A guide to developing and evaluating a college readiness screener

John Hughes
Yaacov Petscher
Florida State University

A college readiness screener can help colleges and school districts better identify students who are not ready for college credit courses. This guide describes the steps for developing a college readiness screener. For colleges that already have a screener, this guide discusses several issues to consider in evaluating its accuracy.

Why this guide?

Half of all undergraduates take one or more developmental education courses (sometimes called remedial courses), at an average annual cost of $7 billion nationally (Scott-Clayton, Crosta, & Belfield, 2014). The high rate of students taking developmental education courses suggests that many students graduate from high school unready to meet college expectations. Many colleges, particularly two-year institutions, use placement test scores to determine whether a student requires a developmental education course (Hughes & Scott-Clayton, 2011). However, placement tests have been criticized, especially when they serve as the primary or only placement criterion (see, for example, Hodara, Jaggars, & Karp, 2012; Scott-Clayton et al., 2014). To improve placement accuracy, colleges that currently rely solely on placement test scores may wish to consider a broader screening tool that incorporates other student information.

This guide describes core ideas for colleges to consider when developing a screener for estimating college readiness. A key focal point is a discussion of ways to improve how well a screener identifies individuals who need developmental education, along with key considerations for a user or developer of such a tool. Specifically, the guide includes seven steps:

1. Creating a definition of college readiness.
2. Selecting a measure of readiness.

3. Identifying potential predictors of college readiness.
4. Prioritizing types of classification error.
5. Collecting and organizing the necessary data.
6. Developing predictive models.
7. Evaluating the screening results and selecting the final model.

Although most two-year colleges use placement tests as their only screener (Parsad & Lewis, 2004), some do not. For example, in Florida recent high school graduates who attend a two-year college cannot be required to take a placement test. Those colleges must assess readiness in an alternative way.

The primary audience for this guide is leaders and staff at colleges that are seeking to revise or evaluate their screening methods or to develop a college readiness screener if they do not already have one. Two interrelated groups are commonly involved in developing a screener. College leaders, particularly in academic affairs and student support services, create the institution's definition of college readiness (step 1), select a measure for the definition of readiness (step 2), and prioritize the types of classification error (step 4). Institutional researchers or similar college staff identify potential predictors of college readiness (step 3), collect and organize the necessary data (step 5), and develop predictive models (step 6). College staff may do the analytic work (steps 5 and 6), or they may seek outside assistance with the main analyses to inform their decisions. Both groups are involved in evaluating the screening results and selecting the final model (step 7).

The examples in this report use a placement test score as the single predictor of college readiness. This reflects the most common approach nationally and simplifies interpretation of the examples. However, as steps 2 and 5 indicate, colleges should consider multiple predictors of readiness in developing and evaluating their placement process.

Step 1: Creating a definition of college readiness

The first step for a college is to define college (and career) readiness. Readiness is most often defined as being prepared to succeed in college (that is, eventually graduate). For example, a recent guide from the National Forum on Education Statistics proposed the following conceptual definition:

> A student is college and career ready when he or she has attained the knowledge, skills, and disposition needed to succeed in credit-bearing (non-remedial) postsecondary coursework or a workforce training program in order to earn the credentials necessary to qualify for a meaningful career aligned to his or her goals and offering a competitive salary (National Forum on Education Statistics, 2015, p. 1).

An operational definition takes a general concept and creates a definition that can be measured systematically and consistently. In this case an operational definition of college readiness allows colleges to establish a measurable basis for determining whether a student is prepared for college credit courses and is likely to graduate. However, that definition often relies on activities in which the student has not yet engaged. For example, a high school student entering college typically has not yet earned credits for college coursework. As such, it is impossible to know whether the student was ready until after he or she has taken a course. Because many factors influence graduation rates, such as support services, financial aid policies, and student nonacademic factors, it is challenging to use graduation as part of an operational definition of readiness. Instead, colleges need to select a measurable definition of readiness that is close in time to college entry but still reflects a key milestone on the path toward graduation.

One such definition focuses on expected level of performance in one or more specific courses, usually gateway courses. Gateway courses are foundational, meaning they serve as an entry point into a major or other courses. Introductory courses for majors are not suitable because they are often taken later in college. Instead, colleges typically focus on gateway courses that have high enrollments, are usually taken in the first few semesters, and serve as prerequisites for other courses within the institution. Introductory English and math courses are common gateway courses. As a result, colleges will typically have at least two operational definitions of readiness, one for English courses and one for math courses.
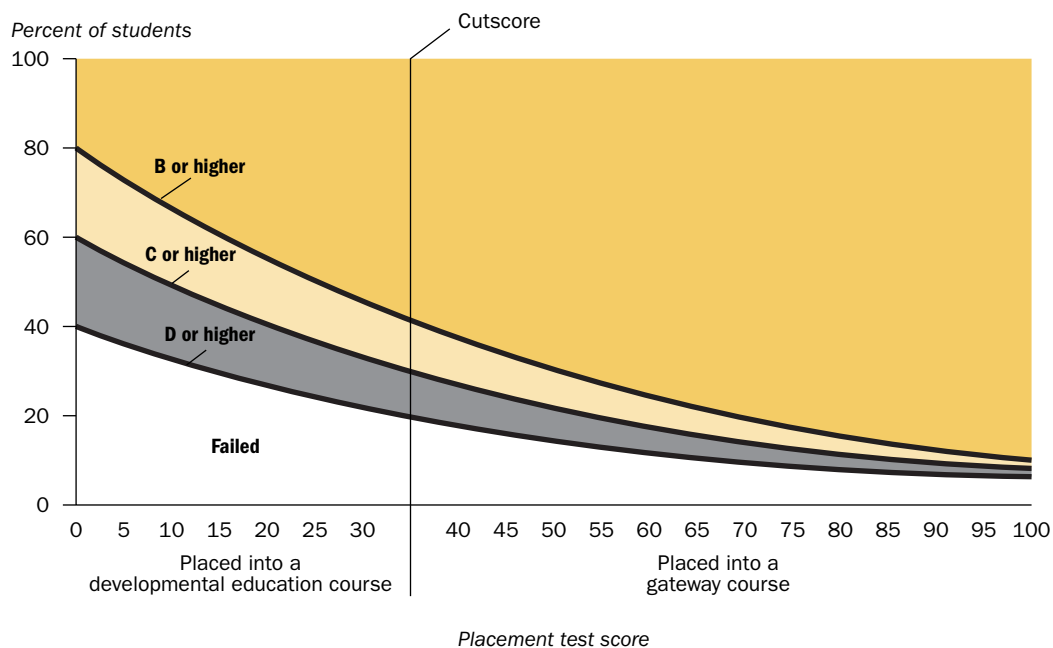
Step 2: Selecting a measure of readiness

