June 2014

Using evidence-based decision trees instead of formulas to identify at-risk readers

Sharon Koon
Yaacov Petscher
Barbara R. Foorman
Florida Center for
Reading Research
at the Florida State University

Key findings

Educators need to understand how students are identified as at risk for reading problems. This study found that the classification and regression tree (CART) model—a type of predictive modeling that presents results in an easy-to-interpret "tree" format—predicted poor performance on the reading comprehension subtest of the Stanford Achievement Test as accurately as the logistic regression model, which is more difficult to interpret. The CART model's ease of communication enables parents, teachers, principals, and school district leaders to better understand how a student is predicted to be at risk.

ies NATIONAL CENTER FOR EDUCATION EVALUATION AND REGIONAL ASSISTANCE
Institute of Education Sciences
U.S. Department of Education

REL SOUTHEAST
Regional Educational Laboratory
At Florida State University

REL 2014–036

The National Center for Education Evaluation and Regional Assistance (NCEE) conducts unbiased large-scale evaluations of education programs and practices supported by federal funds; provides research-based technical assistance to educators and policymakers; and supports the synthesis and the widespread dissemination of the results of research and evaluation throughout the United States.

June 2014

This report was prepared for the Institute of Education Sciences (IES) under Contract ED-IES-12-C-0011 by Regional Educational Laboratory Southeast administered by the Florida Center for Reading Research, Florida State University. The content of the publication does not necessarily reflect the views or policies of IES or the U.S. Department of Education nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government.

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.

This report is available on the Regional Educational Laboratory website at http://ies.ed.gov/ncee/edlabs.

Summary

This study examines whether the classification and regression tree (CART) model improves the early identification of students at risk for reading comprehension difficulties compared with the more difficult to interpret logistic regression model. CART is a type of predictive modeling that relies on nonparametric techniques. It presents results in an easy-to-interpret "tree" format, enabling parents, teachers, principals, and school district leaders to better understand how a student is predicted to be at risk.

Using data from a sample of Florida public school students in grades 1 and 2 in 2012/13, the study found that the CART model predicted poor performance on the reading comprehension subtest of the Stanford Achievement Test as accurately as logistic regression while using fewer or the same number of variables. This research is motivated by state education leaders' interest in maintaining high classification accuracy while simultaneously improving practitioner understanding of the rules used to identify students as at-risk or not at-risk readers.

Contents

Why this study?

Since 2002/03 Florida school districts have been required to administer an interim reading assessment aligned to state standards for grade-level reading performance. The Florida Assessments for Instruction in Reading (FAIR) 1.0 was developed to meet this requirement and is administered by most school districts to monitor student progress and predict student performance on the end-of-year summative assessment—the Stanford Achievement Test Series, Tenth Edition (SAT-10), in grades 1 and 2 and the Florida Comprehensive Assessment Test (FCAT) 2.0 in grades 3–10. In 2012/13 FAIR 1.0 was administered to more than 250,000 students in grades 1 and 2. In recent years the Florida Department of Education has received numerous queries from school officials on how FAIR 1.0 identifies students as at risk or not at risk. The assessment's probability-of-success formula, derived from a logistic regression, is cumbersome to explain.

The classification and regression tree model yields a classification flowchart that clearly shows how a student may be identified as at risk or not at risk for reading comprehension difficulties

The Florida Center for Reading Research has developed a new version of FAIR 1.0 to be aligned to the state's new content standards for 2014/15. This new version, the Florida Assessments for Instruction in Reading–Florida Standards (FAIR–FS), will be available for license to the Florida Department of Education to use as the reading component skill battery.

Members of the Regional Educational Laboratory Southeast (REL-SE) Improving Literacy Alliance requested a preliminary analysis of the FAIR–FS's classification accuracy in grades 1 and 2 using an approach that would help practitioners understand how students are identified as at risk.

A classification and regression tree (CART) model may meet practitioners' needs. The CART model identifies students as at risk or not at risk based on their scores on predictor variables (for example, FAIR–FS scores). The CART model searches for the optimal split on the predictor variables and partitions the sample into binary subsamples called nodes. In a visual representation of the model, the nodes form either a rectangular box or an oval (figure 1). Boxes are referred to as terminal nodes, which do not split further, and ovals are referred to as nonterminal nodes, which split again when there is another meaningful difference between students on the predictor variables.

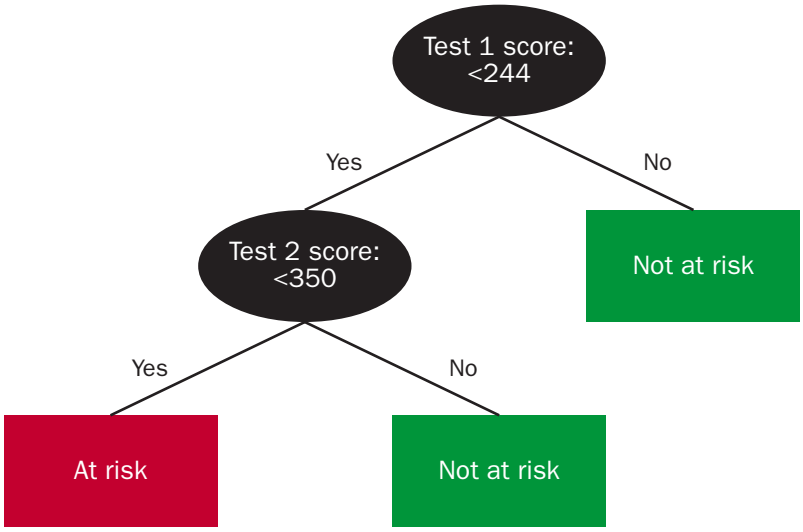
The CART model yields a classification flowchart that clearly shows how a student may be identified as at risk or not at risk for reading comprehension difficulties. For example, in figure 1, students who score below 244 on test 1 and below 350 on test 2 (with score ranges of 100–1,000) are identified as at risk. Decision trees with a limited number of splits lend themselves to easier interpretation than those with many splits.

Logistic regression, the traditional approach to early identification, relies on empirically estimated coefficients, Euler's constant (the base of the natural log [*e*], equal to 2.718), and the transformation of log odds to a predictive probability. Results are a by-product of the following type of equation:

$$ln\left[\frac{p}{1-p}\right] = \beta_0 + \beta_1(X_1) + \beta_2(X_2).$$

This complicated process can be difficult for parents, teachers, principals, and school district leaders to understand.

