



February 2015

Applied Research Methods

Comparing methodologies for developing an early warning system: Classification and regression tree model versus logistic regression

Sharon Koon

Yaacov Petscher

Florida Center for Reading Research
at the Florida State University

Key findings

The classification and regression tree (CART) model is an emerging tool in the development of early warning systems for identifying students at risk of poor performance in reading. This study finds that CART results are consistent with those of logistic regression on all measures of classification accuracy while using fewer or the same number of variables and making fewer model assumptions.

REL 2015–077

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.

February 2015

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 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.

This REL report is in the public domain. While permission to reprint this publication is not necessary, it should be cited as:

Koon, S., & Petscher, Y. (2015). *Comparing methodologies for developing an early warning system: Classification and regression tree model versus logistic regression* (REL 2015–077). 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

Early warning systems provide opportunities for student interventions or differentiated instruction that may prevent an anticipated negative outcome such as failing a state summative assessment. Negative outcomes often interrupt a student's education progress. For example, a student who fails a summative assessment on reading could be retained in the current grade, or a student with poor academic performance could drop out of school. To identify students likely to experience a negative outcome, two prominent statistical methodologies are available: logistic regression and classification and regression tree (CART) models. The CART model is an emerging tool in the field of education, and limited research exists on its comparability with the more widely used logistic regression when using multivariate assessments of reading to screen for reading difficulties.

Regional Educational Laboratory (REL) Southeast recently released a technical report summarizing the results of a comparison between logistic regression and CART models in identifying at-risk readers in Florida (Koon, Petscher, & Foorman, 2014). The technical report emphasized testing the classification accuracy of a new screening assessment in Florida using the two statistical approaches. In that report the two statistical approaches were introduced to help readers understand how logistic regression and CART models are constructed. This report poses the same research question but focuses on the details of each method. The additional details are meant to help analysts interested in developing early warning systems use the CART model. Highlights of the model complexities and evaluation criteria will also help state education leaders understand the need for expert assistance when developing both logistic regression and CART models to determine which students are at risk.

Using data from a sample of Florida public school students in grades 1 and 2 in the 2012/13 school year, the study found that the CART model performed comparably with logistic regression on all measures of classification accuracy while using fewer or the same number of variables and making fewer model assumptions. In addition, evidence from the health care field suggests that CART models may be easier for some practitioners to understand and implement quickly than logistic regression. Their use in early warning systems could be studied to determine whether school staff find them easier to use than logistic regression.

Contents

Summary	i
Why this study?	1
Two models for developing early warning systems	1
Use of the CART model in education	3
What the study examined	4
Logistic regression is accurate, but results can be complex to interpret	4
Classification and regression tree model: How to develop a classification tree with high negative predictive power and a simple structure	5
How the study was conducted	11
Logistic regression	11
Classification and regression tree	12
What the study found	12
Grade 1 logistic regression results	13
Grade 1 classification and regression tree results	14
Grade 2 logistic regression results	16
Grade 2 classification and regression tree results	17
Implications of the study	19
Limitations of the study	19
Appendix A. Measures of classification accuracy	A-1
Appendix B. Data, measures, and outliers and missing data	B-1
Appendix C. Classification tables	C-1
Notes	Notes-1
References	Ref-1
Boxes	
1 Logistic regression model	5
2 Classification and regression tree model parameters	8
Figures	
1 CART example model 1	7
2 CART example model 2	9
3 CART example model 3	10
4 Grade 1 classification and regression tree complexity parameter values by cross-validation relative error	15
5 Grade 1 CART decision rules	15

6	Grade 2 classification and regression tree complexity parameter values by cross-validation relative error	17
7	Grade 2 CART decision rules, model 1	18
8	Grade 2 CART decision rules, model 2	18

Tables

1	Sample classification table for CART model 1	7
2	Sample classification table for CART model 2	9
3	Sample classification table for CART model 3	10
4	Summary of results by model	13
5	Grade 1 logistic regression model evaluation	13
6	Grade 1 logistic regression final model	14
7	Grade 2 logistic regression model evaluation	16
8	Grade 2 logistic regression final model	16
A1	Sample 2×2 classification table	A-1
B1	Grade 1 missing data statistics ($n = 986$)	B-2
B2	Grade 2 missing data statistics ($n = 887$)	B-2
C1	Grade 1 logistic regression classification table ($n = 206$)	C-1
C2	Grade 1 CART classification table ($n = 206$)	C-1
C3	Grade 2 logistic regression classification table ($n = 181$)	C-1
C4	Grade 2 CART classification table, model 1 ($n = 181$)	C-1
C5	Grade 2 CART classification table, model 2 ($n = 181$)	C-2

Why this study?

Research has shown that students unable to read by the end of grade 3 are likely to falter in the later grades and ultimately drop out of high school. Hernandez (2011) found that grade 3 students who do not read proficiently are four times more likely not to graduate from high school on time than grade 3 students who read proficiently.¹ Because of the general consensus among education researchers that students must be ready to “read to learn” by grade 4, state legislatures around the country have enacted laws requiring that students demonstrate an acceptable level of reading comprehension on the state assessment of reading in grade 3 or face possible retention. School districts must also monitor and report the progress of their students in reading before entering grade 3 so that students who need interventions can be identified.

These laws have led to the need for early warning systems that can identify students who may be at risk for reading difficulties in grade 3 (Good III, Simmons, & Kame’enui, 2001; Shapiro, Solari, & Petscher, 2008). Given the importance of these decisions for students, these systems must have high rates of classification accuracy. Because reading is a multi-dimensional skill, systems based on multiple tests or measures are recommended to improve accuracy over single tests or measures (Fletcher et al., 2002; Francis et al., 2005). Therefore, early warning systems in reading are best built using multivariate assessments of reading skills.

The term “early warning system” is often associated with the identification of students at risk for dropping out of high school (Bruce, Bridgeland, Fox, & Balfanz, 2011; Carl, Richardson, Cheng, Kim & Meyer, 2013; Davis, Herzog, & Legters, 2013; Johnson & Semmelroth, 2010; Neild, Balfanz, & Herzog, 2007), but early warning systems may be viewed more broadly to include sets of assessments used to classify students at risk or not at risk for a particular benchmark (for example, dropping out of school, not passing a state math achievement test, or having a learning disability). Education research focusing on early warning systems for younger students tends to use alternative nomenclature such as classification accuracy, diagnostic accuracy, screening accuracy, and diagnostic systems.

While the language of early identification systems varies, the methodology used to identify at-risk individuals is often the same. Most studies focused on early identification employ logistic regression (Beach & O’Connor, in press; Catts, Fey, Zhang, & Tomblin, 2001; Piasta, Petscher, & Justice, 2012). Few studies have used classification and regression tree analysis (CART) (Compton, Fuchs, Fuchs, & Bryant, 2006; Fuchs, Compton, Fuchs, Bryant, & Davis, 2008; Fuchs, Fuchs, & Hamlett, 2007), a potentially useful method for classifying individuals that is often used in the health care field. Because the CART model allows for a simpler presentation of how students are classified as at risk or not at risk, it was of interest to evaluate this technique against logistic regression to understand the extent to which both methods converged on similar classification rules and levels of diagnostic accuracy.

Two models for developing early warning systems

Parametric and nonparametric statistical methods can be used to develop early warning systems based on a set of predictor variables (multivariate assessments). Unlike nonparametric methods, parametric methods make assumptions about the underlying data. Common parametric methods include multivariate analyses (logistic regression), path

Most studies focused on early identification employ logistic regression; few studies have used classification and regression tree analysis

analysis/structural equation models, and hierarchical linear models. Because early warning systems are often created to predict a dichotomous outcome (such as passing or failing a test), logistic regression is frequently used to develop classification rules. Logistic regression estimates the direct effects of a set of predictors on the outcome.

Logistic regression models. Logistic regression models with multiple predictors are most often specified in a stepwise or hierarchical fashion, in which the contribution of each predictor in explaining the variation in the outcome can be evaluated. Predictors that do not statistically contribute or improve the model fit are typically not retained in the model. Once an optimal solution is reached, the model produces a mean log-odds for the likelihood of achieving one of the two categories for the selected outcome (for example, to pass or fail a test) conditional on one or more predictor variables in the model.

Catts et al. (2001), one of the first empirical works that allowed practitioners to fully use a multivariate assessment to estimate a student's likelihood of having reading difficulties, applied a stepwise logistic regression analysis to identify risk factors in kindergarten students. Language, early literacy, and nonverbal cognitive measures were used as predictors of reading difficulties in grade 2. Five measures (letter identification, sentence imitation, phonological awareness, rapid naming, and mother's education) were found to uniquely predict grade 2 reading outcomes. With the estimated logistic regression formula, practitioners can estimate a student's probability of having reading difficulties in grade 2.