An important consideration after defining readiness is how to measure readiness. Thus, the second step is to take the operational definition, convert it into something that can be directly measured, and select a target level to represent readiness. Given the need to accurately place students into developmental education or gateway courses, readiness for success is often operationalized as the probability of success in a gateway course. This is the approach used by major college placement exams and for the examples in this report.

However, that definition needs more precision. What exactly constitutes success in a gateway course? Is it a grade of D or higher, C or higher, or something else? A grade of D is often sufficient to pass a course but may not be an acceptable standard for being college ready. This question is critical because the grade that defines success directly relates to the probability that a student will be successful—an intuitive but critical point. For example, defining success as achieving a D or higher in a gateway course or as achieving a C or higher substantively changes the number of students placed into a developmental education course.

This can be shown by charting the probability of success in a course against the placement test score. Figure 1 depicts three outcome options—earning a B or higher, earning a C or higher, and earning a D or higher. The area above each line represents the students who would earn the respective grade or higher, and the area below each line represents the students who would earn a lower grade. Across the bottom is the score on a placement test. At the left side, at a score of 0, about 20 percent of students would earn a B or higher in a gateway course, and 80 percent would earn below a B; about 40 percent of students would earn a C or higher; and about 60 percent of students would earn a D or higher. At the far right of the figure, about 90 percent of students with a score 100 on the placement test would earn a B or higher, and slightly more than that would earn a C or higher or a D or higher. Essentially, almost all students who score 100 on the placement test would earn a B or higher.

Changing the definition of success can greatly affect placement rates. Defining success as a B or higher is much different from defining success as a D or higher. Using the placement test from figure 1 and a cutscore of 35 means that anyone who scores at or above 35 is placed into a gateway course and anyone who scores below 35 is placed into a developmental education course. Among students in figure 1 who score at the cutscore, about 59 percent would be expected to earn a B or higher and about 80 percent would be expected to earn a D or higher. Thus, changing the definition of success from a grade of B or higher to a grade of D or higher increases the percentage of students expected to be successful at that cutscore by 21 percentage points. This change directly affects placement: fewer students would be placed into a developmental education course because more would be expected to be successful in a gateway course (see step 4 for a discussion of the tradeoffs associated with higher and lower definitions of success).

Figure 1. Hypothetical relationship between placement scores and earning a specific grade in a college gateway course

Percent of students

Cutscore

B or higher

C or higher

D or higher

Failed

Placed into a
developmental education course

Placed into a
gateway course

Placement test score

Source: Adapted from Scott-Clayton (2012).

Step 3: Identifying potential predictors of college readiness

The third step is to identify key predictors of the outcome chosen in step 2. Most colleges, particularly open-access institutions, use placement test scores, with COMPASS® and ACCUPLACER® the most common (Hughes & Scott-Clayton, 2011).[1] Placement tests have the advantage of providing quick, objective, and consistent measures for predicting college readiness. Their ubiquitous use often leads to scores from their being considered synonymous with measures of college readiness.

However, relying on a single test score raises several concerns. Researchers have found that students may not understand the purpose of the placement test or take it seriously, resulting in artificially low scores, and that some students are not prepared for the format of the exam, which also could reduce scores (Venezia, Bracco, & Nodine, 2010). Students who score lower on the placement test because of those kinds of factors run the risk of being misclassified (see step 4 for more detail). Indeed, some research suggests that incorporating more than a placement test score could improve placement accuracy (Belfield & Crosta, 2012; Hodara et al., 2012; Johnson, Jenkins, & Petscher, 2010). For example, combining high school grade point average and placement test score has been found to reduce error rates and improve placement rates. Including other high school variables, such as number of honors and college-level courses, may also provide marginal improvement in placement accuracy (Belfield & Crosta, 2012).

Moreover, failure in college may be due to more than academic readiness. Noncognitive factors such as study habits, confidence, and resilience may play a key role, along with social and financial support (Hodara et al., 2012). Those factors can be difficult to measure directly, but high school transcripts may offer indirect evidence that could be predictive of college success.

For all these reasons colleges should at least consider incorporating high school grades and other outcomes from high school transcripts to improve placement results. In particular, Scott-Clayton (2012) found that high school grade point average was as good as or better than placement tests and that other predictors, such as college preparatory credits earned in high school and time since high school graduation, can incrementally improve placement accuracy. While multiple predictors for readiness is uncommon among colleges (Hughes & Scott-Clayton, 2011), some research indicates that using more than one predictor could improve classification precision (Belfield & Crosta, 2012).

Colleges could use high school transcript data to develop and test other potential predictors, such as:
- Grades for selected courses such as college preparatory courses or subject-specific courses.
- Total credit accumulation in specific subjects or in college preparatory courses.
- Number of courses failed or ratio of credits attempted to credits earned.
- Timing of key courses, such as when a student took Algebra 1.
- End-of-course exam scores.
- End-of-grade exam scores.

From this list or from other sources colleges can select a set of potential predictors for which data are readily available. Most colleges will have access to high school transcript data, but not all will, and what is available may vary. Moreover, although some predictors, such as high school grades, are intuitively obvious, others might not be. The key is to identify as many potential predictors as possible to test in order to maximize the accuracy of the screener. The optimal mix of predictors will vary from college to college; thus, colleges should identify and ultimately test a range of predictors (see steps 6 and 7).

