IES Institute of Education Sciences

APPENDIX
August 2020

Regional Educational Laboratory Southeast
At Florida State University

Identifying North Carolina Students at Risk of Scoring below Proficient in Reading at the End of Grade 3

See https://go.usa.gov/xfAZa for the full report.

Appendix A. Methods

The appendix provides details on the study sample and the classification and regression tree (CART) analyses used in the study.

Sample

The dataset consisted of students who took the state reading assessment at the end of grade 3 in 2017/18 for the first time and had data from the middle-of-the-year interim assessments in kindergarten in 2014/15 as well as data from the beginning-of-the-year and middle-of-the-year interim assessments in grade 1 in 2015/16, the beginning-of-the-year interim assessments in grade 2 in 2016/17, and the beginning-of-the-year interim assessments in grade 3 in 2017/18. The study sample included 91,855 students, or about 77 percent of the 120,029 grade 3 students statewide in 2017/18. Students may have been excluded from the study sample because they transferred in or out of the public school system during the study period, because they were absent during the testing sessions, or because they were exempt from testing due to learning disabilities or limited English proficiency. The demographic characteristics of the statewide population and the study sample were very similar (table A1). However, because the study sample excluded 23 percent of grade 3 students statewide and because the sample was not randomly selected, there may be important unmeasured differences between the study sample and the statewide population. Thus, the study findings may not generalize to the statewide population of grade 3 students.

Table A1. Demographic characteristics of the statewide population of grade 3 students and the study sample, 2017/18

	Statewide population of grade 3 students (*N* = 120,029)		Study sample (*n* = 91,955)		Difference between the statewide population and the study sample (standard deviation units)
Characteristic	Number	Percent	Number	Percent	
Male students	61,579	51	46,788	51	0.00
Female students	58,450	49	45,067	49	0.00
Special education students	14,216	12	9,900	11	0.06
English learner students	12,486	10	9,351	10	0.00
Economically disadvantaged students	57,016	48	44,179	48	0.00
Black students	30,984	26	23,560	26	0.00
Hispanic students	22,657	19	16,985	19	0.00
White students	55,278	46	43,243	47	–0.02

Source: Authors' analysis of data from the North Carolina Department of Public Instruction.

Classification and regression tree analyses

CART analysis is a statistical technique that classifies individuals into mutually exclusive subgroups and presents the results visually in a decision tree (Breiman, Friedman, Olshen, & Stone, 1984). It does this by identifying the best predictors and levels of those predictors that most efficiently split the sample into the most similar subgroups of individuals. A variable can appear in a CART model multiple times with different cutoffs because the search for the single variable that best divides the data includes all variables at each split point (Therneau & Atkinson, 2013). Previous studies have found CART results to be consistent with those from logistic regression (Koon, Petscher, & Foorman, 2014) and easier to understand because of CART's graphic format.

The predictors in the CART analyses for the current study were scores from administrations of interim assessments at five points in time (see table 1 in the main text) and from the North Carolina Beginning-of-Grade 3 English Language Arts/Reading Test (BOG3 assessment), a separate state assessment. The interim assessments were mCLASS 3D Reading assessments, which consist of Acadience™ Reading (Dynamic Measurement Group, 2018; formerly DIBELS Next®) measures and a set of reading comprehension passages called Text Reading and Comprehension (Amplify, 2015). All the assessments are aligned to the intended curriculum in North Carolina and have demonstrated good reliability (table A2).

Acadience™ Reading provides a raw score for individual assessments and a composite score that is the sum of all Acadience™ Reading assessment scores given at that time point in that grade. Only the individual assessment scores were used as predictors in the CART models. Raw scores are the number of items that a student answered correctly in one minute, unless noted otherwise here. For the Nonsense Word Fluency assessment there are two raw scores: Correct Letter Sounds and Whole Words Read. For the Oral Reading Fluency assessment there are three raw scores: Words Correct Per Minute, Accuracy (the percentage of words correct), and Retell. For the Daze assessment the raw score is the number of correct words out of the three words in the maze boxes circled in three minutes. For the Text Reading and Comprehension assessment the score is the text difficulty level at which a student was proficient (that is, Print Concepts, Reading Behaviors, or a level from A to Z). The reliability estimates for the assessment scores are above .80, except for Phonemic Segmentation Fluency in kindergarten (.70) and grade 1 (.78), Oral Reading Fluency Retell in grade 2 (.68; Dynamic Measurement Group, 2019), and Text Reading and Comprehension Reading Behaviors in kindergarten (0.62; Amplify, 2015; see table A2).