Figure 1. Sample classification and regression tree flowchart



Note: Scores range from 100 to 1,000.

Source: Authors' creation.

Logistic regression results can be used to generate more straightforward contingency tables. A full table based on two tests with score ranges of 100–1,000 would consist of one column for each possible test 1 score and one row for each possible test 2 score. A snapshot of the table can capture the scores at which a student transitions to at-risk status. For example, in table 1, as in figure 1, students with test 1 scores below 244 and test 2 scores below 350 are identified as at risk (through red shading), while other students are identified as not at risk (through green shading).

When presenting scores from only one or two subtests, logistic regression contingency tables are easy to interpret. But with more subtests and score ranges, contingency tables become more complex and difficult to explain, not lending themselves to simple snapshots.

The CART model thus offers advantages over logistic regression in statistical parsimony (fewer predictors in the classification model) and ease of communication. But does it classify students as accurately?

Table 1. Contingency table snapshot

	Test 1 score				
Test 2 score	**242**	**243**	**244**	**245**	**246**
348	**0.15**	**0.15**	**0.30**	**0.60**	**0.80**
349	**0.15**	**0.15**	**0.30**	**0.60**	**0.80**
350	**0.30**	**0.30**	**0.30**	**0.60**	**0.80**
351	**0.60**	**0.60**	**0.60**	**0.60**	**0.80**
352	**0.80**	**0.80**	**0.80**	**0.80**	**0.80**

Note: Scores range from 100 to 1,000.

Source: Authors' creation.

Several traditional indexes of classification accuracy can be used to evaluate results from logistic regression and CART models (Schatschneider, Petscher, & Williams, 2008). These indexes are derived from a 2x2 contingency table that provides counts of students in four categories resulting from student performance on a screening assessment and an outcome assessment (table 2).

The first index, sensitivity, is the proportion of students who are identified as at risk on the screening assessment among all students who fail the outcome assessment—or the number of true positives divided by the sum of the true positives and false negatives (A/[A+C]). The second index, specificity, is the proportion of students who are identified as not at risk among all students who pass the outcome assessment—or the number of true negatives divided by the sum of true negatives and false positives (D/[D+B]). The third index, positive predictive power, is the proportion of students who fail the outcome assessment among all students who are identified as at risk on the screening assessment—or the number true positives divided by the sum of true positives and false positives (A/[A+B]). The fourth index, negative predictive power, is the proportion of students who pass the outcome assessment among all students who are identified as not at risk on the screening assessment—or the number of true negatives divided by the sum of false negatives and true negatives (D/[C+D]).

The REL-SE Improving Literacy Alliance is most interested in maximizing negative predictive power to minimize false negatives so that at-risk students can receive timely interventions

The REL-SE Improving Literacy Alliance is most interested in maximizing negative predictive power (Petscher, Kim, & Foorman, 2011). The goal of this strategy is to minimize false negatives (that is, not underidentifying students) so that at-risk students can receive timely interventions. Consistent with FAIR 1.0 (Florida Department of Education, 2009), a negative predictive power of .85 is the expected minimum standard for the FAIR–FS (that is, no more than 15 percent of students are underidentified).

If the CART model demonstrates equal or better classification accuracy than logistic regression (the current method in Florida) does, the REL-SE Improving Literacy Alliance may wish to advocate using the CART model in establishing new classification rules for use with the FAIR–FS.

Table 2. Sample 2 × 2 contingency table

	Outcome assessment	
Screening assessment	**Fail**	**Pass**
At risk	A: True positive	B: False positive
Not at risk	C: False negative	D: True negative

Source: Authors' illustration.

What the study examined

This study used data from a sample of students in grades 1 and 2 in Florida public schools during the 2012/13 academic year to answer the following research question: How do CART analyses compare with logistic regression methods in predicting poor performance on the reading comprehension subtest of the SAT-10?

Classification accuracy in each grade was based on the accuracy of FAIR–FS tasks in predicting end-of-year reading scores on the SAT-10. (See appendix A for detailed information on the study's data and methods.)

The FAIR–FS tasks in grades K–2 were developed in accordance with research consensus that reading success in the primary grades is predicted by print knowledge (knowledge of letter names and sounds, phonological awareness, word reading, and spelling) and language skills (syntax, vocabulary, and listening comprehension; National Early Literacy Panel, 2008; National Institute of Child Health and Human Development, 2000; Snow, Burns, & Griffin, 1998; Rayner, Foorman, Perfetti, Pesetsky, & Seidenberg, 2001). Measurement of these predictors and how they change over time is essential to accurately identifying students at risk for reading difficulties.

Measurement of print knowledge and language skills and how they change over time is essential to accurately identifying students at risk for reading difficulties

Multivariate screening approaches are preferred to single tests or measures (Fletcher et al., 2002; Francis et al., 2005) because they can assess skills at different points in reading development. The FAIR–FS thus consists of four alphabetic and oral language tasks designed to be administered at different grade levels. The tasks included in each analysis in this study varied by grade as follows:

- Alphabetics
  - *Word reading* (grades 1 and 2): The student is required to pronounce each word displayed on the screen.
  - *Word building* (grade 1): The student sees a word on the screen and is asked to manipulate individual letters at the bottom of the screen to change the word into a different word. For example, "this is the word 'hop'; make the word 'hot'" or "this is the word 'hop'; make the word 'hope.'"
  - *Spelling* (grade 2): The student types a word pronounced by the computer.
- Oral language
  - *Vocabulary pairs* (grades 1 and 2): Three words are displayed on the screen and read aloud and the student is required to identify the two words that go together (for example, "dark," "night," "swim").
  - *Following directions* (grades 1 and 2): The student is required to listen to single and multistep directions from the computer. The student then must respond to the directions by clicking on or moving the specified objects on the computer screen (for example, put the square in front of the chair and then put the circle behind the chair).

Performance on each FAIR–FS task is reported using a developmental scale with scores ranging from 200 to 800, a mean of 500, and a standard deviation of 100.

Findings

The CART model performed comparably to logistic regression in using FAIR–FS tasks to predict risk of poor performance on the SAT-10. CART results were consistent with those from logistic regression on all measures of classification accuracy while using fewer or the same number of variables (table 3).