Step 4: Prioritizing types of classification error

The fourth step is to prioritize the classification error. No matter how good the predictors, all screeners are subject to classification error. At its simplest, classification error refers to whether a student is correctly placed (Sawyer, 1996; Schatschneider, Petscher, & Williams, 2008). In the context of college readiness, accurate placement means not placing college-ready students into a developmental education course and not placing students who are not college ready into a gateway course.

A two-by-two table can illustrate classification error types (table 1). The columns represent whether a student was actually ready for college, and the rows represent whether the student was placed into a developmental education course or a gateway course. Each student thus falls into one of four groups.

Two groups of students—those in cells A and D—were correctly classified and placed. The students in cell A scored below the college's cutscore and were truly not college ready; they were correctly placed into a developmental education course. The students in cell D scored above the cutscore and were truly college ready; they were correctly placed into a gateway course.

Table 1. Two-by-two table classification table for college readiness

		Actual readiness	
		Not college ready	College ready
Screen	Scored below the cutscore (placed into a developmental education course)	A (true positive)	B (false positive)
	Scored above the cutscore (placed into a gateway course)	C (false negative)	D (true negative)

Source: Adapted from Schatschneider et al. (2008).

The students in cells C and B were misclassified and misplaced. The students in cell C scored above the college's cutscore but were not truly college ready; they were placed into a gateway course even though they were not college ready. The students in cell B scored below the cutscore but were truly college ready; they were placed into a developmental education course even though they were ready for a gateway course.

The overall classification accuracy can be calculated using table 1 (Schatschneider et al., 2008). Cells A and D are considered correct screening classifications (or placements), while cells B and C are considered incorrect. The total number of students in cells A and D divided by the total number of students in all cells provides the overall accuracy rate. The total number of students in cells B and C divided by the total number of students in all cells provides the overall error rate.

Although increasing the accuracy rate and decreasing the error rate is always important for a screener, considering the types of errors is also important. The students in cells B and C both represent misclassifications, but they are different kinds of misclassifications—specifically, overplacement and underplacement (Scott-Clayton, 2012):

- The potential for a student who is not college ready to be placed into a gateway course and to ultimately fail can be considered overplacement because the student is placed above the level of coursework in which he or she could be successful (cell C).
- The potential for a student who is college ready to be placed into a developmental education course, resulting in wasted time and money, can be considered underplacement because the student is placed below the level of coursework in which he or she could be successful (cell B).
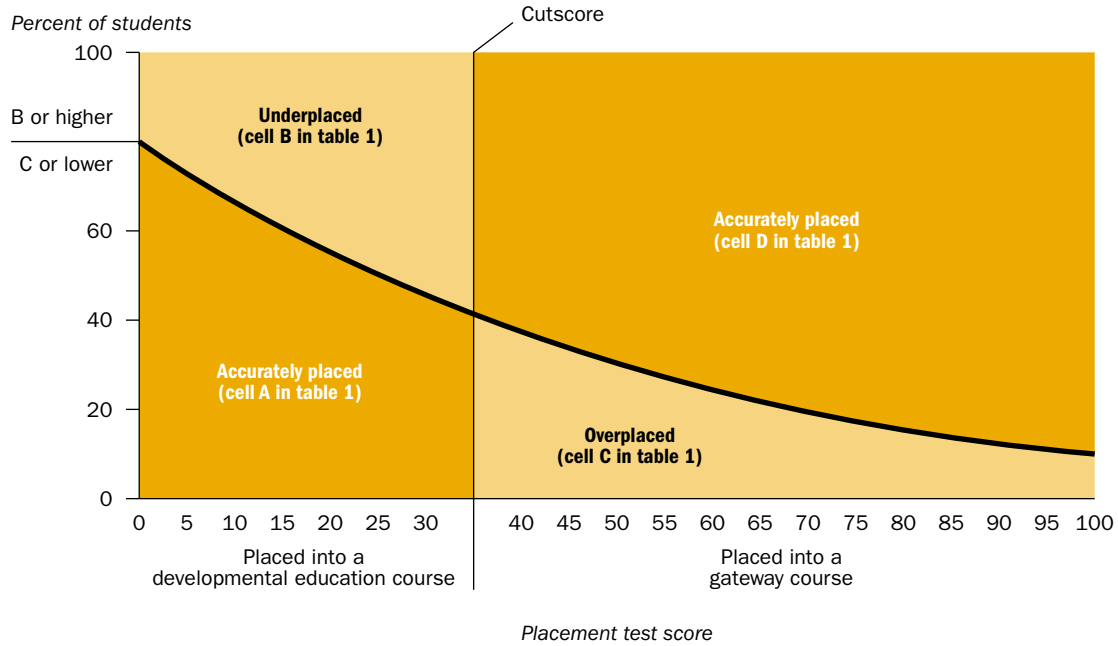
Thus, overplacement and underplacement reflect two different types of errors, and colleges need to take both into account when screening students for college readiness.

Overplacement and underplacement can be visualized using figure 2, which combines the information from table 1 and figure 1. In this example the expected outcome is a grade of B or higher. The lower left and upper right quadrants represent accurate placements, and the other two quadrants represent misclassifications.