Table A2. Score range and reliability of assessment scores, by grade

Assessment score	Score range	Reliability			
		Kindergarten	Grade1	Grade 2	Grade 3
First Sound Fluency	0–60	.93[b]	na	na	na
Letter Name Fluency	0–110	.95[b]	nr	na	na
Phonemic Segmentation Fluency	0–80	.70[b]	.78[b]	na	na
Nonsense Word Fluency Correct Letter Sounds	0–143	.88[b]	.94[b]	nr	na
Nonsense Word Fluency Whole Words Read	0–50	na	.96[b]	nr	na
Oral Reading Fluency Words Correct Per Minute	0–254	na	.98[b]	.96[b]	.97[b]
Oral Reading Fluency Accuracy	0–100	na	.88[b]	.83[b]	.80[b]
Oral Reading Fluency Retell	0–94	na	na	.68[b]	.81[b]
Daze	0–51	na	Na	na	.93[b]
Text Reading and Comprehension					
Print Concepts	[a]	.80[c]	na	na	na
Reading Behaviors	[a]	.62[c]	na	na	na
A–Z	[a]	.93[c]	.97[c]	.93[c]	.96[c]
BOG3 assessment (composite)	408–461	na	na	na	.91[c]

na is not applicable because the assessment is not administered at the given grade level in North Carolina. nr is not reported. BOG3 is the North Carolina Beginning-of-Grade 3 English Language Arts/Reading Test.

a. Text Reading and Comprehension results are reported on a non-numerical scale in which the lowest level is Print Concepts, followed by Reading Behaviors, then a scale of levels from A to Z.

b. Alternate form reliability estimate reported by the test publisher.

c. Internal consistency reliability estimate reported by the test publisher.

Source: Amplify, 2015; Dynamic Measurement Group, 2019; North Carolina Department of Public Instruction, 2014.

A small proportion of students in the study sample were missing data for some assessment scores (ranging from 0 percent in kindergarten to 3 percent in grade 3; table A3). In those cases the CART model used Text Reading and

Comprehension scores in place of the splitting variable to predict the outcome, if the splitting variable was not the Text Reading and Comprehension score.

Table A3. Missing data, by model and assessment score, 2014/15–2017/18

Model and predictor	Missing statistics (*n* = 91,855) Number	Percent
Kindergarten middle-of-the-year interim assessments		
First Sound Fluency	0	0.0
Letter Name Fluency	0	0.0
Phonemic Segmentation Fluency	0	0.0
Nonsense Word Fluency Correct Letter Sounds	0	0.0
Text Reading and Comprehension	0	0.0
State reading assessment at the end of grade 3 (outcome variable)	0	0.0
Grade 1 beginning-of-the-year interim assessments		
Letter Name Fluency	1,581	1.7
Phonemic Segmentation Fluency	1,581	1.7
Nonsense Word Fluency Correct Letter Sounds	1,581	1.7
Nonsense Word Fluency Whole Words Read	1,581	1.7
Text Reading and Comprehension	0	0.0
State reading assessment at the end of grade 3 (outcome variable)	0	0.0
Grade 1 middle-of-the-year interim assessments		
Nonsense Word Fluency Correct Letter Sounds	1,575	1.7
Nonsense Word Fluency Whole Words Read	1,575	1.7
Oral Reading Fluency Words Correct Per Minute	1,575	1.7
Oral Reading Fluency Accuracy	1,575	1.7
Text Reading and Comprehension	0	0.0
State reading assessment at the end of grade 3 (outcome variable)	0	0.0
Grade 2 beginning-of-the-year interim assessments		
Nonsense Word Fluency Correct Letter Sounds	2,315	2.5
Nonsense Word Fluency Whole Words Read	2,315	2.5
Oral Reading Fluency Words Correct Per Minute	2,315	2.5
Oral Reading Fluency Accuracy	2,315	2.5
Text Reading and Comprehension	0	0.0
State reading assessment at the end of grade 3 (outcome variable)	0	0.0
Grade 3 beginning-of-the-year interim assessments		
Oral Reading Fluency Words Correct Per Minute	2,762	3.0
Oral Reading Fluency Accuracy	2,762	3.0
Oral Reading Fluency Retell	2,762	3.0
Daze	2,762	3.0
Text Reading and Comprehension	0	0.0
State reading assessment at the end of grade 3 (outcome variable)	0	0.0
Grade 3 beginning-of-the-year interim assessments and BOG3 assessment		
Oral Reading Fluency Words Correct Per Minute	2,762	3.0
Oral Reading Fluency Accuracy	2,762	3.0
Oral Reading Fluency Retell	2,762	3.0
Daze	2,762	3.0
Text Reading and Comprehension	0	0.0
BOG3 assessment (composite score)	0	0.0
State reading assessment at the end of grade 3 (outcome variable)	0	0.0

BOG3 assessment is the North Carolina Beginning-of-Grade 3 English Language Arts/Reading Test.
Source: Authors' analysis of data from the North Carolina Department of Public Instruction.