Classification and regression tree model. Unlike logistic regression, the CART model is a nonparametric approach. Nonparametric approaches make fewer assumptions about the underlying populations from which the data are obtained, are relatively less sensitive to outlying observations, and are often easier for practitioners to understand. Specifically, the CART model does not make distributional assumptions, does not require any functional form for the predictors, and does not assume additivity of the predictors, which allows the identification of complex interactions (Gordon, 2013).

The estimation process of the CART model is similar to logistic regression: students are classified as "at risk" or "not at risk" on a future outcome according to their observed performance on one or more assessments within a testing battery. Unlike logistic regression, the CART model classifies students as at risk or not at risk based on a set of if-then statements, rather than using beta coefficient weights to estimate log-odds or predicted probabilities. The CART model results are presented in a useful and easy-to-interpret "tree" format (Breiman, Friedman, Olshen, & Stone, 1984; Berk, 2008; Lewis, 2000).

CART models are used mainly in medical applications (diagnosis and prognosis of illnesses) because of the method's suitability in generating clinical decision rules (Lewis, 2000). For example, Jarvis et al. (2013) used a CART model to develop and validate an early warning system for hospital mortality in emergency medical admissions that was easier and faster than the complex existing system, which used logistic regression based on seven laboratory tests. Using the CART model, Jarvis et al. developed a simple paper-based system that emergency staff could use to make quick decisions to identify patients at risk of in-hospital mortality.

Similarly, Takahashi et al. (2006) overcame challenging clinical prediction rules established through logistic regression with a simple decision-tree model of in-hospital mortality

Unlike logistic regression, the CART model is a nonparametric approach, which makes fewer assumptions about the underlying populations, is relatively less sensitive to outlying observations, and is often easier for practitioners to understand

risk stratification for intensive-care patients. Twenty-nine predictors, representing data available at the time of initial patient evaluation, were used in the model. The final CART model retained 3 of the 29 predictors in stratifying patients into four levels of risk. The CART model was found to have slightly higher classification accuracy than the logistic regression model and was selected for practical use by clinicians.

In a nursing research study, a CART model illustrated associations between variables that provided insight into previously unknown patterns in the data (Kuhn, Page, Ward, & Worrall-Carter, in press). The study team predicted that CART models would play an important role in future nursing research designed to improve patient care.

Both models working together. Parametric and nonparametric models can work together, contributing different advantages. Kuhnert, Do, and McClure (2000) studied the use of three models—CART, multivariate adaptive regression splines (MARS), and logistic regression—in the context of motor vehicle injuries. A key finding was that nonparametric methods (CART and MARS) can serve both as primary modeling tools and as exploratory tools before using parametric methods like logistic regression. Combining parametric and nonparametric methods allows the analyst to build on the strengths of each approach. For example, the study team found that the CART and MARS models produced variable importance rankings that may be used in the stepwise specifications of logistic regression. The MARS model was able to identify interactions between predictors as well as a group of outliers, both of which have important implications in logistic regression. The CART model was able to identify informative primary splits, which could be used in logistic regression to separate samples into distinct groups. Both the CART and MARS methods contributed to identifying important variables and understanding their contribution to the modeling of the outcome variable. Although the CART model did not outperform the MARS model in identifying interactions in this study, most splitting criteria in a CART model represent interaction effects (Berk, 2008).

While the CART model is more frequently found in medical literature, it has begun to emerge in education research

Use of the CART model in education

While the CART model is more frequently found in medical literature, it has begun to emerge in education research. Compton et al. (2006) compared a CART model with logistic regression in the identification of early warning signs of at-risk readers in grade 1 to examine ways to improve the accuracy of predicting which children should enter secondary intervention because of elevated risk for developing reading difficulties. The diagnostic accuracy was significantly better for the CART model than for the logistic regression prediction model. The improvement in diagnostic accuracy may be in part due to the complexity of the resulting decision trees, a factor that often decreases the interpretability of the CART results. Like 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 classification tree using model specifications, so that the resulting tree is not overfit to the data. A limitation of this study is that the study team did not seek the most parsimonious CART model in its comparison.

Although Compton et al. (2006) introduced the CART methodology in identifying at-risk readers, it is still an emerging tool in education. To show its limited use in education, it is instructive to look at dropout prevention, which has a long history of research designed to identify early warning signs of negative student outcomes.

In a review of peer-reviewed studies published over 25 years (1983–2007), Rumberger and Lim (2008) found that the majority of the 389 analyses were conducted using logistic regression, with other methods including path analysis and multilevel modeling. None of the reviewed studies used CART models. A study of four states with early warning systems of students at risk for dropping out found that all four relied on systems based on logistic regression models (Ryan, 2011).

Recently, Knowles (2014) studied the use of 28 separate algorithms, including CART models, in developing a dropout early warning system in Wisconsin. The state wanted to identify early warning systems that were both accurate and easy to communicate to support decisions that improve student outcomes. Results from a generalized linear model analysis were taken as the baseline. Algorithms were compared using the area-under-the-curve metric. Most of the algorithms, including CART models and generalized linear models, resulted in an area under the curve of between .83 and .87, showing comparable accuracy in making classification decisions. While the generalized linear model was chosen for use in Wisconsin, partly because of its familiarity to most quantitative researchers, the inclusion of CART models extends the literature on dropout-prevention early warning systems.

This study used data from a sample of students in grades 1 and 2 in Florida public schools during 2012/13 to examine how CART models compare with logistic regression models in predicting poor performance on the reading comprehension subtest of the Stanford Achievement Test

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 school year to answer this research question: How do CART models compare with logistic regression models in predicting poor performance on the reading comprehension subtest of the Stanford Achievement Test?

Model comparisons are based on several traditional indexes of classification accuracy, including sensitivity, specificity, negative predictive power, and positive predictive power. Classification accuracy in each grade was based on the accuracy of Florida Assessments for Instruction in Reading–Florida Standards (FAIR-FS) tests in predicting end-of-year reading scores on the Stanford Achievement Test Series, Tenth Edition (SAT-10). Measures of classification accuracy are described in appendix A.

Logistic regression is accurate, but results can be complex to interpret

Logistic regression is an extension of simple or multiple regression, in which a dichotomously scored dependent variable is regressed on one or more selected independent variables. This technique is widely used to predict log-odds of success on the dependent variables. It is also used to study the rates of true and false positives and negatives as they relate to the classification of individuals as at risk or not at risk for achieving success. When presenting scores from only one or two tests, logistic regression contingency tables are easy to interpret (box 1). But with more tests and score ranges, contingency tables become more complex and difficult to explain, not lending themselves to simple snapshots.

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

Box 1. Logistic regression model

The logistic regression approach 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:

$$\ln \frac{p}{1-p} = \beta_0 + \beta_1 X_1 + \beta_2 X_2.$$

The beta coefficients ($\beta_0, \beta_1, \beta_2$) can be used to estimate predicted log-odds, which can be converted to a predicted probability of success on the outcome. Logistic regression results can be used to generate more straightforward contingency tables, such as shown in the table. Similar to what is seen in figure 1 in the main text, students with test 1 scores below 244 and test 2 scores below 350 are identified as at risk (shaded cells in table below), while other students are identified as not at risk.

Logistic regression contingency table example

Test 2	Test 1				
	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

Source: Author’s illustration.

Classification and regression tree model: How to develop a classification tree with high negative predictive power and a simple structure

A CART model approach to early warning systems in education could maintain numerous benefits from both statistical and practical vantage points. Although several published studies have evaluated CART with other methods of early warning, the body of work in this area is limited. Subsequently, a goal of this study is to add to the literature and compare the diagnostic accuracy of CART and logistic regression so that others may see the merit of each application.

This study used the Florida data to develop a specific CART model for each grade that would capture all or most students at risk and have a simple reporting structure. Prior to discussing the Florida data, this section explains the basic process of fitting and pruning a CART model that achieves parsimony. Because CART models are known for overfitting the data, it is important to understand how to prune the classification tree so that it can be expected to be robust (Lawrence & Wright, 2001). In this context, a robust CART model has practical implications—for example, students who need help will not be screened out, teachers can understand the decision rules—and meets technical requirements—for example, negative predictive power is at least 0.85 while maintaining acceptable sensitivity (proportion of true positives) and specificity (proportion of true negatives).

The classification tree. The CART model classifies individuals into mutually exclusive subgroups of a population using a nonparametric approach that results in a classification

tree (Breiman et al., 1984). The CART model searches for the optimal split on the predictor variables and splits the sample into binary subsamples called nodes. In a visual representation of the model, when the split between the samples occurs, the nodes form either a rectangular box or a circle (figure 1). Boxes, referred to as terminal nodes, do not split further, but circles, referred to as nonterminal nodes, split again when there is a difference between students on the predictor variables or when a stopping rule has been reached.