- The lower left quadrant reflects students who scored below the cutscore and who would not have earned a B or higher. These students would be accurately placed into a developmental education course.
- The upper right quadrant represents students who scored above the cutscore and earned a B or higher. They would be accurately placed into a gateway course.
- The lower right quadrant reflects overplacements—students who scored above the cutscore but who would not have earned a B or higher and would be incorrectly placed into a gateway course.
- The upper left quadrant reflects underplacements—students who scored below the cutscore and would have earned a B in a gateway course but would be incorrectly placed into a developmental education course. In this example 20 percent of those scoring 0 on the placement test would have gone on to earn a B.
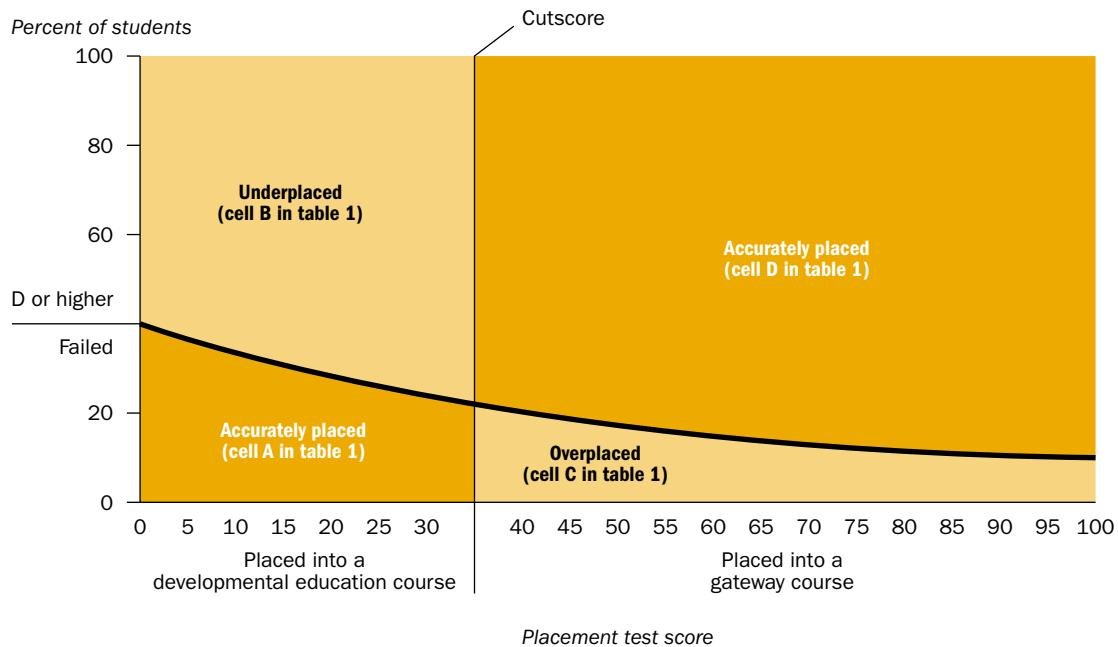
There is a direct tradeoff between overplacement and underplacement: reducing one will increase the other. As a result, minimizing error is partly a process of selecting which type of error to minimize. Compare figure 2 with figure 3. Figure 3 uses an expected grade of D or higher as the definition of success. This results in fewer overplacements and more underplacements (if the cutscore is not adjusted). The lower expectation for performance in a gateway course decreases the potential error when placing students into a gateway course but increases the potential errors when placing students into a developmental course. Similarly, having a higher expected outcome, as shown in figure 2, increases the potential error when placing students into a gateway course but reduces the error when placing students into a developmental course.

Figure 2. Classification accuracy based on an expected grade of B or higher in a gateway course

Percent of students

Cutscore

Underplaced
(cell B in table 1)

Accurately placed
(cell D in table 1)

Accurately placed
(cell A in table 1)

Overplaced
(cell C in table 1)

B or higher
C or lower

100
60
40
20
0

0 5 10 15 20 25 30 40 45 50 55 60 65 70 75 80 85 90 95 100

Placed into a
developmental education course

Placed into a
gateway course

Placement test score

Figure 3. Classification accuracy based on an expected grade of D or higher in a gateway course

Percent of students

Cutscore

Underplaced
(cell B in table 1)

Accurately placed
(cell D in table 1)

Accurately placed
(cell A in table 1)

Overplaced
(cell C in table 1)

D or higher
Failed

100
80
60
20
0

0 5 10 15 20 25 30 40 45 50 55 60 65 70 75 80 85 90 95 100

Placed into a
developmental education course

Placed into a
gateway course

Placement test score

The same principle holds for changing the cutscore. Setting a higher or lower cutscore will increase the chance of one type of error while decreasing the chance of the other. This can be visualized by imagining moving the line to the left or the right. For example, moving it to the right (raising the cutscore) will make the overplacement section smaller and the underplacement section larger.

Policymakers must decide whether overplacement or underplacement poses a larger cost for the institution and the students. In this case, costs broadly encompass the actual costs of courses as well as the time lost when students spend a semester in the wrong course and the psychological costs of taking a course that may be too hard or too easy. The associated costs differ for each type of error. Whether the costs of taking an unneeded developmental course are higher or lower than the costs of failing a course that a student was not prepared for is a question for policymakers at each institution, who must determine which type of error is more consequential or less consequential for their students.

The costs depend on the chosen definition of college readiness. Placing a college-ready student into a developmental education course (underplacement) has a higher opportunity cost for a student who would have earned an A or B than for a student who would have earned a C or D. For example, placing a student into a development education course has a higher cost for a student who was likely to earn a B than for a student who was likely to earn a D (who would have passed but who now might earn a higher grade). The reverse is also true. Overplacing a student into a gateway course has more consequences for a student who fails that course than for one who would earn a D and thus still be able to move on to other college credit courses. Thus, having a higher or lower definition of readiness affects the potential costs and should be part of evaluating the predictive models discussed later in this report.

Step 5: Collecting and organizing the necessary data

In the fifth step the college will need data on the outcome measure and the various predictors. Actions to consider when collecting and organizing the data include:
- Collecting the grades for each identified gateway course for at least a semester—ideally, for multiple semesters.
- Collecting the section identification number for each course and linking it to the instructor.
- For each student with grades from a gateway course, collecting the data for each predictor.[2]

Next, the data need to be organized. For the kinds of analyses described in this guide, the data typically need to be organized as one record per student per outcome. For each student there should be one line of data that contains the predictors to be used in the model (table 2). This would be repeated for each gateway course or subject. In this example the outcome is the student's grade in an introductory algebra course. The predictors are high school grade point average, placement test score, and grades in high school Algebra I and II. Any number of additional predictors of interest could be used, based on local interest and data availability.