Researchers have proposed different threshold values for sensitivity and specificity; many look for levels of at least .80, and some recommend at least .90 (Compton, Fuchs, Fuchs, & Bryant, 2006; Jenkins, 2003). Jenkins (2003) suggested that 5–10 percent represents an acceptable level of false negatives (that is, a negative predictive power of .90–.95) and that 90–95 percent represents an acceptable level of true positives (that is, sensitivity of .90–.95). All final models in this study yielded negative predictive power meeting or exceeding the .90 standard—false negative rates ranged from 4 to 8 percent, with minimal differences between methods within each grade. Sensitivity fell below the recommended standard, except for the grade 1 CART model. However, as discussed earlier, this study emphasized maximizing negative predictive power. Specificity was at or near .90. Positive predictive power was much lower for all models, also reflecting the emphasis on negative predictive power.

In grade 1 the classification and regression tree results were better than the logistic regression results on all indexes of classification accuracy

Grade 1

In grade 1 the CART results were better than the logistic regression results on all indexes of classification accuracy. Both methods resulted in models that retained three of the four available tasks, but each model used a different combination of three: the CART model retained word reading, vocabulary pairs, and following directions, and the logistic regression model retained word reading, vocabulary pairs, and word building. A variable may appear in the CART model many times (Therneau & Atkinson, 2013) because the search for the single variable that will result in the best subsequent split to the data includes all variables at each split.

Classification and regression tree model. Based on the CART results, students would be identified as at risk under either of the following conditions (figure 2):

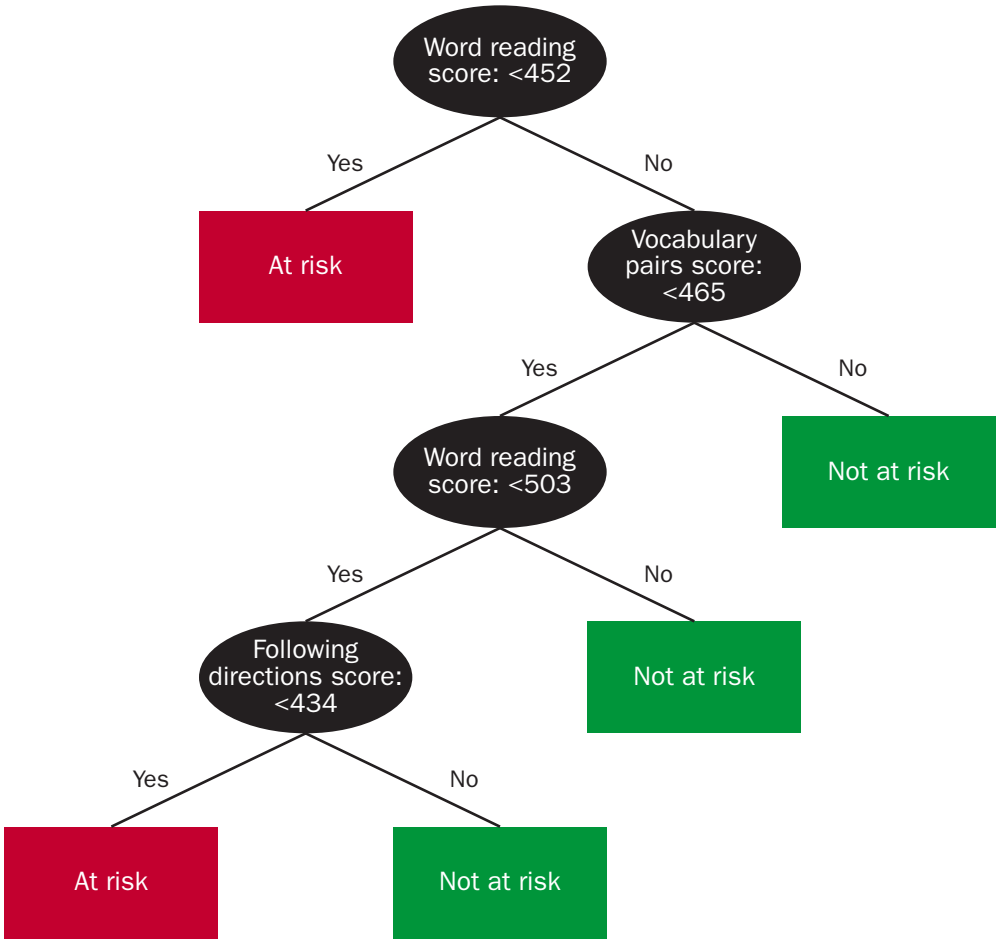
- The student achieved a word reading score below 452.
- The student achieved a vocabulary pairs score below 465, a word reading score of 452–502, and a following directions score below 434.

Table 3. Summary of results by model

Grade and model	Sensitivity	Specificity	Positive predictive power	Negative predictive power	Overall proportion correct
Grade 1 (*n* = 206)					
Classification and regression tree	.92	.90	.79	.96	.90
Logistic regression	.87	.90	.78	.94	.89
Grade 2 (*n* = 181)					
Classification and regression tree model 1	.70	.89	.74	.87	.83
Classification and regression tree model 2	.82	.86	.73	.92	.85
Logistic regression	.84	.88	.76	.92	.87

Source: Authors' analysis of data from the Florida Center for Reading Research.

Figure 2. Grade 1 classification rules under the classification and regression tree model



Note: The decision trees presented in this report are simplified and therefore do not provide the number of students correctly classified within each terminal node. Decision trees that include this additional classification information are available from the authors upon request. Classification tables for each model are provided in appendix A. Scores range from 200 to 800.

Source: Authors' analysis of data from the Florida Center for Reading Research.

Alternatively, students would be identified as not at risk under any of the following conditions:

- The student achieved a word reading score of 452 or above and a vocabulary pairs score of 465 or above.
- The student achieved a word reading score of 503 or above and a vocabulary pairs score below 465.
- The student achieved a word reading score of 452–502, a vocabulary pairs score below 465, and a following directions score of 434 or above.

The negative predictive power of .96 indicates that of the 135 students predicted to be not at risk after applying these decision rules, the CART model correctly identifies 130 (table A3 in appendix A). The remaining five students represent the model's false negatives, which would be found in one of the green-shaded rectangles. Of the 71 students predicted to be at risk, 56 were correctly identified, reflecting the model's positive predictive power

of .79. The remaining 15 students represent the model's false positives, and they would be found in the red-shaded rectangles.