The subgroup splits in the CART model are determined by the software program (such as the Recursive Partitioning and Regression Trees [rpart] package) to improve the overall classification accuracy. The CART model uses an exhaustive subgroup comparison to identify the best predictors and predictor levels that most efficiently split the sample into the most homogeneous subgroups of individuals who are identified as at risk or not at risk based on their observed scores. A variable may appear in the CART model many times because the search for the single variable that will result in the best subsequent split to the data includes all variables at each split (Therneau & Atkinson, 2013).

In this study the CART model yields a classification flowchart that clearly shows how a student is identified as at risk or not at risk for reading comprehension difficulties. For example, suppose that we have 100 students with data on four hypothetical screening assessments that all have score ranges of 100–1,000. In addition, there is an outcome assessment with a pass/fail indicator. The results of a CART model using these data are shown in figure 1.

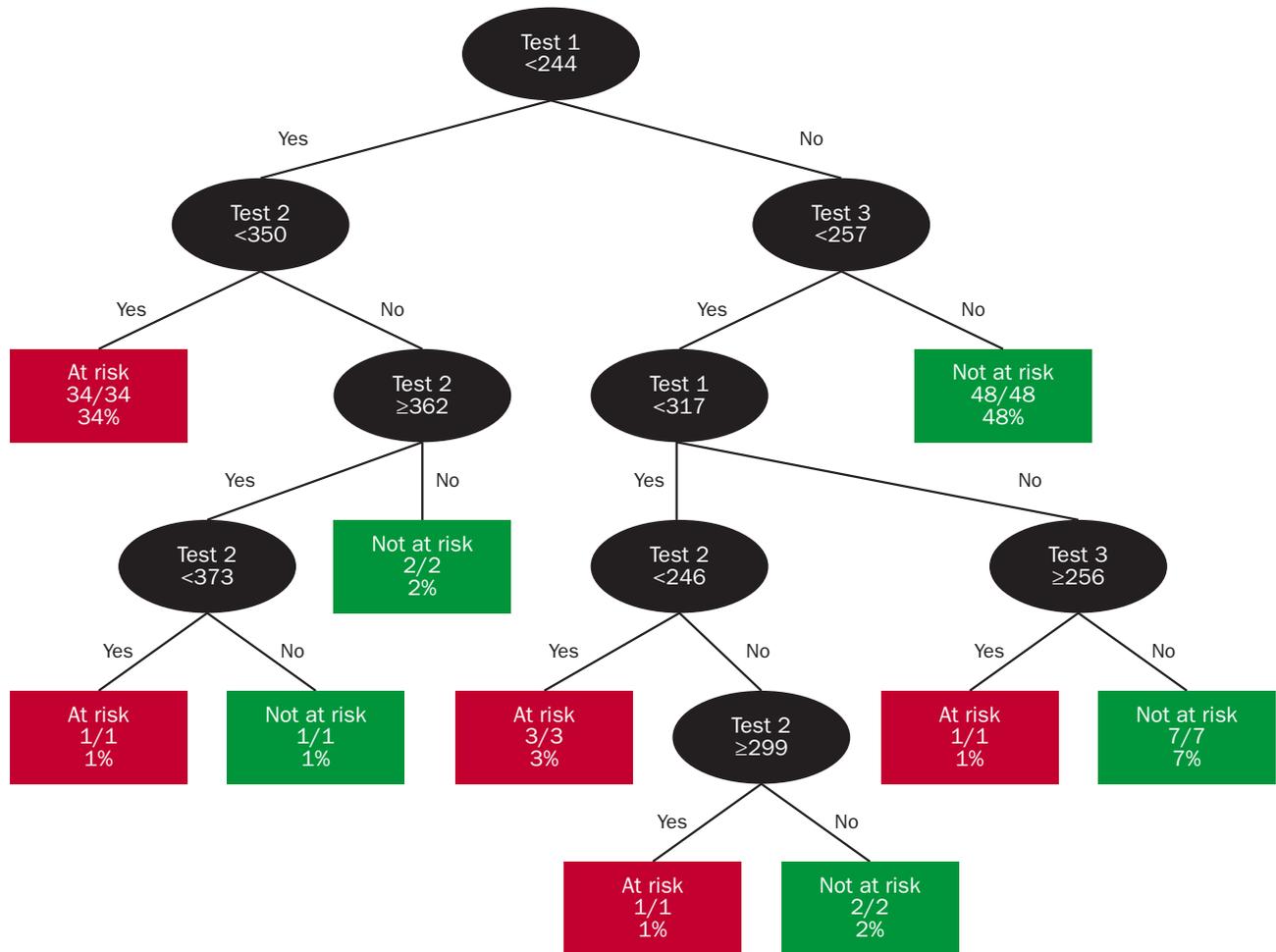
In this study the CART model yields a classification flowchart that clearly shows how a student is identified as at risk or not at risk for reading comprehension difficulties

Rules for identifying students at risk. There are five rules for identifying students as at risk and five rules for identifying students as not at risk for failing the outcome assessment. The rules are derived using the criterion specified in each nonterminal node followed with a “yes” or “no” answer. All “yes” answers split to the left, and all “no” answers split to the right, which is accomplished by changing the signs in the nonterminal nodes (“<” or “≥”) if necessary. For example, if splitting to the left, students who score less than 244 on test 1 and less than 350 on test 2 are identified as at risk according to 1 of the 10 rules. Splitting to the right, students who score at least 244 on test 1 and at least 257 on test 3 are identified as not at risk.

Information on the classification accuracy of the tree is also shown in figure 1. Under each terminal node the number of students correctly identified out of the total number in that node is provided, as well as the percentage of the 100 students placed in that terminal node. So, for instance, 34 students with test 1 scores less than 244 and test 2 scores less than 350 are correctly identified as at risk, since all students meeting this rule failed the outcome assessment (34/34). Thirty-four percent of the 100 students tested are found in this terminal node.

Model 1: Accurate but complex. All 100 students are correctly classified by the model 1 decision tree. The classification accuracy results for model 1 are summarized in table 1, which shows that all measures of classification accuracy are maximized. In addition, the *R*-squared for this tree is 1.0, indicating a perfect fit of the model to the sample data. However, these results should be considered exploratory, as the complexity of the classification rules may not generalize well to other samples because it is a saturated model. On inspection, some of the decision rules are nonsensical and are most likely a result of overfitting the model to the data. For example, students with higher test scores are classified as at risk by several rules.

Figure 1. CART example model 1



CART is classification and regression tree.

Source: Authors' illustration.

Table 1. Sample classification table for CART model 1

Screening assessment	Outcome assessment	
	Fail	Pass
At risk	40	0
Not at risk	0	60

CART is classification and regression tree.

Source: Authors' illustration.

Generalizing the model to other samples can be accomplished using two methods: *v*-fold cross-validation within the analysis and testing the decision rules using a separate validation sample. The *v*-fold cross-validation method is used for estimating error rates without test data and thus is preferred when the dataset is small and the use of a validation sample is not possible (Salford Systems, n.d.). This method partitions the sample into a specified number of subsamples (“*v*”) and estimates error rates for a tree with the maximum number of splits, the greatest complexity, and subtrees with fewer splits. If a separate validation

sample is feasible, the decision-tree rules can be applied to the validation sample, and classification accuracy can be assessed. For this example it is assumed that model 1 is overfit to the small sample of students in the dataset and should be revised.

Pruning the classification tree. Therneau and Atkinson (2013) outline various parameters that control the fit of the model to the data and, therefore, the complexity of the tree. For this study the following parameters are pertinent: minimum split, minimum complexity parameter, and loss matrix (box 2).

The minimum split and complexity parameters were set to zero in model 1, and the default loss matrix was not changed. These specifications resulted in the maximum number of splits, with the tree splitting until there were no differences among students on the outcome assessment in each terminal node. All students in the at-risk terminal nodes failed the outcome assessment, and all students in the not-at-risk terminal nodes passed the outcome assessment.

Model 2: Easy to read but with false negatives. The specifications were changed in model 2 to attempt to fit a more parsimonious model (figure 2). Guided by Compton et al. (2006), a minimum split size of three students was specified given a sample size of 100. Increasing the minimum split size from 0 to 3 in this example will decrease the number of splits. In addition, for illustration purposes, the number of splits was limited by specifying a minimum complexity parameter of .02.

Decision trees with a limited number of splits, such as the one in figure 2, are easier to interpret than those with many splits. Model 2 has only two rules for identifying students as not at risk and only one rule for identifying students as at risk. While this model is less complex, changes in classification accuracy and model fit must be evaluated. In this case there is a reduction in complexity without sacrificing too much relevant information.

Decision trees with a limited number of splits are easier to interpret than those with many splits

Box 2. Classification and regression tree model parameters

Minimum split. This parameter specifies the minimum number of cases that must exist in a node for a split to be attempted. Increases in the minimum split generally decrease the number of splits.