The study team coded scores on the state reading assessment at the end of grade 3 (the outcome variable) to indicate whether a student met the proficiency standard. Scores at or above the standard were coded 1 for proficient, and scores below the standard were coded 0 for not proficient. The dataset was split into a calibration dataset used to build the CART models, consisting of a random sample of 73,607 students (approximately 80 percent of the study sample), and a validation dataset used to test the CART models, consisting of the remaining 18,248 students (approximately 20 percent of the study sample). An 80/20 split is a common division in statistical learning and data mining when conducting cross-validation (Salford Systems, n.d.). CART analyses were run using the Recursive Partitioning and Regression Trees package (rpart; R 3.5.1 package). SPSS was used to split the sample and to report on all descriptive statistics.

A minimum split size of 100 students was specified in all models, so that all decision rules would apply to at least 100 students. In addition, tenfold cross-validation was specified for use in evaluating the quality of the prediction tree and determining the appropriate minimum complexity parameter, which is the minimum improvement in the model (relative error) required for each node (Breiman et al., 1984; Kohavi, 1995). CART analysis accommodates the use of both continuous and categorical predictors without additional specifications.

As with other statistical methods, the principle of parsimony is applicable to CART models. This principle suggests that the simplest model that fits the data is often the best model. In a CART model this principle is applied by pruning the decision tree using model specifications so that the resulting tree is not overly specific to the sample data. Each additional split in a tree adds complexity to the model being estimated. To control the size of the tree, the study team specified a complexity parameter. The nodes that do not add to model improvement (in other words, do not meet the criterion of minimum improvement) are redundant and can be pruned. In deciding the complexity parameter, the study team consulted plots of the cross-validation relative error against minimum complexity parameter values and a table of cross-validation results. The default minimum complexity parameter value to prune a tree is .01, which means that each additional split must reduce the relative error by 1 percent. Generally, a smaller complexity parameter leads to a bigger tree with greater complexity, while a larger complexity parameter leads to a smaller tree with less complexity. For each model the complexity parameter values that provided the best fit to the data using the fewest splits in the tree are as follows: kindergarten middle-of-the-year interim assessments, 0.044; grade 1 beginning-of-the-year interim assessments, 0.0048; grade 1 middle-of-the-year interim assessments, 0.0086; grade 2 beginning-of-the-year interim assessments, 0.007; grade 3 beginning-of-the-year interim assessments, 0.0054; grade 3 beginning-of-the-year interim assessments with the BOG3 assessment, 0.051.

The classification rules were applied to the validation dataset to predict group membership and to derive the classification table (table A4).

Table A4. Classification table based on the validation dataset, 2014/15–2017/18

Predicted classification by model	Proficiency status on the state reading assessment at the end of grade 3	
	Not proficient	Proficient
Kindergarten middle-of-the-year interim assessments		
At risk	4,988	2,797
Not at risk	2,914	7,549
Grade 1 beginning-of-the-year interim assessments		
At risk	5,181	2,179
Not at risk	2,721	8,167
Grade 1 middle-of-the-year interim assessments		
At risk	4,849	1,498
Not at risk	3,053	8,848
Grade 2 beginning-of-the-year interim assessments		
At risk	5,136	1,564
Not at risk	2,766	8,782
Grade 3 beginning-of-the-year interim assessments		
At risk	5,426	1,593
Not at risk	2,476	8,753
Grade 3 beginning-of-the-year interim assessments and BOG3 assessment		
At risk	6,544	1,692
Not at risk	1,358	8,654

BOG3 is the North Carolina Beginning-of-Grade 3 English Language Arts/Reading Test.
Note: Results are based on the validation dataset (*n* = 18,248).
Source: Authors' analysis of data from the North Carolina Department of Public Instruction.