Logistic regression model. The results of the logistic regression model are often best represented by an equation, such as the following from the grade 1 analysis:

$$\text{Logit} = -21.749 + .032*\text{word reading score} + .009*\text{vocabulary pairs score} + .006*\text{word building score}.$$

Because the three independent variables in the logistic regression were on the same scale, the variable weights can be directly compared without using standardized estimates. Consistent with the CART model, the word reading task contributed the most to predicting a student's classification, followed by vocabulary pairs.

The estimated logistic regression model would be used to calculate predicted SAT-10 logit scores for each future student, which could then be transformed to predicted probabilities. Probabilities below .50 would be identified as at risk for scoring below the 40th percentile on the SAT-10 assessment.[1]

In grade 2 the logistic regression and classification and regression tree results were comparable, after the addition of a loss matrix in which false negatives were treated as two times the cost of false positives to the CART model specifications

Grade 2

In grade 2 the logistic regression and CART results were comparable, after the addition of a loss matrix in which false negatives were treated as two times the cost of false positives to the CART model specifications (model 2; see appendix A). Although the logistic regression results were better, CART model 2 may be more parsimonious, with only three predictors retained instead of the four in the logistic regression model. Consistent with the grade 1 results, the word reading task contributed the most in both methods to predicting a student's classification.

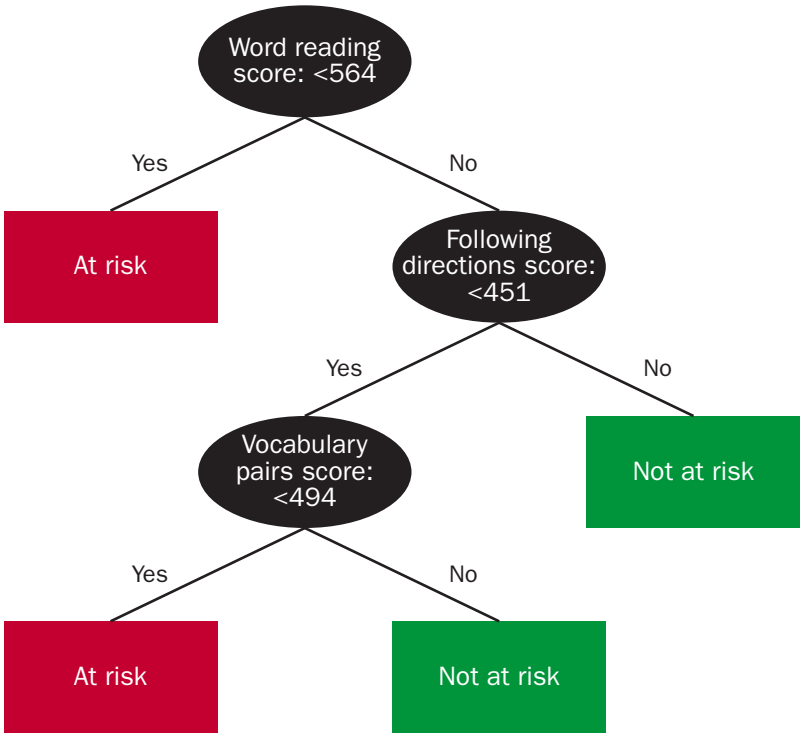
Classification and regression tree model. Based on the CART results, students would be identified as at risk under either of the following conditions (figure 3):

- The student achieved a word reading score below 564.
- The student achieved a word reading score of 564 or above, a following directions score below 451, and a vocabulary pairs score below 494.

Alternatively, students would be identified as not at risk under either of the following conditions:

- The student achieved a word reading score of 564 or above and a following directions score of 451 or above.
- The student achieved a word reading score of 564 or above, a following directions score below 451, and a vocabulary pairs score of 494 or above.

Figure 3. Grade 2 classification rules under classification and regression tree model 2



Note: The decision trees presented in this report are simplified and therefore do not provide the number of students correctly classified within each terminal node. Decision trees that include this additional classification information are available from the authors upon request. Classification tables for each model are provided in appendix A. Scores range from 200 to 800.

Source: Authors' analysis of data from the Florida Center for Reading Research.

Logistic regression model. As in grade 1, the prediction equation resulting from the logistic regression model,

$$\text{Logit} = -21.984 + .017*\text{word reading score} + .011*\text{spelling score} + .007*\text{vocabulary pairs score} + .007*\text{following directions score},$$

would be used to calculate predicted SAT-10 logit scores for each future student, which could then be transformed to probabilities. Probabilities below .50 would be identified as at risk for scoring below the 40th percentile on the SAT-10.

Implications of the findings

Based on the study results, the CART model can be recommended for use in grades 1 and 2 with the new FAIR–FS in Florida for several reasons. The CART results are comparable to those of logistic regression—both yield negative predictive power above the recommended standard of .90. But CART results are easier to communicate and use. Practitioners can identify a student as at risk or not at risk using the decision tree and know which assessment—and, therefore, which component skill placed the student in an at-risk category—without complicated mathematical operations. In addition, computer applications using decision rules instead of equations are often much easier to implement.

The CART model also holds several technical advantages over logistic regression. First, as a nonparametric method, the CART model is not sensitive to the presence of outliers, unlike logistic regression. Second, the CART model is not sensitive to collinearity between the variables. Third, it models complex interactions among predictors that may be difficult or impossible to estimate in the regression framework. A disadvantage of the CART model, however, is that it is sensitive to missing data, so either listwise deletion or data imputation are required to estimate the model.

Study limitations

This study has several limitations. First, CART results can be greatly affected by small changes in the independent variables (for example, the screening assessments), so any such changes require an updated analysis.

Second, CART decision trees can be generated that correctly identify all students, but the classification rules for these decision trees are far more complex. This study sought a parsimonious, technically adequate method.

Third, sensitivity or specificity could be improved in logistic regression by adjusting the cutscore in group classifications. Specifications could be adjusted in both models. This study used specifications designed to meet or exceed negative predictive power of .85 while maintaining acceptable sensitivity and specificity levels.

The classification and regression tree model is not sensitive to the presence of outliers, it is not sensitive to collinearity between the variables, and it models complex interactions among predictors that may be difficult or impossible to estimate in the regression framework

Appendix A. Data and methodology

This appendix provides detailed information on the study's data sources and methodology.