Minimum complexity parameter. This parameter specifies the minimum decrease in the overall lack of fit that must result from an additional split. Fit is measured by the model's relative error, which is equivalent to $1 - R$ -squared (Steinberg, 2013), as well as the relative error found in the cross-validation samples (the cross-validation relative error) by the number of splits. 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). Tables and plots of the cross-validation results can be consulted to determine an appropriate complexity parameter value. The use of this model-based statistic as a splitting criterion changes the model to an "essentially" non-parametric approach (Harrell, 2001).

Loss matrix. A loss matrix is used to weight classification errors differently. To increase the negative predictive power, the specification would be to view false negatives as more costly. The default specification is to weight all classification errors equally.

Figure 2. CART example model 2



CART is classification and regression tree.

Source: Authors' illustration.

Classification accuracy decreases when comparing the classifications of model 1 to model 2 (figures 1 and 2). In model 2, four students (4 percent of the total) have test 3 scores less than 244 and test 1 scores equal to or greater than 350, putting them in the not-at-risk category. However, only three students out of the four are correctly identified (passed the outcome assessment), with the remaining student representing a false negative (predicted to pass but failed). Similarly, 62 students (62 percent of the total) have scores equal to or greater than 244 on test 1 and are classified in the not-at-risk category. Five of the 62 students are incorrectly identified and therefore represent additional false negatives. The six false negatives are shown in table 2. In addition to a decrease in classification accuracy, there also is a reduction in the *R*-squared for model 2 from 1.0 to 0.85.

Model 3: Still easy to read with some false positives. To minimize false negatives while maintaining high overall classification accuracy, a revision can be made to the model 2 specifications to add a loss matrix (see box 2), in which false negatives are weighted as two times as costly as false positives. The decision of how much weight to apply is based on the inherent costs associated with misclassification (Lewis, 2000).

The revised tree is presented in figure 3. The addition of a loss matrix resulted in two additional splits and a change in the tests used to make the splits. These changes resulted in an *R*-squared of 0.92.

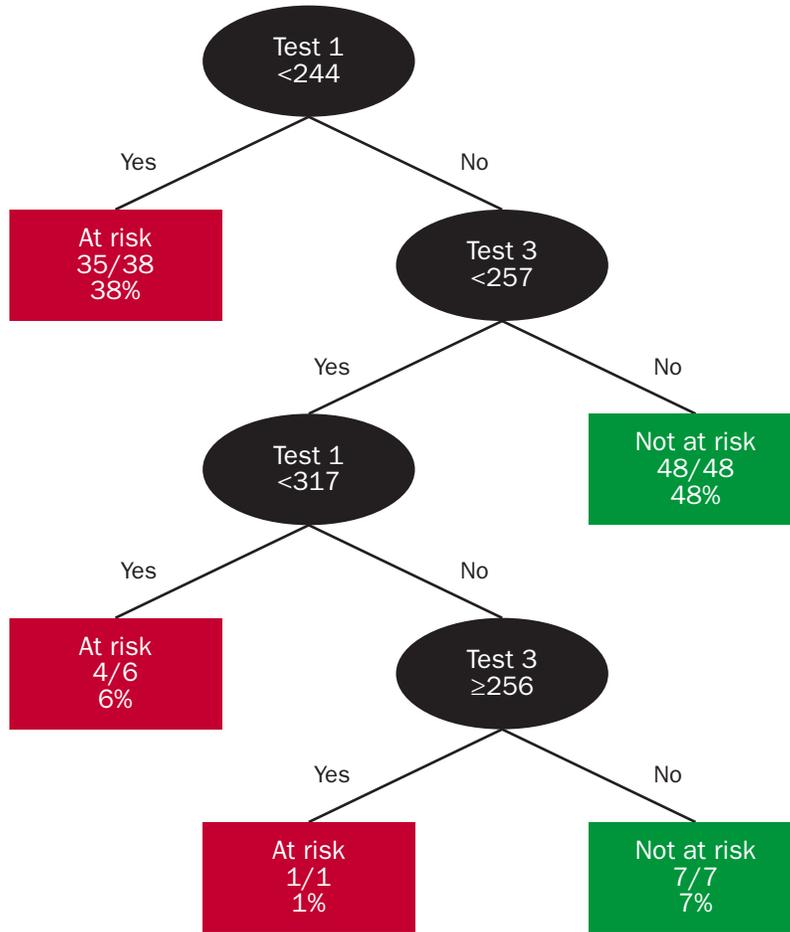
Table 2. Sample classification table for CART model 2

Screening assessment	Outcome assessment	
	Fail	Pass
At risk	34	0
Not at risk	6	60

CART is classification and regression tree.

Source: Authors' illustration.

Figure 3. CART example model 3



CART is classification and regression tree.

Source: Authors' illustration.

The updated contingency table for model 3 (table 3) shows that the misclassification errors are now false positives (five students are classified in the risk category by the classification rules, but did not fail the outcome test). While there is a decrease in the overall classification accuracy, as well as the *R*-squared, model 3 is preferred over models 1 and 2 because of its simple structure (versus model 1) and high negative predictive power (versus model 2).

An interaction effect, in which there are two different cutpoints for each assessment used in the tree, is also shown in figure 3. The CART model allows the predictors to interact with each other, such that different combinations of cutpoints may be used to differentially

Table 3. Sample classification table for CART model 3

Screening assessment	Outcome assessment	
	Fail	Pass
At risk	40	5
Not at risk	0	55

CART is classification and regression tree.

Source: Authors' illustration.

classify a student's level of risk. Though logistic regression allows the estimation of an interaction coefficient among independent variables, the estimation is often difficult to interpret (Lemon, Roy, Clark, Friedmann, & Rakowski, 2003; Steadman et al., 2000).

Another useful piece of information provided by the CART model using rpart software is a determination of a variable's importance on a scale of 1 to 100. Variable importance is based on the number of times a variable is used in making splits and its splitting efficiency. Both primary splits (the black circles) and surrogate splits are considered.² Tests 1 and 3 were determined to have the greatest importance, followed by tests 4 and 2. While test 1 was determined to be the most important with a value of 40, the remaining tests were all around 20, with test 3 at 21, test 4 at 20, and test 2 at 19.

How the study was conducted

The “developmental ability” scores from each FAIR-FS test were used in a series of logistic regression and CART models to predict end-of-year performance on the SAT-10. Traditional indexes of classification accuracy were used to assess differences in the results between the approaches. Information on the data and measures can be found in appendix B.

Before conducting the analyses, the grade-based correlations among the FAIR-FS test scores from the individual literacy components were examined for multicollinearity in each of the imputed files. None of the Pearson correlations was higher than 0.80, eliminating concerns of redundancy in the subsequent logistic regression analyses. After 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 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. An 80/20 split is a common division in statistical learning and data mining when conducting cross-validation (Salford Systems, n.d.). Both the CART and logistic regression models 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, and overall correct classification. Logistic regression analyses were run using SPSS Statistics 21, and CART analyses were run using rpart (R 2.15.3 package).

Logistic regression

The logistic regression models in this study were developed in a hierarchical manner. Based on the correlations between the individual FAIR-FS tests and the dichotomized SAT-10 variable, the FAIR-FS test scores were entered into the logistic regression ordered by correlational magnitude. FAIR-FS tests, which added at least 2 percent unique variance above the test 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

The “developmental ability” scores from each FAIR-FS test were used in a series of logistic regression and CART models to predict end-of-year performance on the SAT-10

shown that an R -squared between 2 percent 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 maximum likelihood in logistic regression and can be interpreted in the same way as the R -squared estimated in an ordinary least squares regression.

Classification and regression tree

CART models assessed the individual performance of each FAIR-FS test, at every available cutpoint, in classifying students into at-risk and not-at-risk categories. To ensure a parsimonious model, several specifications were used to limit the number of splits, including a minimal split size of three students. In addition, the number of splits was limited by specifying a minimum reduction in the cross-validation relative error (that is, a minimum complexity parameter), identified after running a base model with no minimum specified. Each grade-based model included tenfold cross-validation ($v = 10$) for evaluating the quality of the prediction tree and determining the appropriate minimum complexity parameter (Breiman et al., 1984). The value selected for the minimum complexity parameter was the value resulting in the fewest number of splits with a cross-validation relative error less than one standard error above the minimum cross-validation relative error. Plots of the cross-validation relative error against minimum complexity parameter values were consulted for this decision.

CART results were consistent with those from logistic regression on all measures of classification accuracy while using fewer or the same number of variables

In using the CART model, the intention was to build and prune trees based on maximizing the negative predictive power of .85. To accomplish this, revisions to the model in grade 2 included the specification of a loss matrix. The same process was used in the explanation of the CART examples presented above.

What the study found

CART results were consistent with those from logistic regression on all measures of classification accuracy (table 4) while using fewer or the same number of variables. All final models in this study yielded negative predictive power of or exceeding .85—that is, false negative rates ranged from 4 percent 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 CART results were better than or equal to the logistic regression results on all indexes of classification accuracy. Both methods resulted in models that retained three of the four available tests, 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. In grade 2 the logistic regression and CART results were comparable, after the addition of a loss matrix in which the false negatives were treated as two times the cost of false positives to the CART model specifications (model 2). 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.