Table 2. Hypothetical example of student records organized for analysis

Student	Gateway course	Gateway grade	Course section	Math placement test score	High school grade point average	High school Algebra II grade	High school Algebra I grade
000001	Intro to Algebra	B	0123	46	3.32	B–	A
000002	Intro to Algebra	F	0123	32	2.10	Missing	C
000003	Intro to Algebra	C	9876	35	2.88	C	C

Source: Authors' example.

Step 6: Developing predictive models

The sixth step involves developing the actual predictive models. This section is not intended as a tutorial on these methods; see appendix A for links to resources that elaborate on the methodologies discussed here.

Developing predictive models can take several forms. This guide highlights two methods. Both use predictors to classify success, but the methods are distinguished by their form and complexity. The first approach is logistic regression,[3] and the second is a classification and regression tree (CART) analysis (Koon & Petscher, 2015).

Logistic regression

The most common approach in research studies and screening and placement is some form of logistic regression. This is a statistical approach designed to predict the probability of a given outcome. Logistic regression uses multiple predictors (variables) to estimate the probability of a given outcome. In this case a college might use several predictive variables to estimate the probability of a given student earning a B or better in Introductory English.

Logistic regression is the most common approach to modeling college readiness (see, for example, Belfield & Crosta, 2012, and Scott-Clayton, 2012) because college readiness is a binary outcome—ready or not ready—and logistic regression is designed for use with that type of outcome. Logistic regression can produce an individual-specific score, which can be converted to a predicted probability of college readiness based on a given set of predictors. See appendix A for details on the steps in a logistic regression analysis.

Classification and regression tree analysis

CART analysis offers a second approach to classification, one that is comparable to logistic regression but with results that often are easier to interpret (Koon & Petscher, 2015; Koon, Petscher, & Foorman, 2014). CART analysis also classifies students based on a given outcome, but it does so using a set of if–then statements instead of statistical coefficients. For example, if the student is above this cutscore, he or she is college ready; if the student is below this cutscore, he or she is not college ready. The CART model searches for the best way to split the sample into ready and nonready groups, based on the available predictors.

The resulting CART shows each predictor and cutscore. A hypothetical CART for college readiness might start with high school grade point average. Students with a grade point average of 3.8 or higher would be considered college ready (figure 4). For students with a grade point average lower than 3.8, the CART then considers their placement score; students with a placement score of 35 or higher would be considered college ready. Students with a grade point average lower than 3.8 and a placement score lower than 35 would be considered not college ready. See appendix A for details on the steps involved in CART analysis.

Colleges may consider running separate models for different semesters. For most colleges a large majority of new students enroll during the fall semester. Spring enrollments are not only smaller but potentially reflect a different type of student. As a result, a model that works for fall enrollees might not work as well for spring enrollees.

Figure 4. Hypothetical classification and regression tree analysis for college readiness
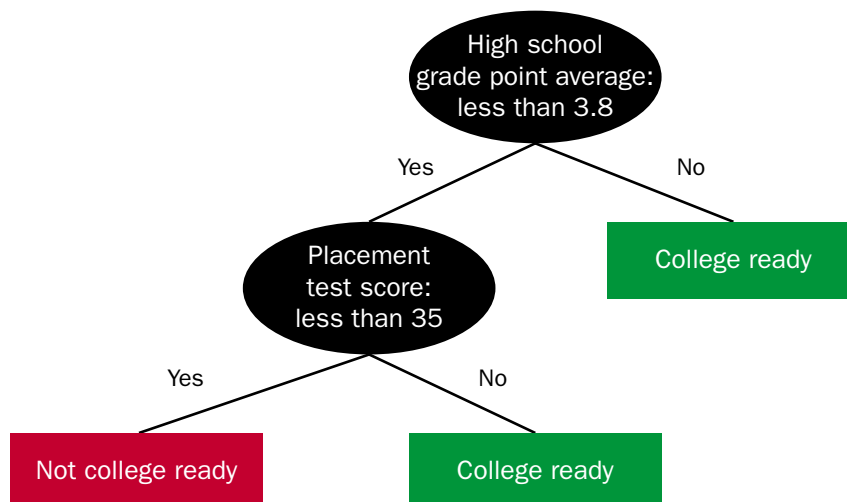
Step 7: Evaluating the screening results and selecting the final model

The seventh step is to evaluate the results of the screener. Colleges have several ways to evaluate the classification accuracy of the models tested (Schatschneider et al., 2008). The most basic measure is the overall classification accuracy. The overall accuracy could be calculated for different models, and the one with the highest accuracy would then be selected. However, this method ignores the risks associated with different types of errors. The college must decide whether it wants to minimize overplacement or underplacement or whether to balance them.

Once that is determined, the college can calculate key measures of screening accuracy—overall accuracy rate, specificity, and sensitivity:

- Overall accuracy rate = (total number accurately placed)/(total number) or (A+D)/A+B+C+D) using table 1.
- Specificity is the proportion of true negatives—the number of students who were accurately determined to be college ready divided by the number of students predicted to be college ready [D/(B+D) from table 1].
- Sensitivity is the proportion of true positives—the number of students accurately determined to be not college ready divided by the number of students predicted to be not college ready [A/(A+C) from table 1].

Specificity and sensitivity are common measures of screening accuracy. They indicate the proportion of placements that are accurate or correct.

Specificity

Specificity is a measure of the proportion of students who are ready for college and gateway courses and who are accurately placed by the screener. This is the inverse of the underplacement rate, which is the proportion of students who are college ready and who are placed into developmental education courses [D/(B+D) from table 1]. Increasing the specificity decreases the underplacement error rate, resulting in fewer college-ready students being placed into developmental education courses.

Sensitivity is a measure of the proportion of students who are not ready for college and who are accurately placed by the screener into a developmental education course. This is the inverse of the overplacement rate, which equals the proportion of students who are not ready for college and who are placed into a gateway course [C/(A+C) from table 1]. Increasing the sensitivity decreases the overplacement error rate. This means that an increase in the sensitivity results in fewer students who are not college ready being placed into a gateway course.