Data

Participant data were obtained from an archive containing FAIR–FS data on 4,500 students in grades K–2 in 28 elementary schools in the Escambia County and Hillsborough County school districts in Florida. The archive is the result of data obtained from a linking study conducted from December 2012 to May 2013 as part of Florida State University's subcontract from the Educational Testing Service's assessment grant in Institute of Education Sciences/National Center for Education Research's Reading for Understanding initiative (Sabatini, PI; R305F100005). FAIR–FS was administered in December 2012–January 2013, and the SAT-10 was administered in April–May 2013. Florida State University's subcontract with the Educational Testing Service makes it clear that Florida State University owns the FAIR–FS and all data produced under the subcontract. These analyses are not part of the Reading for Understanding subcontract.

As part of its current testing practices, Hillsborough County administers the SAT-10 to all students in grades 1 and 2, and Hillsborough agreed to provide the SAT-10 scores for study participants. There was thus no need to administer the SAT-10 in grades 1 or 2. Therefore, this study used only data from the Hillsborough County school district. About 2,000 students in grades 1 and 2 in Hillsborough County, representing 15 schools, took the FAIR–FS between December 3 and January 11 and the SAT-10 between April 2 and April 12.

Outliers and missing data

Grade 1. The initial grade 1 dataset included 1,028 students. Twenty-seven cases were deleted due to missing SAT-10 scores, and one case was deleted due to missing data on all FAIR–FS tasks. An analysis of univariate and multivariate outliers, a requirement of logistic regression, resulted in the deletion of an additional 14 cases. The final dataset included 986 students.

Missing data were not missing completely at random, based on a significant Little's missing completely at random test (p = .000), but a review of the data indicated that the nature of missingness meant that randomness could be assumed. Table A1 summarizes the univariate missing data statistics. To address the missing data, multiple imputation with SAS 9.4 software was used to create a dataset with complete cases for all variables. Logistic regression can analyze and summarize multiply imputed datasets, but there is no accepted procedure for analyzing and summarizing classification trees generated from multiple imputed files. Therefore, a decision was made to conduct 20,000 imputations and then use the mean imputed value for each missing value.

Grade 2. The initial grade 2 dataset included 918 students. Fifteen cases were deleted due to missing SAT-10 scores. An analysis of univariate and multivariate outliers resulted in the deletion of an additional 16 cases. The final dataset included 887 students. The analysis of missing data revealed that missing data were not missing completely at random, based on a significant Little's missing completely at random test (p = .003). Table A2 summarizes the univariate missing data statistics. As in grade 1, a dataset with complete cases for all variables was created by aggregating the results of 20,000 imputations.

Table A1. Grade 1 missing data statistics (*n* = 986)

FAIR–FS task	Total complete cases	Mean	Standard deviation	Missing	
				Count	Percent
Word reading	959	516.06	105.14	27	2.7
Word building	911	504.86	98.26	75	7.6
Vocabulary pairs	888	508.98	109.94	98	9.9
Following directions	967	502.18	111.21	19	1.9

FAIR-FS is Florida Assessments for Instruction in Reading–Florida Standards.

Source: Authors' analysis of data from the Florida Center for Reading Research.

Table A2. Grade 2 missing data statistics (*n* = 887)

FAIR–FS task	Total complete cases	Mean	Standard deviation	Missing	
				Count	Percent
Word reading	866	614.28	112.95	21	2.4
Spelling	853	501.26	102.57	34	3.8
Vocabulary pairs	847	562.02	114.38	40	4.5
Following directions	866	505.46	112.09	21	2.4

FAIR-FS is Florida Assessments for Instruction in Reading–Florida Standards.

Source: Authors' analysis of data from the Florida Center for Reading Research.

Analysis methods

The developmental ability scores from each FAIR–FS task were used in a series of CART and logistic regression analyses in predicting end-of-year performance on the SAT-10. Traditional indexes of classification accuracy were used to assess differences in the results between the approaches.

Prior to conducting the analyses, the grade-based correlations among the FAIR–FS task scores from the individual literacy components were examined for multicollinearity in each of the imputed files. None of the Pearson correlations was higher than .80, eliminating concerns of redundancy in the subsequent logistic regression analyses. Following this step, the SAT-10 scores were dummy-coded to represent proficiency level. Percentile scores on the SAT-10 were dichotomized so that scores at or above the 40th percentile were coded as 1 for "not at risk" and scores below the 40th percentile were coded as 0 for "at risk." A previous report by the American Institutes for Research (2007) demonstrated that the 40th percentile represents a reasonable grade-based target for proficiency in grades K–2.

The final datasets for each grade were then split into a calibration dataset, consisting of a random sample of 80 percent of the students in each grade, and a validation dataset, consisting of the remaining 20 percent. Both the CART and logistic regression analyses were based on the same datasets, with the models built on the calibration dataset and tested on the validation dataset.

The two methods were evaluated using traditional measures of diagnostic accuracy, including sensitivity (proportion of true positives), specificity (proportion of true negatives), positive and negative predictive power, overall correct classification, and the receiver operating characteristic area under the curve.

CART analyses were run using the R 2.15.3 package rpart, and logistic regression analyses were run using SPSS Statistics 21. SPSS Statistics 21 was used to calculate both the CART and logistic regression receiver operating characteristic area under the curve estimates.

Classification and regression tree model. The CART model classifies individuals into mutually exclusive subgroups of a population using a nonparametric approach that results in a classification tree (see figure 1 in the main report). The subgroup splits in the CART model are determined by the software program (the R rpart package) to improve the overall predictive accuracy. The CART model uses an exhaustive subgroup comparison to identify the best predictors and predictor levels that most efficiently split the learning sample into the most homogeneous subgroups of students who are identified as at risk or not at risk based on their observed scores.

In this study, the CART model assessed the individual performance of each FAIR–FS task at every available cutscore to classify students into at-risk and not-at-risk categories. To ensure a parsimonious model, several specifications were used to limit the number of splits. Guided by Compton et al. (2006), the analysis specified a stopping rule of a minimal parent-node size of three students. The number of splits was also limited by specifying a minimum reduction in the relative error (approximately equivalent to 1–*R*-squared), identified after running a base model with no minimum specified. Each grade-based model included tenfold cross-validation to evaluate the quality of the prediction tree and determine the appropriate minimum complexity parameter (Breiman, Friedman, Olshen, & Stone, 1984). A recommended minimum standard is the value of the complexity parameter that results in a cross-validation relative error less than one standard error above the minimum cross-validation relative error (Therneau, Atkinson, & Ripley, 2013).