Table 4. Summary of results by model

Model	Sensitivity	Specificity	Positive predictive power	Negative predictive power	Overall proportion correct
Grade 1					
Classification and regression tree	.92	.90	.79	.96	.90
Logistic regression	.87	.90	.78	.94	.89
Grade 2					
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

Note: Sample size is 206.

Source: Authors' analysis of data from the Florida Center for Reading Research; see appendix B for details.

Grade 1 logistic regression results

Model building. Based on the correlations between the individual FAIR-FS tests and performance on the SAT-10, the FAIR-FS test 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 ($r = .41$). All correlations are significant at the 0.01 level. The results based on the calibration dataset are provided in table 5.

Because of the minimal increase in the explained variance based on the Nagelkerke pseudo R -squared (1.7 percent), the “following directions” test was deleted from the model. The final model coefficients are provided in table 6, with a Nagelkerke pseudo R -squared of 72 percent.

Table 5. Grade 1 logistic regression model evaluation

Block	Variable	Hosmer and Lemeshow test p value	Nagelkerke pseudo R squared	Change in Nagelkerke pseudo R squared	Overall percentage correct
0	Constant	na	na	na	70.1
1	Word reading	.49	.66	na	86.5
2	Word building	.67	.68	.023	87.1
3	Vocabulary pairs	.91	.72	.039	88.6
4	Following directions	.72	.74	.017	88.5

na is not applicable.

Note: Sample size is 780.

Source: Authors' analysis of data from the Florida Center for Reading Research; see appendix B for details.

Table 6. Grade 1 logistic regression final model

Variable	Coefficient (β)	Standard error	Wald statistic	Degrees of freedom	Significance level	Exp(β)	95 percent confidence interval for Exp(β)	
							Lower	Upper
Word reading	0.03	0.00	102.51	1	.00	1.03	1.03	1.04
Word building	0.01	0.00	11.00	1	.00	1.01	1.00	1.01
Vocabulary pairs	0.01	0.00	36.30	1	.00	1.01	1.01	1.01
Constant	-21.75	1.82	142.87	1	.00	0.00	na	na

na is not applicable.

Note: Sample size is 780.

Source: Authors' analysis of data from the Florida Center for Reading Research; see appendix B for details.

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

$$\text{Logit} = -21.749 + 0.032*(\text{word reading score}) + 0.006*(\text{word building score}) + 0.009*(\text{vocabulary pairs score}),$$

to calculate predicted SAT-10 logit scores for each case in the validation dataset, which were then transformed to probabilities. Probabilities equal to or greater than .5 were recoded as 1 (scoring at or above the 40th percentile on the SAT-10),³ and all other values were coded as 0. The results were used to generate a classification table for use in calculating indices of classification accuracy (see table C1 in appendix C).

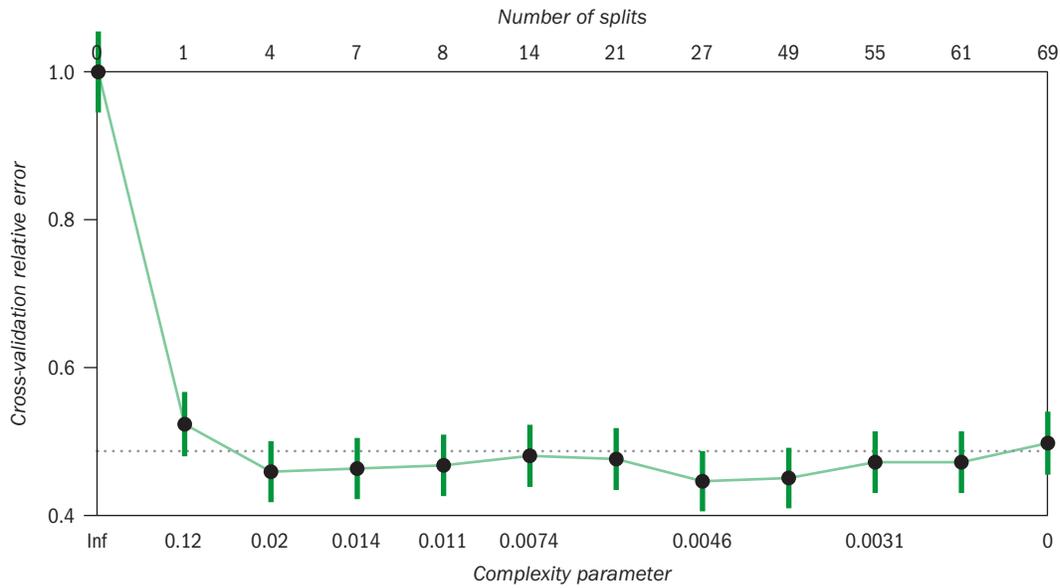
Grade 1 classification and regression tree results

Model building. All four FAIR-FS tests 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 loss 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 0.02 (figure 4). Selecting a complexity parameter of 0.02 results in a cross-validation relative error below the recommended standard of one standard error above the minimum cross-validation relative error, indicated by the dotted line. Additional cutpoints could be selected, but they would result in a larger number of splits.

The pruned tree with the final classification rules is shown in figure 5. The model *R*-squared is 0.64. Word reading was the variable found to have the greatest importance with a value of 62, followed by vocabulary pairs at 16, word building at 15, and following directions at 7. Based on the CART results, students would be identified as at risk under either of the following conditions (figure 5): the student achieved a word reading score below 452, or the student achieved a vocabulary pairs score below 465, a word reading score of 452–502, and a following directions score below 434.

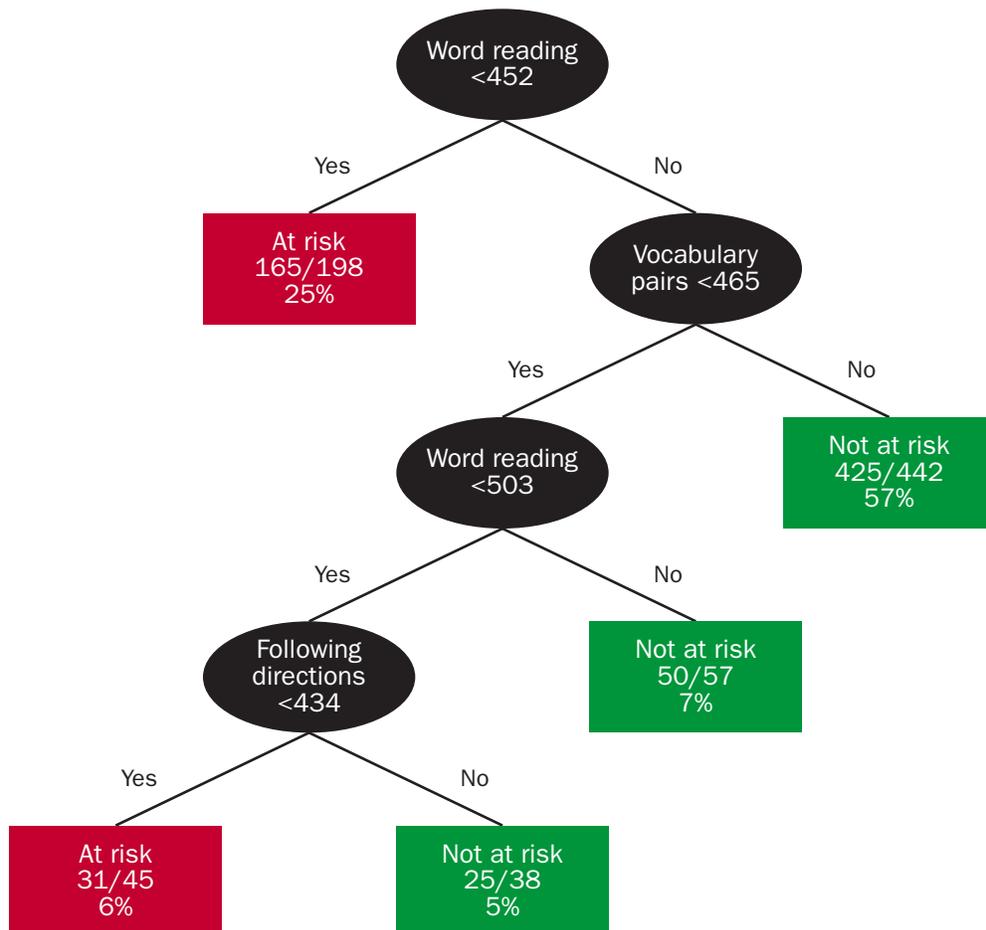
Model testing. The classification rules were applied to the validation dataset to predict group membership as well as probabilities associated with membership in each group. Using these results, a classification table was generated (see table C2 in appendix C).

Figure 4. Grade 1 classification and regression tree complexity parameter values by cross-validation relative error



Source: Authors' analysis of data from the Florida Center for Reading Research; see appendix B for details.