Combined, specificity and sensitivity can provide diagnostic information about the accuracy of the screening models. Research has used a generally established accuracy threshold of .80–.90 for each of these measures (Piasta, Petscher, & Justice, 2012). That equates to overplacement and underplacement error rates of .10–.20. Within that range colleges should seek to minimize the type of error deemed most problematic based on the nature of the target outcome and the institution's priorities. The college can use classification accuracy to select the final model for screening students.

Diagnostic analysis using receiver operating characteristic curves

Receiver operating characteristic (ROC) curves are a common diagnostic test for examining the fit of a screener (Petscher & Kim, 2011a, 2011b). A ROC curve plots the true positive rate (sensitivity) against the false positive rate (specificity). The shape of the graph provides a visual indication of how much better (or worse) the screener does compared with guessing or simply placing a student randomly. See appendix A for more details on ROC curve analyses.

Concluding considerations

College readiness screeners currently in use tend to focus on a single placement score. The research described in this guide suggests that using multiple predictors of college readiness can improve screening results. In addition, all screeners have implicit assumptions about the definition of college readiness and the appropriate tradeoffs between overplacements and underplacements. Colleges that lack a screening process may wish to develop one, and those that have one may want to evaluate its accuracy. This guide can be used for both purposes and can help colleges understand and address the various challenges and tradeoffs associated with developing and evaluating a screener.

Developing a screener is a process of selecting tradeoffs between accuracy and simplicity. One hundred percent accuracy is never possible, but additional data or analyses almost always ensure greater accuracy. For example, a logistic regression almost always conveys some piece of data that could improve the predictive capability of the model, but the expense and effort of collecting those data might outweigh the benefits. Similarly, in a CART model, adding another branch will almost always improve accuracy. However, at some point the tree becomes too difficult to interpret and may not generalize to new populations, such as the next year's freshman class. Colleges will need to determine the most parsimonious solution, balancing accuracy with simplicity of interpretation and generalizability.

Developing a screener is also an iterative process and ongoing process. In the simplest form this can mean comparing actual outcomes to predictions. But it also means re-evaluating the models on a regular basis. The nature of the students who enroll can shift over time, particularly if there are changes in K–12 policy around high school graduation requirements. Models developed in the past might not work the same in the future.

Appendix A. Methodological guidance and resources

This guide outlines two basic screening methodologies—logistic regression and classification and regression tree (CART) analysis. This appendix offers more detail on the basic processes used in those methodologies.

Logistic regression

Logistic regression is commonly used when the outcome of interest is binary. When only two outcomes are possible, such as ready or not ready for college, other analytic techniques, such as ordinary least squares regression, are less appropriate. Although a detailed explanation about how to use logistic regression is beyond the scope of this guide, these are the basic steps after data screening:

- Create a binary variable to define college success. For example, a student with a gateway course grade of at least B could be coded as 1, and a student with a grade lower than B could be coded as 0.
- Select the predictor variables and the college readiness measure. For example, a model might include placement test scores, high school grade point average, and math end-of-course exam scores.
- Run the initial logistic regression model with all the variables.
- Evaluate model fit using tests such as the chi-square goodness of fit, Hosmer-Lemeshow, classification table, and pseudo-R^2 statistic.
- Remove predictors that are not statistically significant, taking into account the fit of the model—that is, remove one or more predictor variables, re-run the model, and check the model fit. Any variables whose removal notably reduces model fit should be kept (put back in the model).
- Repeat until an optimal model is identified based on the desired classification accuracy and error rates.
- Use the model to generate a predicted score for each student in the dataset.
- Convert the predicted score, typically a log-odds value, to a predicted probability score.

Once those basic steps are completed, the predicted probability score can be used to create a table that displays the probability value of college success for each possible score (or combination of scores) of the predictors (Koon & Petscher, 2015). That can be done for several different models (or combinations of predictors), and the results can be evaluated to determine the best all-around model (see below for discussion of evaluating models).

Classification and regression tree analysis

CART analysis can provide an alternative to logistic regression (Koon & Petscher, 2015; Koon, Petscher, & Foorman, 2014). Like logistic regression, a CART analysis classifies students based on a given outcome, but it does so using a set of if–then statements instead of statistical coefficients. Each if–then statement creates a branching tree that splits students into groups, such as ready or not ready for college. The splits can be based on data, such as test scores and grades. The analysis identifies the optimal variables and splits.

However, a CART analysis with too many branches can become too complex to interpret easily and may no longer be generalizable. A CART with too many branches becomes specific to the population of students in the given dataset, and the results may not generalize to a different group of students. Models must be pruned to maximize accuracy while minimizing complexity and error rates. The following steps outline the process for conducting a CART analysis:

- Select the predictor variables and the college readiness measure. For example, a model might include placement test scores, high school grade point average, and math end-of-course exam scores.

- Select appropriate stopping rules to determine how many splits will be made. In theory, a CART analysis can create as many splits as needed until every individual in the dataset is correctly predicted. Having more splits increases accuracy but also increases complexity and reduces generalizability to different student populations, such as a different incoming class.
- Select an appropriate number of cross-validations that use subsets of the data to test the fit for a different sample of students. A cross-validation splits the full sample of students into smaller subsets, develops a classification tree for one sample, and then tests it on a different sample.
- Choose the default or desired weighting of errors.
- Review the initial model and prune the tree by choosing an appropriate complexity parameter. In this step a classification tree is reviewed, and some splits may be pruned to maximize the overall accuracy while minimizing the highest risk error and the total number of branches.