The study team used the classification table to calculate the percentage of below-proficient students correctly identified as at risk (referred to as predictive ability in this study and referred to as the percentage of true positives or a model's sensitivity in the literature), the percentage of proficient students correctly identified as not at risk (commonly referred to as the percentage of true negatives or a model's specificity), and the overall percentage of students correctly identified (the number of below-proficient students correctly identified as at risk plus the number of proficient students correctly identified as not at risk, divided by the total number of students; table A5). The study team also calculated the *R*-squared for each model, which represents the reduction in the relative error (rather than the percentage of explained variance or the coefficient of determination, as in a regression context; Steinberg, 2013).

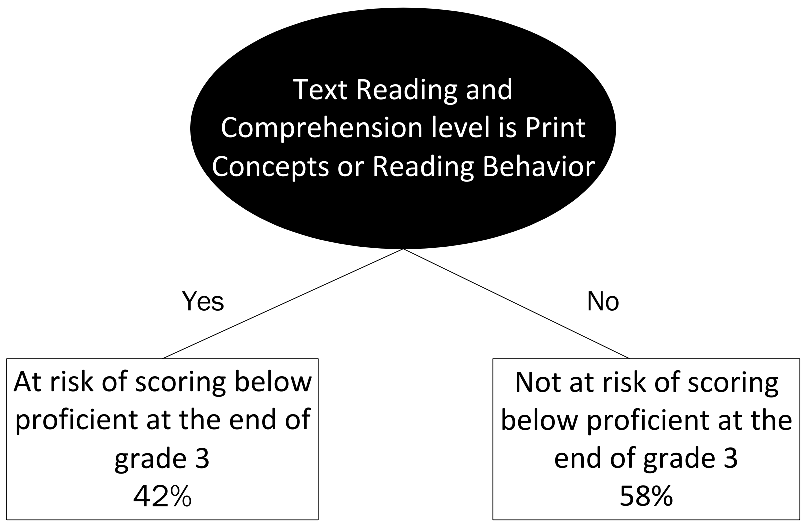
Table A5. Classification accuracy results, by model, 2014/15–2017/18

Model	Percentage of true positives (predictive ability or sensitivity)	Percentage of true negatives (specificity)	Overall percentage correctly identified	*R*-squared
Kindergarten middle-of-the-year interim assessments	63	73	69	0.28
Grade 1 beginning-of-the-year interim assessments	66	79	73	0.37
Grade 1 middle-of-the-year interim assessments	61	86	75	0.44
Grade 2 beginning-of-the-year interim assessments	65	85	76	0.46
Grade 3 beginning-of-the-year interim assessments	69	85	78	0.49
Grade 3 beginning-of-the-year interim assessments and BOG3 assessment	83	84	83	0.61

BOG3 is the North Carolina Beginning-of-Grade 3 English Language Arts/Reading Test.
Note: Results are based on the validation dataset (*n* = 18,248).
Source: Authors' analysis of data from the North Carolina Department of Public Instruction.