Figure 5. Grade 1 CART decision rules



CART is classification and regression tree.

Source: Authors' illustration.

Grade 2 logistic regression results

Model building. Based on the correlations between the individual FAIR-FS tests and performance on the SAT-10, the FAIR-FS test 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 ($r = .40$). All correlations are significant at the 0.01 level. The results based on the calibration dataset are provided in table 7.

All FAIR-FS tests were found to contribute to explaining a significant and practically important percentage of variance and were kept in the final model (table 8). About 70 percent of the variance in the logit of SAT-10 scores was explained by the FAIR-FS tests, as indicated by the Nagelkerke pseudo R -squared of .70.

Model testing. The model coefficients from the final model were used in a prediction equation to calculate predicted SAT-10 logit scores for each case in the validation dataset and the classification table for calculating indices of classification accuracy (see table C3 in appendix C).

Table 7. Grade 2 logistic regression model evaluation

Block	Variable	Hosmer and Lemeshow test p value	Nagelkerke pseudo R squared	Change in Nagelkerke pseudo R squared	Overall percentage correct
0	Constant	na	na	na	66.7
1	Word reading	.03	.58	na	84.1
2	Word building	.27	.61	.027	83.7
3	Vocabulary pairs	.26	.66	.057	85.8
4	Following directions	.50	.70	.029	85.4

na is not applicable.

Note: Sample size is 706.

Source: Authors' analysis of data from the Florida Center for Reading Research; see appendix B for details.

Table 8. Grade 2 logistic regression final model

Variable	Coefficient (β)	Standard error	Wald statistic	Degrees of freedom	Significance level	Exp(β)	95 percent confidence interval for Exp(β)	
							Lower	Upper
Word reading	0.02	0.00	44.32	1	.00	1.02	1.01	1.02
Spelling	0.01	0.00	21.92	1	.00	1.01	1.01	1.02
Vocabulary pairs	0.01	0.00	26.29	1	.00	1.01	1.01	1.01
Following directions	0.01	0.00	25.10	1	.00	1.01	1.01	1.01
Constant	-21.98	1.84	142.89	1	.00	0.00	na	na

na is not applicable.

Note: Sample size is 706.

Source: Authors' analysis of data from the Florida Center for Reading Research; see appendix B for details.

Grade 2 classification and regression tree results

Model building. All four FAIR-FS tests 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. Based on the cross-validation results from the base model, the tree was pruned by specifying a complexity parameter of 0.016 (figure 6), which was chosen over 0.096 because of greater confidence that the cross-validation error is below the threshold.

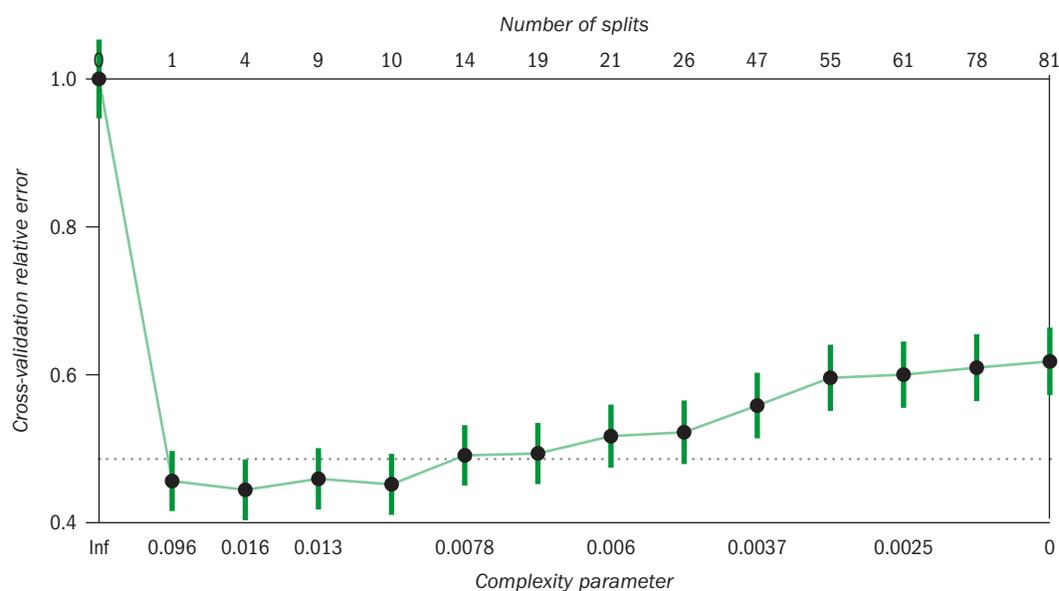
The pruned tree resulting from this specification is shown in figure 7. As expected from the table of complexity parameter values, classification rules are based on four splits. The *R*-squared for model 1 is 0.60.

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. Using these results, a classification table was generated (see table C4 in appendix C).

Model revision and testing. 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 based on a judgment of the cost of underidentifying students at risk. The pruned tree resulting from this revision is shown in figure 8. Word reading was rated as the most important variable with a value of 48, followed by spelling at 26, vocabulary pairs at 16, and following directions at 9. The *R*-squared for model 2 is 0.71.

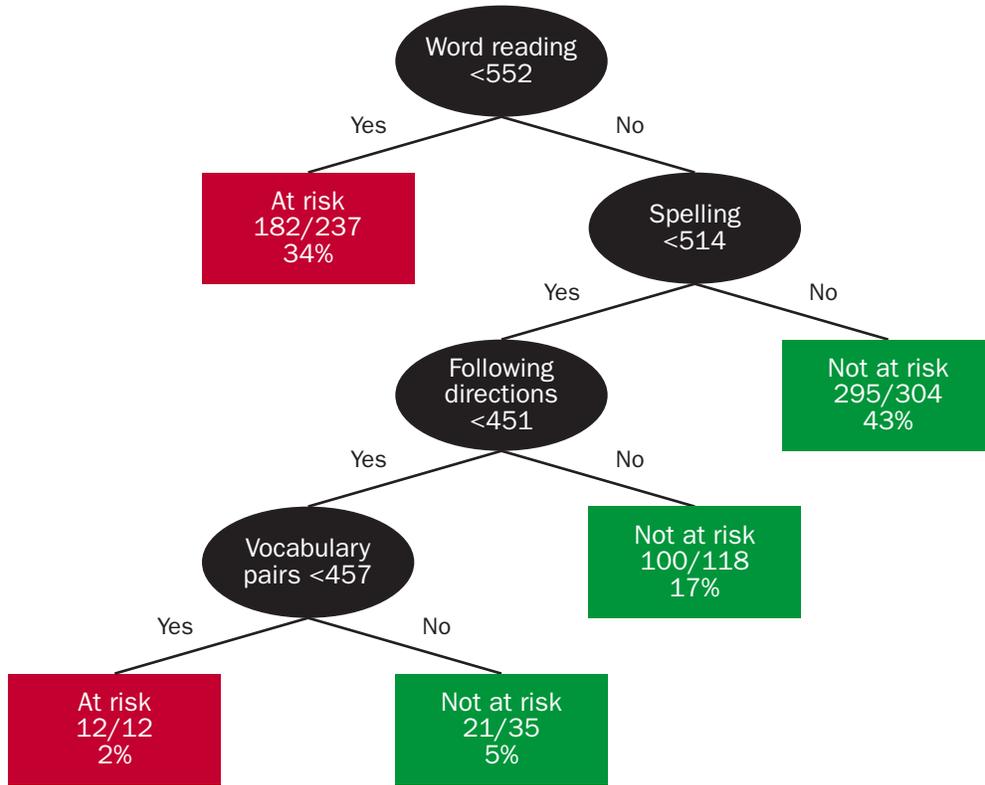
Based on the CART results, students would be identified as at risk under either of the two following conditions (see figure 8): the student achieved a word reading score below 564, or the student achieved a word reading score of 564 or above, a following directions score below 451, and a vocabulary pairs score below 494. The classification table that resulted from applying the model to the validation dataset is provided in table C5 in appendix C.

Figure 6. Grade 2 classification and regression tree complexity parameter values by cross-validation relative error



Source: Authors' analysis of data from the Florida Center for Reading Research; see appendix B for details.