Receiver operating characteristic curve analyses

A receiver operating characteristic (ROC) curve plots the true positive rate (sensitivity) against the false positive rate (specificity). The shape of the resulting curve provides a visual indication of the improvement provided by a screening instrument. A simple diagonal line represents what would happen with guessing or completely random placement (figure A1). That placement will be correct 50 percent of the time (represented by the shaded area below the line), so the true positive rate and the false positive rate are equal. By contrast, if a screener operates perfectly, the true positive rate will be 100 percent and the false positive rate will be 0 percent, resulting in a straight line along both axes (figure A2).
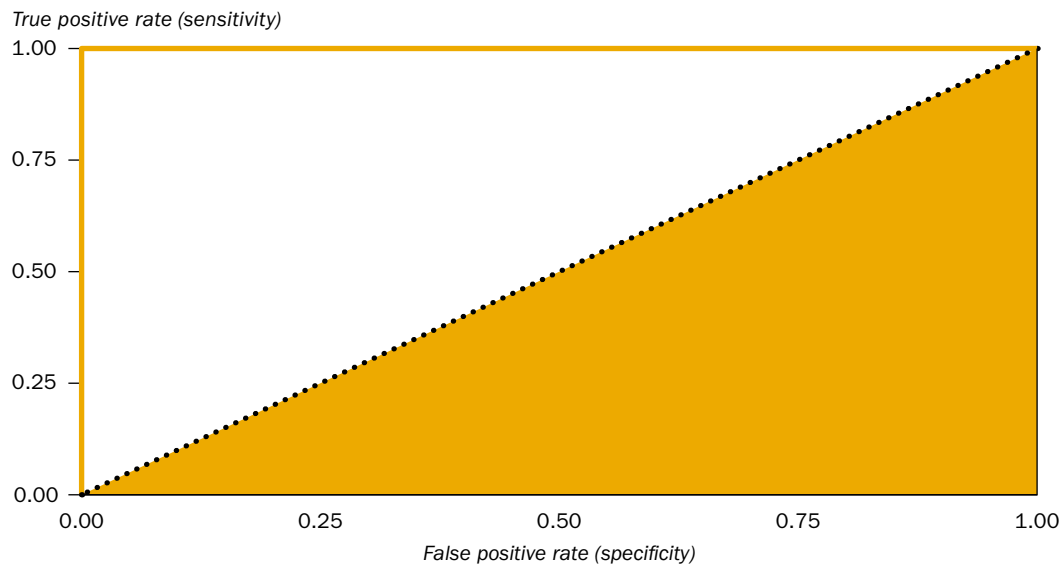
In reality, a ROC analysis will yield a hump-shaped curve. The size and shape of the hump show the improvement over the random-guessing model. The area below the diagonal line represents the results of guessing, whereas the area above that line but below the ROC curve shows the improvement offered by the screener (the area above the dotted line in figure A3). By plotting the results for different models, ROC curve analyses can show how much improvement a screener offers and how the results of different screeners compare.

Figure A1. Receiver operating characteristic curve analysis with random screener assignment
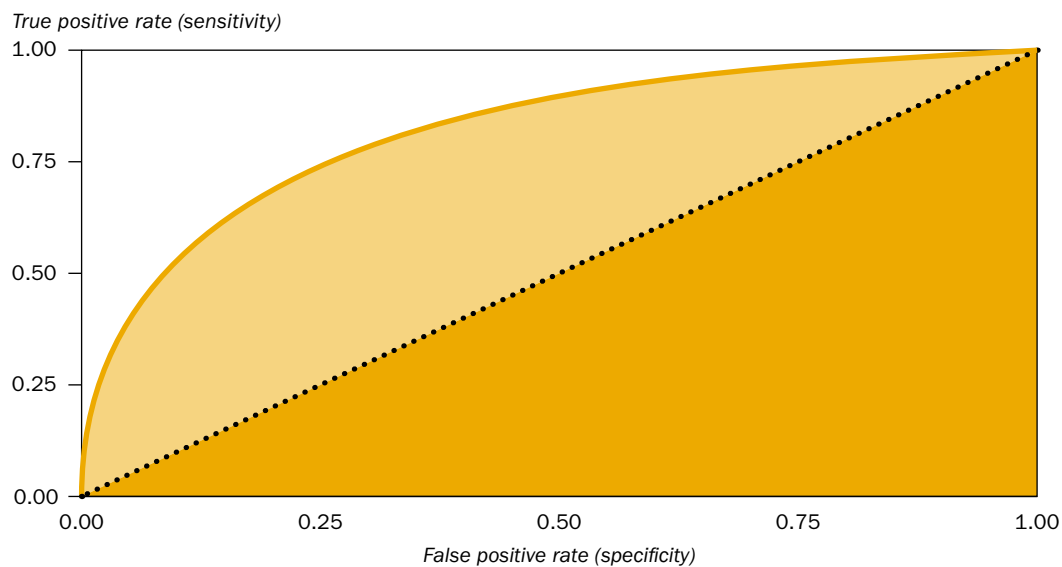

True positive rate (sensitivity)

False positive rate (specificity)
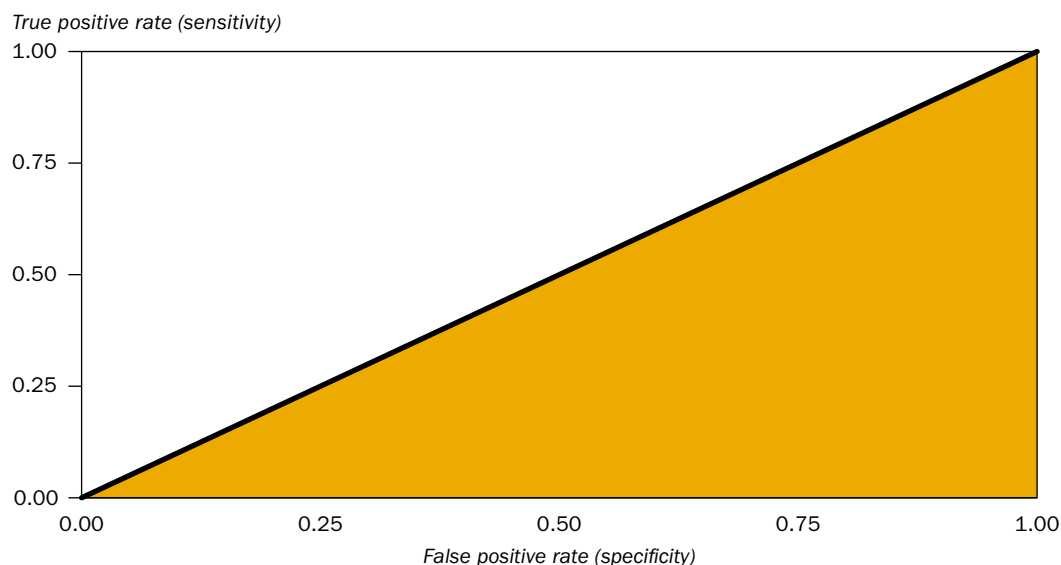
Source: Authors' example.

