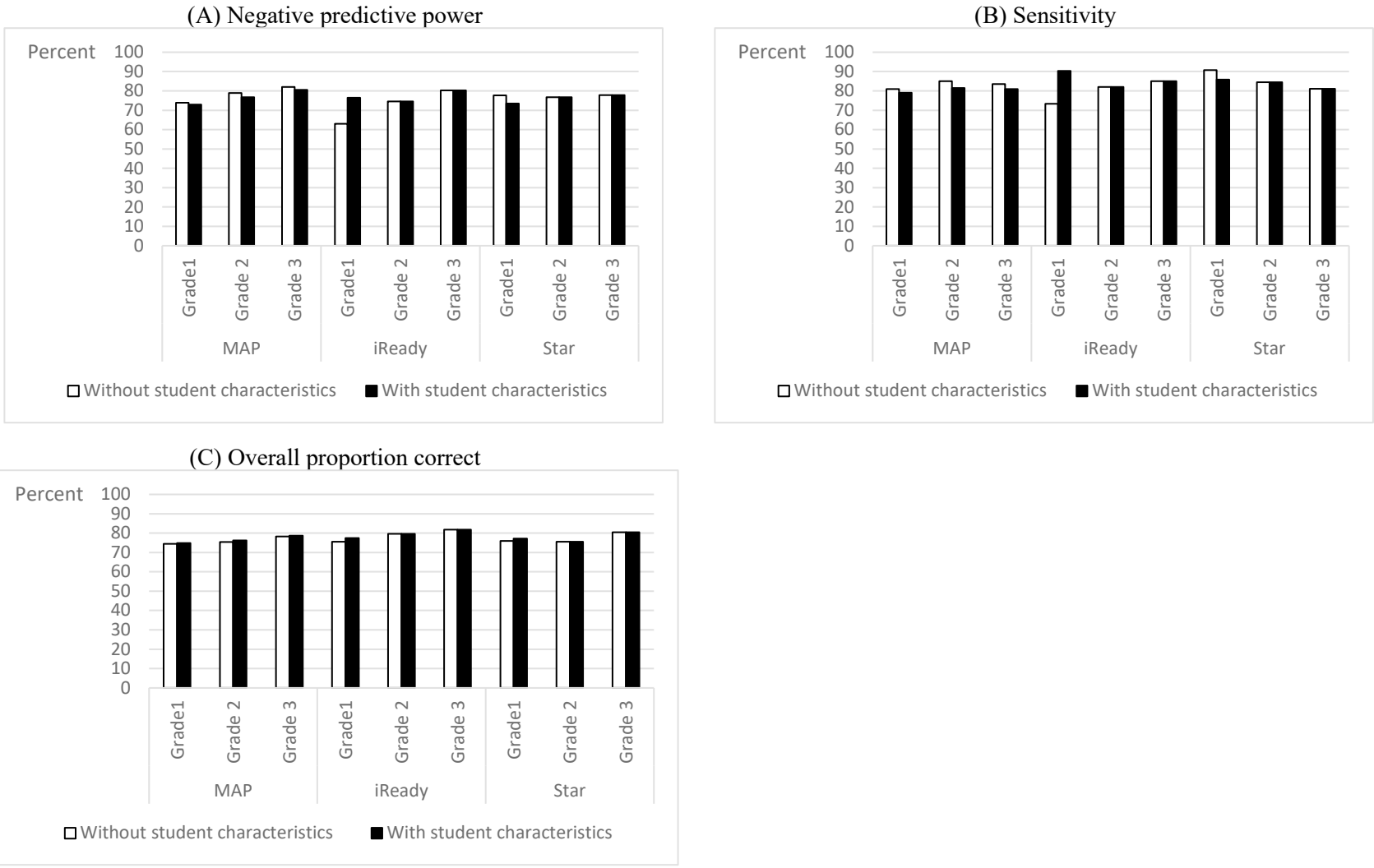


(B) Star (1st grade)



Note: In each node, the label on the top is the predicted performance category on 3rd grade K-Prep math, the decimal number in the middle is the observed proportion of students who perform below grade level on 3rd grade K-Prep math, and the bottom percentage is the percent of total sample that is partitioned into the node. *TestPercentile* is student screener score, *frl_status* equals 1 if a student was eligible for subsidized meals and 0 otherwise. *ic_specialeducation* equals 1 if a student received special education and 0 otherwise.