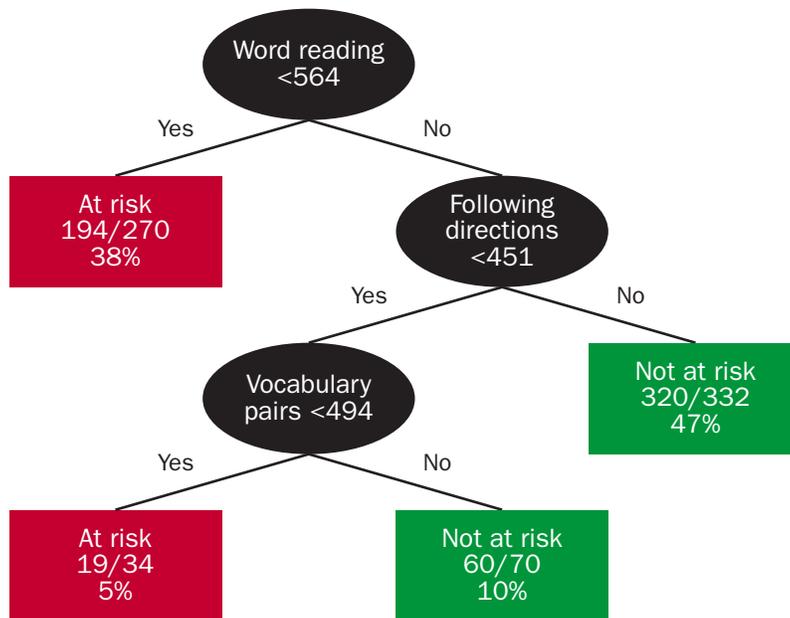
Figure 7. Grade 2 CART decision rules, model 1



CART is classification and regression tree.

Source: Authors' illustration.

Figure 8. Grade 2 CART decision rules, model 2



CART is classification and regression tree.

Source: Authors' illustration.

Implications of the study

The CART results were found to be comparable with those of logistic regression, with the results of both methods yielding negative predictive power greater than the recommended standard of .90. Given the comparability, the CART model may be more appealing in an education context because of the ease with which the results can be communicated to and used by practitioners. For instance, practitioners would be able to identify a student as at risk or not at risk by using the decision tree and would be able to know which assessments, and therefore which component skills, placed the student in an at-risk category. They could see this information in a simple paper-and-pencil format. Evidence from the health care field suggests CART models may be easier for some practitioners to understand and implement quickly. Their use in early warning systems could be studied to determine whether school staff find them easier to use than logistic regression.

CART models also hold several technical advantages over logistic regression. First, as a nonparametric method, a CART model is not sensitive to the presence of the outliers, unlike logistic regression, which is a parametric method. Second, it is not sensitive to collinearity between the variables. Third, a CART model is able to reveal complex interactions among predictors, which may be difficult or impossible to estimate in the regression framework unless the interaction terms are specified a priori.

Limitations of the study

At the same time CART has several limitations when compared with logistic regression models. For example, a notable disadvantage is that the CART model is sensitive to the presence of missing data and, thus, requires either listwise deletion or data imputation. Combining decision trees from multiple imputed files is not easily conducted in a CART model, whereas parameter estimates from multiple imputed files can be averaged in logistic regression. In addition, with logistic regression, improvements in sensitivity or specificity could be expected by adjusting the cutscores used in group classifications. Depending on the importance of one decision over another, adjustments to the model specifications could be made in both models when used in practice. The specifications used in this study were designed to meet or exceed negative predictive power of 0.85, while maintaining acceptable levels of sensitivity and specificity.

Finally, both CART and logistic regression may be used complementarily in developing an early warning system, since each method provides different tools to the researcher. Logistic regression focuses on the relative statistical significance of the predictors, while CART emphasizes the absolute effects. Both types of outcomes may be important in analyzing and interpreting the results of the models and identifying at-risk students.

Appendix A. Measures of classification accuracy

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 2×2 contingency or classification table that provides counts of individuals in four categories. In this study, students are categorized based on their performance on a screening assessment (the Florida Assessments for Instruction in Reading—Florida Standards, or FAIR-FS) and an outcome assessment (the Stanford Achievement Test Series, Tenth Edition, or SAT-10) (table A1).

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—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—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 of 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 ($D/[C+D]$).

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 et al., 2006; Jenkins, 2003). Jenkins suggested that screening assessments should demonstrate a negative predictive power of .90–.95 and a sensitivity level of .90–.95.

The developers of early warning systems are often most interested in maximizing the negative predictive power while maintaining high overall classification accuracy (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. 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.

Table A1. Sample 2×2 classification table

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

Source: Authors' illustration.

Appendix B. Data, measures, and outliers and missing data

This appendix describes the data source, the Florida Assessment of Instruction in Reading–Florida Standards (FAIR–FS), as well as outliers and missing data.

Data

Participant data were obtained from an archive containing FAIR-FS data on approximately 2,000 students in grades 1 and 2 in 15 elementary schools in the Hillsborough County school district in Florida. The archive is the result of data obtained from a linking study conducted from December 2012 to April 2013 as part of Florida State University’s subaward from the Educational Testing Service’s assessment grant in the Institute of Education Sciences/National Center for Educational Research’s Reading for Understanding initiative (Sabatini, PI; R305F100005). FAIR-FS was administered between December 3, 2012, and January 11, 2013, and the Stanford Achievement Test Series, Tenth Edition (SAT-10) between April 2 and April 12, 2013. As part of its current testing practices, Hillsborough County administered 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. The subcontract that Florida State University has with the Educational Testing Service makes it clear that the university owns the FAIR-FS and all data produced under the subcontract. These analyses are not part of the Reading for Understanding subaward.

Measures

FAIR-FS tests in grades K–2 were developed to measure print knowledge—that is, alphabetic knowledge of letter names and sounds, phonological awareness, word reading, and spelling—and language skills—that is, syntax, vocabulary, and listening comprehension. The FAIR-FS consists of alphabetic and oral language tests designed for different grade levels. The tests included in each grade-specific analysis in this study varied by grade as follows: word reading (grades 1 and 2), word building (grade 1), spelling (grade 2), vocabulary pairs (grades 1 and 2), and following directions (grades 1 and 2).

Performance on each FAIR-FS test is reported using a developmental scale. Scores range from 200 to 800, with a mean of 500 and a standard deviation of 100.

Outliers and missing data

Grade 1. The initial grade 1 dataset included 1,028 students. Twenty-seven cases were deleted because of missing SAT-10 scores, and one case was deleted because of missing data on all FAIR-FS tests. An analysis of univariate and multivariate outliers, strongly recommended before conducting logistic regression, resulted in the deletion of an additional 14 cases. The final dataset included 986 students.

The analysis of missing data in the final dataset of 986 students revealed that each test had less than 10 percent missing data (table B1). 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 multiple imputed datasets, but there is no accepted procedure for analyzing and summarizing classification trees generated

Table B1. Grade 1 missing data statistics (n = 986)

FAIR FS test	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; see appendix B for details.

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 because of 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, with missing data rates for each test of less than 5 percent (table B2). As in grade 1, a dataset with complete cases for all variables was created by aggregating the results of 20,000 imputations.

Table B2. Grade 2 missing data statistics (n = 887)

FAIR FS test	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; see appendix B for details.

Appendix C. Classification tables

Comparisons between the logistic regression and CART models are based on indexes of classification accuracy, including sensitivity, specificity, negative predictive power, and positive predictive power (see appendix A). These indexes are derived from a 2×2 classification table that provides counts of individuals in four categories. In this study, students are categorized based on their predicted performance on the SAT-10 (from the statistical model) and their observed performance on the SAT-10. This appendix provides classification tables resulting from all the models.

Table C1. Grade 1 logistic regression classification table ($n = 206$)

		SAT 10 score: observed		Total
		0 (at risk)	1 (not at risk)	
SAT-10 score: predicted	0 (at risk)	53	15	68
	1 (not at risk)	8	130	138
Total		61	145	206

Source: Authors' analysis of data from the Florida Center for Reading Research; see appendix B for details.

Table C2. Grade 1 CART classification table ($n = 206$)

		SAT 10 score: observed		Total
		0 (at risk)	1 (not at risk)	
SAT-10 score: predicted	0 (at risk)	56	15	71
	1 (not at risk)	5	130	135
Total		61	145	206

Source: Authors' analysis of data from the Florida Center for Reading Research; see appendix B for details.

Table C3. Grade 2 logistic regression classification table ($n = 181$)

		SAT 10 score: observed		Total
		0 (at risk)	1 (not at risk)	
SAT-10 score: predicted	0 (at risk)	47	15	62
	1 (not at risk)	9	110	119
Total		56	125	181

Source: Authors' analysis of data from the Florida Center for Reading Research; see appendix B for details.

Table C4. Grade 2 CART classification table, model 1 ($n = 181$)

		SAT 10 score: observed		Total
		0 (at risk)	1 (not at risk)	
SAT-10 score: predicted	0 (at risk)	39	14	53
	1 (not at risk)	17	111	128
Total		56	125	181

Source: Authors' analysis of data from the Florida Center for Reading Research; see appendix B for details.

Table C5. Grade 2 CART classification table, model 2 (*n* = 181)

		SAT 10 score: observed		Total
		0 (at risk)	1 (not at risk)	
SAT-10 score:	0 (at risk)	46	17	63
predicted	1 (not at risk)	10	108	118
Total		56	125	181

Source: Authors' analysis of data from the Florida Center for Reading Research; see appendix B for details.