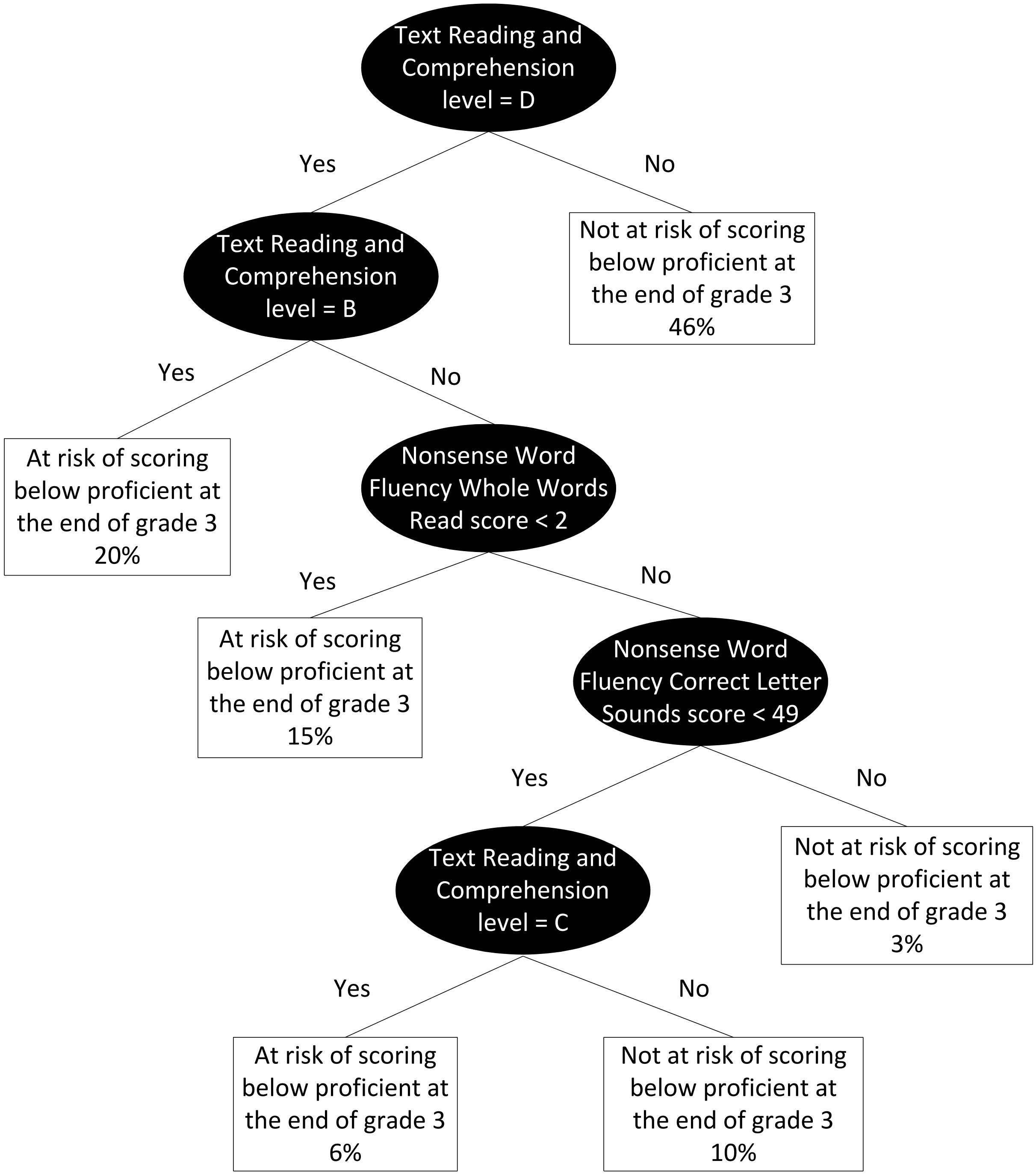
The model that included the beginning-of-the-year interim assessments in grade 3 and the BOG3 assessment as potential predictors was the only model with adequate predictive ability to identify students with reading difficulties at the end of grade 3, and the pruned decision tree is presented in the main text (see figure 1). While the model considered the use of the interim assessments as predictors, they were not selected as predictors in the pruned decision tree because they did not increase the predictive ability above that of the BOG3 alone. Decision trees for the other models, which did not meet the study criterion for predictive ability, are presented in figures A1–A5.

Figure A1. Decision tree for classifying North Carolina students as at risk of scoring below proficient on the state reading assessment at the end of grade 3 based on scores on the middle-of-the-year interim assessments in kindergarten, 2014/15 and 2017/18