Figure 8. Changes in classification accuracy by adding student characteristics, unweighted, by screener type and grade



Note: Unweighted means that equal weights were assigned to false positives and false negatives.

Table 1. Characteristics of elementary schools, by whether they administered math screeners among K-3 students: averaged across 2015-2016 and 2019-2020 school years

	Elementary schools that administered K-3 math screeners		Elementary schools that did not administer K-3 math screeners	
Characteristic	Mean	Standard deviation	Mean	Standard deviation
Female	0.484	0.026	0.471	0.076
Proportion of students				
Black	0.095***	0.115	0.151***	0.189
Hispanic	0.064	0.071	0.073	0.087
White	0.934***	0.096	0.871***	0.181
Female	0.484	0.026	0.471	0.076
Subsidized meals eligible	0.686	0.149	0.678	0.187
English learner	0.039	0.063	0.053	0.090
Special education	0.214	0.051	0.224	0.135
School characteristics				
Enrollment size	375	135	362	179
Rural	0.545	0.500	0.475	0.500
Town	0.265*	0.443	0.147*	0.354
Suburban	0.091	0.289	0.139	0.347
Urban	0.083***	0.277	0.226***	0.419
Number of unique students	97,921		366,368	
Number of unique schools	145		647	

* Significant at $p < .05$; *** significant at $p < .001$.

Note: One screener school has no school characteristics information in the Kentucky Department of Education data.

Table 2. Number of students with both a math screener score and a 3rd grade K-Prep math score, by screener type and grade: 2015-2016 and 2019-2020

Screener	K	Grade 1	Grade 2	Grade 3
MAP	3073	**10653**	**10459**	**13079**
iReady	216	**992**	**2024**	**2906**
Star	206	**1024**	**1907**	**2307**
Discovery Ed	346	849	922	990
Other	138	243	245	234

Note: Bolded cells constitute the main analytic sample in this study.

Table 3. An example of 2X2 contingency table

	Grade 3 K-Prep math (actual)		
Screener (prediction)	**Below grade level**	**At or above grade level**	
Below grade level	**True positive (TP)**	**False positive (FP)**	Positive predictive power = TP/(TP+FP)
At or above grade level	**False negative (FN)**	**True negative (TN)**	Negative predictive power = TN/(TN+FN)
	Sensitivity = TP/(TP+FN)	Specificity = TN/(TN+FP)	Overall proportion correct = (TP+TN)/(TP+TN+FP+FN)

Source: Based on Koon, Petscher, & Foorman (2014)

Appendix:

Table A1a. Performance level congruence between 1st grade math screener and 3rd grade K-Prep math test, by screener type.

	3rd grade K-Prep math (Actual)	
1st grade screener (Prediction)	Below grade level	At or above grade level
MAP		
Below grade level	3264 (30.64%)	712 (6.68%)
At or above grade level	2614 (24.54%)	4063 (38.14%)
iReady		
Below grade level	395 (39.82%)	48 (4.84%)
At or above grade level	236 (23.79%)	313 (31.55%)
Star		
Below grade level	185 (18.07%)	15 (1.46%)
At or above grade level	447 (43.65%)	377 (36.82%)

Table A1b. Performance level congruence between 2nd grade math screener and 3rd grade K-Prep math test, by screener type.

	3rd grade K-Prep math (Actual)	
2nd grade screener (Prediction)	Below grade level	At or above grade level
MAP		
Below grade level	3484 (33.31%)	863 (8.25%)
At or above grade level	2106 (20.14%)	4006 (38.30%)
iReady		
Below grade level	772 (38.14%)	96 (4.74%)
At or above grade level	424 (20.95%)	732 (36.17%)
Star		
Below grade level	526 (27.58%)	97 (5.09%)
At or above grade level	534 (28.00%)	750 (39.33%)

Table A1c. Performance level congruence between 3rd grade math screener and 3rd grade K-Prep math test, by screener type.