Notes

1. The study identified reading test scores that corresponded roughly to the proficiency levels set by the National Assessment of Education Progress (see National Assessment Governing Board, 2012).
2. Surrogate splits provide an alternative to the primary split for those individuals missing a value on the primary split. For example, in figure 3 it may be the case that when these rules are applied in the future, a student may be missing test 3 scores. Splits are provided on the remaining tests that serve as an alternative to test 3 scores. See Therneau and Atkinson (2013) for a detailed discussion of surrogate variables.
3. The default cutscore of .5 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 increases sensitivity and negative predictive power, decreases specificity and positive predictive power, and reduces the overall percentage correct.

References

- American Institutes for Research. (2007). *Reading First state APR data*. Washington, DC: Author. Retrieved January 9, 2014, from <http://www2.ed.gov/programs/readingfirst/state-data/achievement-data.pdf>.
- Beach, K. D., & O'Connor, R. E. (in press). Early response-to-intervention measures and criteria as predictors of reading disability in the beginning of third grade. *Journal of Learning Disabilities*.
- Berk, R. A. (2008). *Statistical learning from a regression perspective*. New York: Springer.
- Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees*. Belmont, CA: Wadsworth International Group.
- Bruce, M., Bridgeland, J. M., Fox, J. H., & Balfanz, R. (2011). *On track for success: The use of early warning indicator and intervention systems to build a grad nation*. Washington, DC: Civic Enterprises. <http://eric.ed.gov/?id=ED526421>
- Carl, B., Richardson, J. T., Cheng, E., Kim, H., & Meyer, R. H. (2013). Theory and application of early warning systems for high school and beyond. *Journal of Education for Students Placed at Risk*, 18(1), 29–49.
- Catts, H. W., Fey, M. E., Zhang, X., & Tomblin, J. B. (2001). Estimating the risk of future reading difficulties in kindergarten children: A research-based model and its clinical implementation. *Language, Speech, and Hearing Services in Schools*, 32(1), 38–50.
- Cohen, J. (1992). A power primer. *Psychological Bulletin*, 112(1), 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>
- Davis, M., Herzog, L., & Legters, N. (2013). Organizing schools to address early warning indicators (EWIs): Common practices and challenges. *Journal of Education for Students Placed at Risk*, 18(1), 84–100.
- 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>
- 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>

- Fuchs, D., Compton, D. L., Fuchs, L. S., Bryant, J., & Davis, G. N. (2008). Making “secondary intervention” work in a three-tier responsiveness-to-intervention model: Findings from the first-grade longitudinal reading study of the National Research Center on Learning Disabilities. *Reading and Writing, 21*(4), 413–436.
- Fuchs, L. S., Fuchs, D., & Hamlett, C. L. (2007). Using curriculum-based measurement to inform reading instruction. *Reading and Writing, 20*(6), 553–567.
- Good III, R. H., Simmons, D. C., & Kame’enui, E. J. (2001). The importance and decision-making utility of a continuum of fluency-based indicators of foundational reading skills for third-grade high-stakes outcomes. *Scientific Studies of Reading, 5*(3), 257–288.
- Gordon, L. (2013). *Using classification and regression trees (CART) in SAS Enterprise Miner for applications in public health* (Paper No. 089–2013). Lexington, KY: University of Kentucky. Retrieved May 21, 2014, from <http://support.sas.com/resources/papers/proceedings13/089–2013.pdf>.
- Harrell, F. E. (2001). *Regression modeling strategies: With applications to linear models, logistic regression, and survival analysis*. New York: Springer.
- Hernandez, D. J. (2011). *Double jeopardy: How third-grade reading skills and poverty influence high school graduation*. Baltimore, MD: Annie E. Casey Foundation.
- Jarvis, S. W., Kovacs, C., Badriyah, T., Briggs, J., Mohammed, M. A., Meredith, P., et al. (2013). Development and validation of a decision tree early warning score based on routine laboratory test results for the discrimination of hospital mortality in emergency medical admissions. *Resuscitation, 84*(11), 1494–1499.
- 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.nrld.org/symposium2003/jenkins/index.html>.
- Johnson, E., & Semmelroth, C. (2010). The predictive validity of the early warning system tool. *NASSP Bulletin, 94*(2), 120–134.
- Knowles, J. E. (2014). *Of needles and haystacks: Building an accurate statewide dropout early warning system in Wisconsin* (draft). Madison, WI: Wisconsin Department of Public Instruction. Retrieved March 13, 2014, from <http://itp.wceruw.org/documents/KnowlesDropoutEarlyWarningSystemforITPJan2013.pdf>.
- Koon, S., Petscher, Y., & Foorman, B. R. (2014). *Identifying at-risk readers in Florida using two methodologies* (REL 2015–036). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Southeast.
- Kuhn, L., Page, K., Ward, J., & Worrall-Carter, L. (in press). The process and utility of classification and regression tree methodology in nursing research. *Journal of Advanced Nursing*.

- Kuhnert, P. M., Do, K. A., & McClure, R. (2000). Combining non-parametric models with logistic regression: An application to motor vehicle injury data. *Computational Statistics & Data Analysis*, 34(3), 371–386.
- Lawrence, R. L., & Wright, A. (2001). Rule-based classification systems using classification and regression tree (CART) analysis. *Photogrammetric engineering and remote sensing*, 67(10), 1137–1142.
- Lemon, S. C., Roy, J., Clark, M. A., Friedmann, P. D., & Rakowski, W. (2003). Classification and regression tree analysis in public health: Methodological review and comparison with logistic regression. *Annals of Behavioral Medicine*, 26(3), 172–181.
- Lewis, R. J. (2000, May). *An introduction to classification and regression tree (CART) analysis*. Paper presented at the Annual Meeting of the Society for Academic Emergency Medicine, San Francisco, CA.
- National Assessment Governing Board. (2012). “NAEP achievement levels.” Retrieved July 2, 2012, from <http://nces.ed.gov/nationsreportcard/achievement.aspx>.
- Neild, R. C., Balfanz, R., & Herzog, L. (2007). An early warning system. *Educational leadership*, 65(2), 28–33.
- 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>
- Piasta, S. B., Petscher, Y., & Justice, L. M. (2012). Diagnostic efficiency of preschool letter-naming benchmarks: Relations with first-grade literacy achievement. *Journal of Educational Psychology*, 104(4), 945–958.
- Rumberger, R., & Lim, S. A. (2008). *Why students drop out of school: A review of 25 years of research* (California Dropout Research Project Report No. 15). University of California, Santa Barbara. Retrieved March 13, 2014, from <http://inpathways.net/researchreport15.pdf>.
- Ryan, M. (2011). *Early Warning Indicator Systems*. Denver, CO: Education Commission of the States. Retrieved March 13, 2014, from <http://edsources.org/wp-content/uploads/9436.pdf>.
- Salford Systems. (n.d.). *What is cross validation?* San Diego, CA: Author. Retrieved February 21, 2014, from <https://www.salford-systems.com/products/cart/faqs/item/114-what-is-cross-validation>.
- 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.

- Shapiro, E., Solari, E., & Petscher. (2008). Use of an assessment of reading comprehension in addition to the oral reading fluency on the state high stakes assessment for students in grades 3 through 5. *Journal on Learning and Individual Differences*, 18, 316–328.
- Steadman, H., Silver, E., Monahan, J., Applebaum, P. S., Clark Robbins, P., Mulvey, E. P., et al. (2000). A classification tree approach to the development of actuarial violence risk assessment tools. *Law and Human Behavior*, 24(1), 83–100.
- Steinberg, D. (2013, February 19). Finding R-squared for CART regression trees. San Diego, CA: Salford Systems. Retrieved February 21, 2014, from http://1.salford-systems.com/blog/bid/270082/?utm_source=linkedin&utm_medium=social&utm_content=c9222a78-8b04-4d58-977e-fd5819277b8a.
- Takahashi, O., Cook, E. F., Nakamura, T., Saito, J., Ikawa, F., & Fukui, T. (2006). Risk stratification for in-hospital mortality in spontaneous intracerebral hemorrhage: A classification and regression tree analysis. *Qjm*, 99(11), 743–750.
- Therneau, T. M., & Atkinson, E. J. (2013). *An introduction to recursive partitioning using the RPART routines* (Mayo Foundation technical report). Washington, DC: 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 4.1–8). Retrieved December 16, 2013, from <http://cran.r-project.org/web/packages/rpart/rpart.pdf>.

The Regional Educational Laboratory Program produces 7 types of reports



Making Connections

Studies of correlational relationships



Making an Impact

Studies of cause and effect



What's Happening

Descriptions of policies, programs, implementation status, or data trends



What's Known

Summaries of previous research



Stated Briefly

Summaries of research findings for specific audiences



Applied Research Methods

Research methods for educational settings



Tools

Help for planning, gathering, analyzing, or reporting data or research