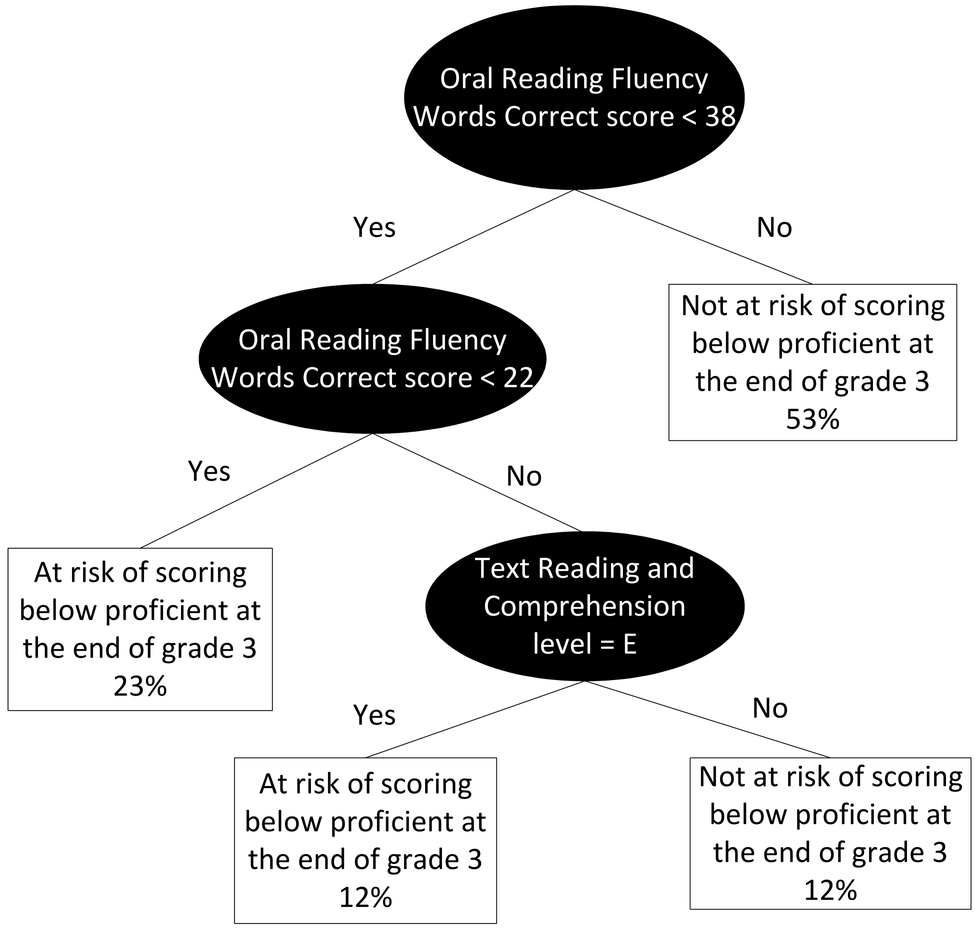
Note: Print Concepts and Reading Behavior are the two lowest levels of performance possible on the Text Reading and Comprehension assessment; see table A2 for score ranges for the middle-of-the-year interim assessments in kindergarten. Results are based on the calibration dataset (*n* = 73,607). Percentages indicate the proportion of students classified as at risk and the proportion classified as not at risk by the model. Data for the kindergarten assessments are for 2014/15; data for the state reading assessment at the end of grade 3 are for 2017/18.
Source: Authors' analysis based on data from the North Carolina Department of Public Instruction.

Figure A2. Decision tree for classifying North Carolina students as at risk of scoring below proficient on the state reading assessment at the end of grade 3 based on scores on the beginning-of-the-year interim assessments in grade 1, 2015/16 and 2017/18



Note: See table A2 for score ranges for the beginning-of-the-year interim assessments in grade 1. Results are based on the calibration dataset (*n* = 73,607). Percentages indicate the proportion of students classified as at risk and the proportion classified as not at risk by the model. Data for the grade 1 assessments are for 2015/16; data for the state reading assessment at the end of grade 3 are for 2017/18.
Source: Authors' analysis based on data from the North Carolina Department of Public Instruction.

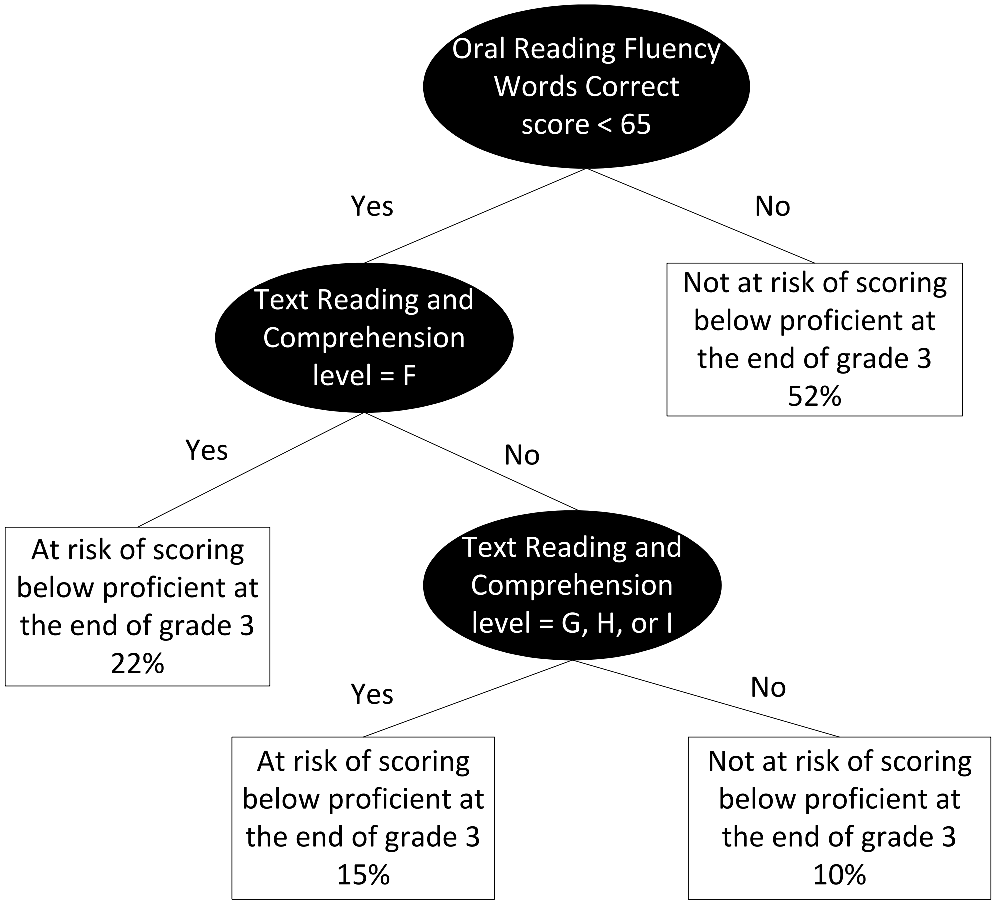
Figure A3. Decision tree for classifying North Carolina students as at risk of scoring below proficient on the state reading assessment at the end of grade 3 based on scores on the middle-of-the-year interim assessments in grade 1, 2015/16 and 2017/18