In using the CART model, the intention was to build and prune trees that would maximize the negative predictive power of .85. To accomplish this, researchers revised the model to specify a loss matrix, through which the program weights classification errors differently. To raise the negative predictive power, the specification would be to view false negatives as more costly.

Logistic regression. Binary logistic regression is an extension of simple or multiple regression, whereby a dichotomously scored dependent variable is regressed on one or more selected independent variables. This technique is widely used not only to predict log odds of success on the dependent variables, but also to study the rates of true and false positives and negatives in classifying individuals as at risk or not at risk. The logistic regression models in this study were developed hierarchically. Based on the correlations between the individual FAIR–FS tasks and the dichotomized SAT-10 variable, the FAIR–FS task scores were entered into the logistic regression ordered by correlational magnitude. FAIR–FS tasks that added at least 2 percent unique variance above the task already in the model, as measured by the Nagelkerke pseudo *R*-squared, were retained for the final classification model from the logistic regression. Cohen (1992) has shown that an *R*-squared between 2 and 14 percent represents a small, practically important contribution to explained variance. This same standard was applied to the increase in Nagelkerke pseudo R-squared, which is estimated by means of maximum likelihood in logistic regression and can be interpreted in the same way as the *R*-squared estimated in an ordinary least squares regression.

Grade 1

Classification and regression tree model

Model building. All four FAIR–FS tasks were specified in a base model using the calibration dataset. Ten cross-validations were specified along with a minimum of three cases required to add another split. A complexity parameter and a cost matrix were not specified, so that the number of splits would not be limited and both types of classification errors would be treated the same. Based on the cross-validation results from the base model, the classification tree was pruned by specifying a complexity parameter of .02, and resulted in a pseudo *R*-squared of .64 (see figure 1 in main report).

Model testing. The classification rules were applied to the validation dataset to predict group membership as well as probabilities associated with membership in each group. These results were used to generate a classification table (table A3).

The area under the curve was estimated to be .94. However, these results were noted by SPSS as systematically underestimated because cases in each observed group had the same value on the predictor variables.

Logistic regression

Model building. Based on the correlations between the individual FAIR–FS tasks and performance on the SAT-10, the FAIR–FS task scores were entered into the logistic regression ordered by correlational magnitude as follows: word reading (r = .64), word building (r = .53), vocabulary pairs (r = .52), and following directions (r = .41). The results based on the calibration dataset are provided in table A4.

Due to the minimal increase in the explained variance based on the Nagelkerke pseudo *R*-squared (1.7 percent), the following directions task was deleted from the model. The final model coefficients are provided in table A5, with a Nagelkerke pseudo *R*-squared of 72.1 percent.

Table A3. Grade 1 classification and regression tree classification table (*n* = 206)

			SAT-10 score: observed		
			0 (at risk)	1 (not at risk)	Total
SAT-10 score: predicted	0 (at risk)	Count	56	15	71
		Percent within predicted	78.9	21.1	100.0
		Percent within observed	91.8	10.3	34.5
		Percent of total	27.2	7.3	34.5
	1 (not at risk)	Count	5	130	135
		Percent within predicted	3.7	96.3	100.0
		Percent within observed	8.2	89.7	65.5
		Percent of total	2.4	63.1	65.5
Total		Count	61	145	206
		Percent of total	29.6	70.4	100.0

SAT-10 is the Stanford Achievement Test Series, Tenth Edition.

Source: Authors' analysis of data from the Florida Center for Reading Research.

Table A4. Grade 1 logistic regression model evaluation (*n* = 780)

Block	Variable	Hosmer and Lemeshow test *p* value	Nagelkerke pseudo *R*-squared	Change in Nagelkerke pseudo *R*-squared	Overall percentage correct
0	Constant	—	—	—	70.1
1	Word reading	.492	.659	—	86.5
2	Word building	.670	.682	.023	87.1
3	Vocabulary pairs	.913	.721	.039	88.6
4	Following directions	.716	.738	.017	88.5

— is not applicable.

Source: Authors' analysis of data from the Florida Center for Reading Research.

Table A5. Grade 1 logistic regression final model (*n* = 780)

Variable	Coefficient (B)	Standard error	Wald statistic	Degrees of freedom	Significance level	Exp(B)	95 percent confidence interval for Exp(B)	
							Lower	Upper
Word reading	.032	.003	102.51	1	.000	1.033	1.026	1.039
Word building	.006	.002	11.00	1	.001	1.006	1.002	1.009
Vocabulary pairs	.009	.002	36.30	1	.000	1.009	1.006	1.013
Constant	–21.749	1.820	142.87	1	.000	.000		

Source: Authors' analysis of data from the Florida Center for Reading Research.

Model testing. The model coefficients from the final model were used in the prediction equation

Logit = –21.749 + .032*word reading score + .006*word building score + .009*vocabulary pairs score

to calculate predicted SAT-10 logit scores for each case in the validation dataset, which were then transformed to probabilities. Probabilities of .5 and above were recoded as 1 (scoring at or above the 40th percentile on the SAT-10), and values below .5 were coded as 0. The results were used to generate a classification table for use in calculating indices of classification accuracy (table A6).

Using these same data, the area under the curve was estimated, using the default cutpoint of .5, and found to be .95, with a standard error of .007.

Grade 2

Classification and regression tree model

Model building. All four FAIR–FS tasks were specified in a base model using the calibration dataset. Ten cross-validations were specified along with a minimum of three cases required to add another split. As noted earlier in the grade 1 base model, a complexity parameter and a cost matrix were not specified, so that the number of splits would not be limited and both types of classification errors would be treated the same. Based on the cross-validation results from the base model, the classification tree was pruned by specifying a complexity parameter of .016 (figure A1).