	3rd grade K-Prep math (Actual)	
3rd grade screener (Prediction)	Below grade level	At or above grade level
MAP		
Below grade level	4003 (30.61%)	764 (5.84%)
At or above grade level	2456 (18.78%)	5856 (44.77%)
iReady		
Below grade level	1057 (36.37%)	123 (4.23%)
At or above grade level	568 (19.55%)	1158 (39.85%)
Star		
Below grade level	504 (21.85%)	51 (2.21%)
At or above grade level	756 (32.77%)	996 (43.17%)

Table A2. CART estimated screener cutoffs and comparisons of classification accuracy between publisher-suggested and optimal cutoffs, by screener type and grade.

	CART cutoff		Overall proportion correct			Negative predictive power			Sensitivity		
Screener (grade)	Unweighted	Weighted	Publisher cutoff	CART cutoff (UW)	CART cutoff (Wtd)	Publisher cutoff	CART cutoff (UW)	CART cutoff (Wtd)	Publisher cutoff	CART cutoff (UW)	CART cutoff (Wtd)
MAP											
1st Grade	62	71	0.688	0.744	0.736	0.609	0.739	0.799	0.555	0.809	0.888
2nd Grade	60	66	0.716	0.754	0.744	0.655	0.789	0.834	0.623	0.851	0.902
3rd Grade	58	59	0.754	0.783	0.780	0.705	0.820	0.837	0.620	0.835	0.860
iReady											
1st Grade	48	66	0.714	0.755	0.774	0.570	0.630	0.716	0.626	0.734	0.857
2nd Grade	54	56	0.743	0.796	0.796	0.633	0.746	0.758	0.645	0.821	0.837
3rd Grade	55	59	0.762	0.818	0.811	0.671	0.803	0.837	0.650	0.850	0.891
Star											
1st Grade	83	83	0.549	0.760	0.760	0.458	0.777	0.777	0.293	0.907	0.907
2nd Grade	70	74	0.669	0.755	0.749	0.584	0.768	0.796	0.496	0.845	0.880
3rd Grade	71	77	0.650	0.805	0.797	0.568	0.778	0.824	0.400	0.811	0.875

Table A3. Comparisons of classification accuracy between using one and two screener scores, unweighted, by screener type and grade.

Screener (grade)	Overall proportion correct		Negative predictive power		Sensitivity		Sample size
	1st score only	Both scores	1st score only	Both scores	1st score only	Both scores	
MAP							
1st Grade	0.736	0.746	0.753	0.755	0.810	0.801	7078
2nd Grade	0.746	0.756	0.821	0.816	0.855	0.840	6697
3rd Grade	0.774	0.791	0.819	0.830	0.829	0.834	7891
iReady							
1st Grade	0.753	0.730	0.691	0.824	0.723	0.917	721
2nd Grade	0.792	0.804	0.751	0.768	0.833	0.845	1069
3rd Grade	0.818	0.832	0.814	0.823	0.848	0.853	1194
Star							
1st Grade	0.734	0.747	0.626	0.703	0.690	0.831	489
2nd Grade	0.753	0.754	0.793	0.792	0.812	0.806	866
3rd Grade	0.806	0.764	0.816	0.721	0.828	0.674	905

Table A4. Comparisons of classification accuracy between using one screener score with and without student characteristics, unweighted, by screener type and grade.

Screener (grade)	Overall proportion correct		Negative predictive power		Sensitivity		Sample size
	Without student characteristics	With student characteristics	Without student characteristics	With student characteristics	Without student characteristics	With student characteristics	
MAP							
1st Grade	0.744	0.749	0.739	0.730	0.809	0.790	10653
2nd Grade	0.754	0.762	0.789	0.768	0.851	0.815	10459
3rd Grade	0.783	0.786	0.820	0.805	0.835	0.810	13079
iReady							
1st Grade	0.755	0.775	0.630	0.765	0.734	0.903	992
2nd Grade	0.796	0.796	0.746	0.746	0.821	0.821	2024
3rd Grade	0.818	0.818	0.803	0.803	0.850	0.850	2906
Star							
1st Grade	0.760	0.772	0.777	0.735	0.907	0.858	1024
2nd Grade	0.755	0.755	0.768	0.768	0.845	0.845	1907
3rd Grade	0.805	0.805	0.778	0.778	0.811	0.811	2307