Note: See table A2 for score ranges for the middle-of-the-year interim assessments in grade 1. Results are based on the calibration dataset (*n* = 73,607). Percentages indicate the proportion of students classified as at risk and the proportion classified as not at risk by the model. Data for the grade 1 assessments are for 2015/16; data for the state reading assessment at the end of grade 3 are for 2017/18.

Source: Authors' analysis based on data from the North Carolina Department of Public Instruction.

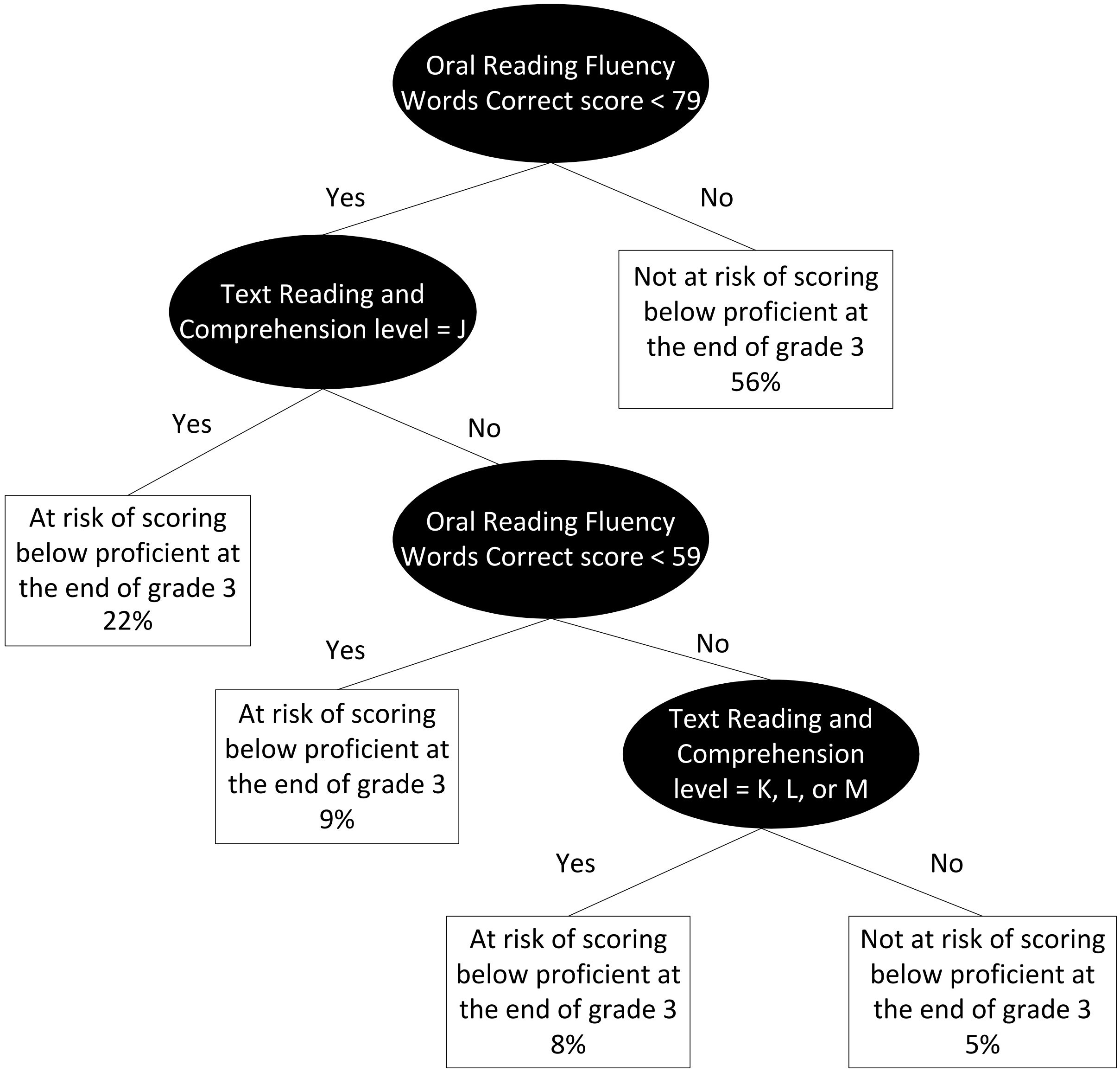
Figure A4. Decision tree for classifying North Carolina students as at risk of scoring below proficient on the state reading assessment at the end of grade 3 based on scores on the beginning-of-the-year interim assessments in grade 2, 2016/17 and 2017/18