Figure A2. Receiver operating characteristic curve analysis with a perfect fit

True positive rate (sensitivity)

False positive rate (specificity)

Figure A3. Receiver operating characteristic curve analysis example fit

True positive rate (sensitivity)

False positive rate (specificity)

Additional resources

The following are online resources that offer more detailed directions for using some of the methodologies.

Learning logistic regression via R:
- http://ww2.coastal.edu/kingw/statistics/R-tutorials/logistic.html
- http://www.r-tutor.com/elementary-statistics/logistic-regression

Learning logistic regression via SAS:
- http://www.ats.ucla.edu/stat/sas/dae/logit.htm

Learning CART via R:
- http://www.stat.cmu.edu/~cshalizi/350/lectures/22/lecture-22.pdf

Learning CART via SAS:
- http://support.sas.com/resources/papers/proceedings13/089–2013.pdf

Learning Receiver operating characteristic (ROC) curve via SAS:
- http://www2.sas.com/proceedings/sugi31/210–31.pdf

Learning ROC curve via R:
- http://blog.yhathq.com/posts/roc-curves.html

Notes

1. The COMPASS placement test is produced by ACT. The ACCUPLACER is produced by the College Board. Both are computerized multiple choice tests that cover English and math.
2. Colleges will have to decide how to handle withdrawals. Those students can be excluded from the analyses or included if the college wants to treat withdrawals as a form of failure.
3. Many related statistical methods are designed for binary and categorical outcomes and other nonlinear models, but for the sake of simplicity this report refers to them all as logistic regression.

References

Belfield, C. R., & Crosta, P. M. (2012). *Predicting success in college: The importance of placement tests and high school transcripts* (CCRC Working Paper No. 42). New York: Columbia University, Teachers College, Community College Research Center. http://eric.ed.gov/?id=ED529827

Hodara, M., Jaggars, S. S., & Karp, M. M. (2012). *Improving developmental education assessment and placement: Lessons from community colleges across the country* (CCRC Working Paper No. 51). New York: Columbia University, Teachers College, Community College Research Center. http://eric.ed.gov/?id=ED537433

Hughes, K. L., & Scott-Clayton, J. (2011). Assessing developmental assessments in community colleges. *Community College Review, 39*(4), 327–351. http://eric.ed.gov/?id=ED516079

Johnson, E. S., Jenkins, J. R., & Petscher, Y. (2010). Improving the accuracy of a direct route screening process. *Assessment for Effective Intervention, 35*(3), 131–140. http://eric.ed.gov/?id=EJ883312

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. http://eric.ed.gov/?id=ED545225

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. http://eric.ed.gov/?id=ED554441

National Forum on Education Statistics. (2015). *Forum guide to college and career ready data* (NFES No. 2015–157). Washington, DC: U.S. Department of Education, National Center for Education Statistics. Retrieved August 20, 2015, from https://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2015157.

Parsad, B., & Lewis, L. (2004). *Remedial education at degree-granting postsecondary institutions in Fall 2000* (NCES No. 2004–010). Washington, DC: U. S. Department of Education, National Center for Education Statistics. http://eric.ed.gov/?id=ED482370

Petscher, Y., & Kim, Y. S. (2011a). Efficiency of predicting risk in word reading using fewer, easier letter names. *Assessment for Effective Intervention, 36*(1), 256–266. http://eric.ed.gov/?id=EJ951739

Petscher, Y., & Kim, Y. S. (2011b). The utility and accuracy of oral reading fluency score types in predicting reading comprehension. *Journal of School Psychology, 49*(1), 107–129. http://eric.ed.gov/?id=EJ911351

Piasta, S. B., Petscher, Y., & Justice, L. M. (2012). How many letters should preschoolers in public programs know? The diagnostic efficiency of various preschool letter-naming benchmarks for predicting first-grade literacy achievement. *Journal of Educational Psychology, 104*(4), 945–958. http://eric.ed.gov/?id=EJ994042

Sawyer, R. (1996). Decision theory models for validating course placement tests. *Journal of Educational Measurement, 33*(3), 271–290. http://eric.ed.gov/?id=EJ535136

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. M. Justice & C. Vukelich (eds.), *Achieving*

Excellence in Preschool Literacy Instruction (pp. 304–316). New York: The Guilford Press. http://eric.ed.gov/?id=ED498268

Scott-Clayton, J. (2012). *Do high-stakes placement exams predict college success?* (CCRC Working Paper No. 41). New York, NY: Columbia University, Teachers College, Community College Research Center. http://eric.ed.gov/?id=ED529866

Scott-Clayton, J., Crosta, P. M., & Belfield, C. R. (2014). Improving the targeting of treatment: Evidence from college remediation. *Educational Evaluation and Policy Analysis, 36*(3) 371–393. http://eric.ed.gov/?id=EJ1042032

Venezia, A., Bracco, K. R., & Nodine, T. (2010). *One-shot deal? Students' perceptions of assessments and course placement in California's community colleges.* San Francisco, CA: WestEd. Retrieved August 20, 2015, from http://www.wested.org/resources/one-shot-deal-students-perceptions-of-assessment-and-course-placement-in-californias-community-colleges/.

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