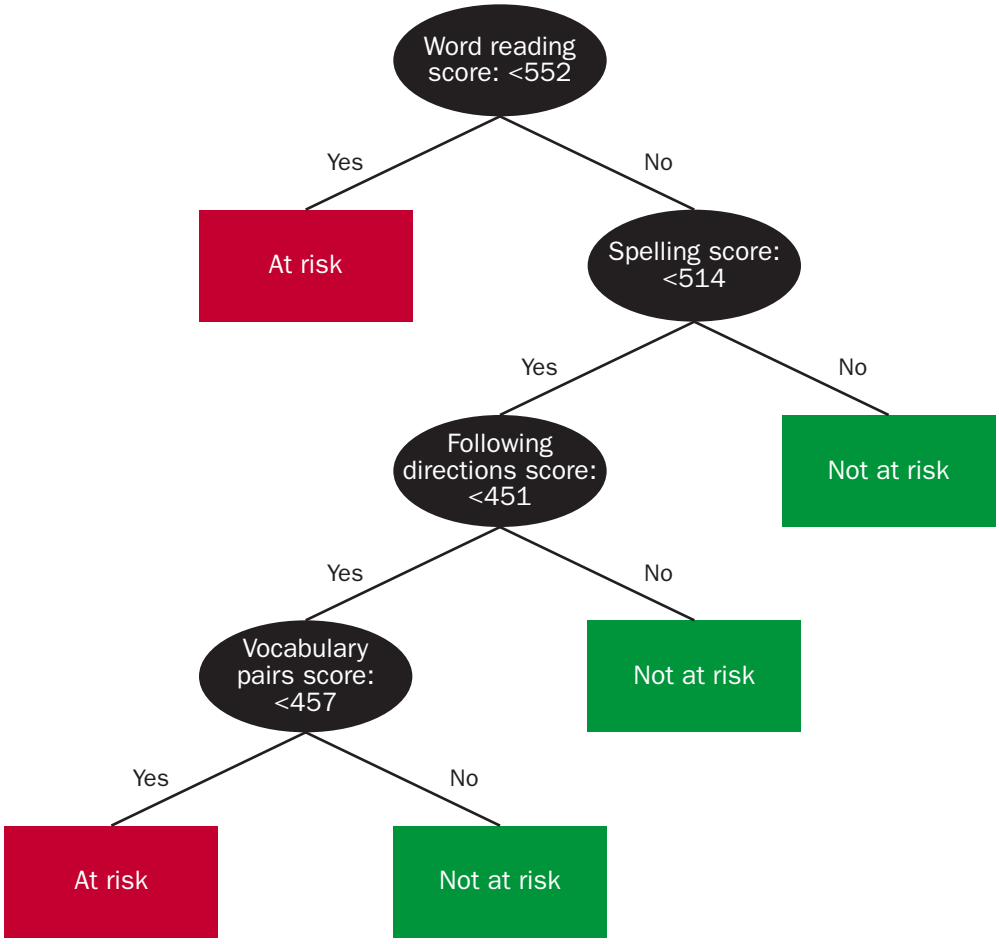
Table A6. Grade 1 logistic regression classification table (*n* = 206)

			SAT-10 score: observed		
			0 (at risk)	1 (not at risk)	Total
SAT-10 score: predicted	0 (at risk)	Count	53	15	68
		Percent within predicted	77.9	22.1	100.0
		Percent within observed	86.9	10.3	33.0
		Percent of total	25.7	7.3	33.0
	1 (not at risk)	Count	8	130	138
		Percent within predicted	5.8	94.2	100.0
		Percent within observed	13.1	89.7	67.0
		Percent of total	3.9	63.1	67.0
Total		Count	61	145	206
		Percent of total	29.6	70.4	100.0

SAT-10 is the Stanford Achievement Test Series, Tenth Edition.

Source: Authors' analysis of data from the Florida Center for Reading Research.

Figure A1. Grade 2 Classification and regression tree model decision rules (model 1)



Note: Scores range from 200 to 800.

Source: Authors' analysis of data from the Florida Center for Reading Research.

Model testing. The classification rules were applied to the validation dataset to predict group membership as well as probabilities associated with membership in each group. These results were used to generate a classification table (table A7).

The area under the curve was estimated to be .853, with a standard error of .031. As noted earlier, these results are somewhat biased.

Because the negative predictive power was only slightly higher than the standard of .85, the calibration model was revised to specify the addition of a loss matrix, where the cost of false negatives would be treated as two times the cost of false positives. The pruned tree resulting from this model is shown in figure 3 of the main report. The pseudo *R*-squared for this model was 0.713.

The classification table that resulted from applying the model to the validation dataset is provided in table A8.

Table A7. Grade 2 classification and regression tree classification table, model 1 (*n* = 181)

			SAT-10 score: observed		
			0 (at risk)	1 (not at risk)	Total
SAT-10 score: predicted	0 (at risk)	Count	39	14	53
		Percent within predicted	73.6	26.4	100.0
		Percent within observed	69.6	11.2	29.3
		Percent of total	21.5	7.7	29.3
	1 (not at risk)	Count	17	111	128
		Percent within predicted	13.3	86.7	100.0
		Percent within observed	30.4	88.8	70.7
		Percent of total	9.4	61.3	70.7
Total		Count	56	125	181
		Percent of total	30.9	69.1	100.0

SAT-10 is the Stanford Achievement Test Series, Tenth Edition.

Source: Authors' analysis of data from the Florida Center for Reading Research.

Table A8. Grade 2 classification and regression tree classification table, model 2 (*n* = 181)

			SAT-10 score: observed		
			0 (at risk)	1 (not at risk)	Total
SAT-10 score: predicted	0 (at risk)	Count	46	17	63
		Percent within predicted	73.0	27.0	100.0
		Percent within observed	82.1	13.6	34.8
		Percent of total	25.4	9.4	34.8
	1 (not at risk)	Count	10	108	118
		Percent within predicted	8.5	91.5	100.0
		Percent within observed	17.9	86.4	65.2
		Percent of total	5.5	59.7	65.2
Total		Count	56	125	181
		Percent of total	30.9	69.1	100.0

SAT-10 is the Stanford Achievement Test Series, Tenth Edition.

Source: Authors' analysis of data from the Florida Center for Reading Research.

Logistic regression

Model building. Based on the correlations between the individual FAIR–FS tasks and performance on the SAT-10, the FAIR–FS task scores were entered into the logistic regression ordered by correlational magnitude as follows: word reading (r = .62), spelling (r = .60), vocabulary pairs (r = .48), and following directions (r = .40). The results based on the calibration dataset are provided in table A9.

All FAIR–FS tasks contributed to explaining a significant and practically important percentage of variance and were kept in the final model. Approximately 70 percent of the variance in the logit of SAT-10 scores was explained by the FAIR–FS tasks, as indicated by the Nagelkerke pseudo *R*-squared of .693. The final model coefficients are provided in table A10.