Note: See table A2 for score ranges for the beginning-of-the-year interim assessments in grade 2. Results are based on the calibration dataset (*n* = 73,607). Percentages indicate the proportion of students classified as at risk and the proportion classified as not at risk by the model. Data for the grade 2 assessments are for 2016/17; data for the state reading assessment at the end of grade 3 are for 2017/18.

Source: Authors' analysis based on data from the North Carolina Department of Public Instruction.

Figure A5. Decision tree for classifying North Carolina students as at risk of scoring below proficient on the state reading assessment at the end of grade 3 based on scores on the beginning-of-the-year interim assessments in grade 3, 2017/18



Note: See table A2 for score ranges for the beginning-of-the-year interim assessments in grade 3. Results are based on the calibration dataset (*n* = 73,607). Percentages indicate the proportion of students classified as at risk and the proportion classified as not at risk by the model.
Source: Authors' analysis based on data from the North Carolina Department of Public Instruction.

References

Amplify. (2015). *mCLASS Reading 3D: Amplify Atlas book set technical manual, 2nd Edition.* New York, NY: Author.

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

Dynamic Measurement Group. (2018). *Acadience™ Reading: Benchmark goals and composite score.* Eugene, OR: Author. Retrieved September 5, 2019, from https://acadiencelearning.org/papers/AcadienceReadingBenchmarkGoals.pdf.

Dynamic Measurement Group. (2019). *Acadience™ Reading K-6: Technical manual.* Eugene, OR: Author. Retrieved September 5, 2019, from https://acadiencelearning.org/Acadience_Reading_K-6_Technical_Manual.pdf.

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. *International Joint Conference on Artificial Intelligence, 14*(2), 1137–1145.

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. Retrieved from http://ies.ed.gov/ncee/edlabs.

North Carolina Department of Public Instruction. (2014). *Reliability of the North Carolina end-of-grade and end-of-course assessments.* Raleigh, NC: Author. Retrieved September 5, 2019, from http://www.ncpublicschools.org/docs/accountability/testing/eogeocreliabilities14.pdf.

Salford Systems. (n.d.). *An introduction to CART® decision trees.* San Diego, CA: Author. Retrieved March 20, 2020, from http://cdn2.hubspot.net/hub/160602/file-249977783-pdf/docs/JSM.

Steinberg, D. (2013). *Finding R-squared for CART regression trees.* San Diego, CA: Salford Systems. Retrieved February 21, 2014, from https://www.salford-systems.com/blog/dan-steinberg/r-squared-for-cart-regression-trees.

Therneau, T. M., & Atkinson, E. J. (2013). *An introduction to recursive partitioning using the RPART routines* (Mayo Foundation technical report). The Comprehensive R Archive Network. Retrieved December 16, 2013, from http://cran.r-project.org/web/packages/rpart/vignettes/longintro.pdf.