Model testing. The model coefficients from the final model were used in the prediction equation

Logit = –21.984 + .017*word reading score +.011*spelling score +
.007*vocabulary pairs score + .007*following directions score

to calculate predicted SAT-10 logit scores for each case in the validation dataset, which were then transformed to probabilities. Probabilities of .5 and above were recoded as 1

Table A9. Grade 2 logistic regression model evaluation (*n* = 706)

Block	Variable	Hosmer and Lemeshow test *p* value	Nagelkerke pseudo *R*-squared	Change in Nagelkerke pseudo *R*-squared	Overall percentage correct
0	Constant	—	—	—	66.7
1	Word reading	.033	.580	—	84.1
2	Spelling	.271	.607	.027	83.7
3	Vocabulary pairs	.260	.664	.057	85.8
4	Following directions	.502	.693	.029	85.4

— is not applicable.

Source: Authors' analysis of data from the Florida Center for Reading Research.

Table A10. Grade 2 logistic regression final model (*n* = 706)

Variable	Coefficient (B)	Standard error	Wald statistic	Degrees of freedom	Significance level	Exp(B)	95 percent confidence interval for Exp(B)	
							Lower	Upper
Word reading	.017	.003	44.32	1	.000	1.017	1.012	1.022
Spelling	.011	.002	21.92	1	.000	1.011	1.006	1.016
Vocabulary pairs	.007	.001	26.29	1	.000	1.007	1.005	1.010
Following directions	.007	.001	25.10	1	.000	1.007	1.005	1.010
Constant	–21.984	1.839	142.89	1	.000	.000		

Source: Authors' analysis of data from the Florida Center for Reading Research.

(scoring at or above the 40th percentile on the SAT-10), and values below .5 were coded as 0. The results were used to generate a classification table for use in calculating indices of classification accuracy (table A11).

Using these same data, the area under the curve was estimated, using the default cutpoint of .5, and was found to be .96, with a standard error of .013.

Table A11. Grade 2 logistic regression classification table (*n* = 181)

			SAT-10 score: observed		
			0 (at risk)	1 (not at risk)	Total
SAT-10 score: predicted	0 (at risk)	Count	47	15	62
		Percent within predicted	75.8	24.2	100.0
		Percent within observed	83.9	12.0	34.3
		Percent of total	26.0	8.3	34.3
	1 (not at risk)	Count	9	110	119
		Percent within predicted	7.6	92.4	100.0
		Percent within observed	16.1	88.0	65.7
		Percent of total	5.0	60.8	65.7
Total		Count	56	125	181
		Percent of total	30.9	69.1	100.0

Source: Authors' analysis of data from the Florida Center for Reading Research.

Note

1. The default cutscore of .50 was used in the logistic regression analyses to evaluate classification accuracy. Another cutscore, such as .70, which represents the base rate for success in the grade 1 sample (or .67 in grade 2), could be used to maximize one index over the other. In grade 1 a change to .70 raises sensitivity and negative predictive power, lowers specificity and positive predictive power, and reduces the overall percentage correct.

References

American Institutes for Research. (2007). *Reading First state APR data*. Washington, DC: American Institutes for Research. Retrieved January 9, 2014, from http://www2.ed.gov/programs/readingfirst/state-data/achievement-data.pdf.

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

Cohen, J. (1992). A power primer. *Psychological Bulletin*, 112, 155–159.

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. http://eric.ed.gov/?id=EJ742190

Fletcher, J. M., Foorman, B. R., Boudousquie, A., Barnes, M., Schatschneider, C., & Francis, D. J. (2002). Assessment of reading and learning disabilities: A research-based, treatment-oriented approach. *Journal of School Psychology*, 40(1), 27–63. http://eric.ed.gov/?id=EJ642598

Florida Department of Education. (2009). *FAIR 3–12 technical manual*. Tallahassee, FL: Florida Department of Education. Retrieved April 10, 2013, from www.fcrr.org/FAIR/Technical manual - 3–12-FINAL_2012.pdf.

Francis, D., Fletcher, J., Stuebing, K., Lyon, G. R., Shaywitz, B. A., & Shaywitz, S. E. (2005). Psychometric approaches to the identification of LD: IQ and achievement scores are not sufficient. *Journal of Learning Disabilities*, 38(2), 98–108. http://eric.ed.gov/?id=EJ695597

Jenkins, J. R. (2003, December). *Candidate measures for screening at-risk students*. Paper presented at the National Research Center on Learning Disabilities' Responsiveness-to-Intervention Symposium, Kansas City, MO. Retrieved December 9, 2013, from http://www.nrcld.org/symposium2003/jenkins/index.html.

National Early Literacy Panel. (2008). *Developing early literacy: Report of the National Early Literacy Panel*. Washington, DC: National Institute for Literacy. Retrieved January 9, 2014, from http://lincs.ed.gov/publications/pdf/NELPReport09.pdf.

National Institute of Child Health and Human Development. (2000). *National Reading Panel–Teaching children to read: Reports of the subgroups* (NIH Pub. No. 00–4754). Washington, DC: U.S. Department of Health and Human Services. Retrieved January 9, 2014, from https://www.nichd.nih.gov/publications/pubs/nrp/Pages/report.aspx.

Petscher, Y., Kim, Y. S., & Foorman, B. R. (2011). The importance of predictive power in early screening assessments: Implications for placement in the RTI framework. *Assessment for Effective Intervention, 36*(3), 158–166. http://eric.ed.gov/?id=EJ925613

Rayner, K., Foorman, B. R., Perfetti, C. A., Pesetsky, & Seidenberg, M. S. (2001). How psychological science informs the teaching of reading. *Psychological Science in the Public Interest*, 2(2), 31–74.

Schatschneider, C., Petscher, Y., & Williams, K. M. (2008). How to evaluate a screening process: The vocabulary of screening and what educators need to know. In L. Justice & C. Vukelic (Eds.), *Every moment counts: Achieving excellence in preschool language and literacy instruction* (pp. 304–317). New York: Guilford Press.

Snow, C. E., Burns, M. S., & Griffin, P., eds. (1998). *Preventing reading difficulties in young children.* Washington, DC: National Academy Press. Retrieved January 9, 2014, from http://www.nap.edu/readingroom/books/reading/.

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

Therneau, T. M., Atkinson, B. & Ripley, B. (2013). *rpart: Recursive Partitioning.* R package version 4.1–4. Retrieved December 16, 2013, from http://cran.r-project.org/web/packages/rpart/rpart.pdf.