# Determining the Success of a Test for Program Admission[1]

A. Alexander Beaujean
Alex_Beaujean@baylor.edu
Baylor Psychometric Laboratory (Baylor University)

Report BPL–2014-ACAD-ENROLL-1
Last updated: January 23, 2014

## Contents

[1] Prepared for January 2014 meeting of Baylor University's Task Force on Academic Program Enrollment.

# 1. Introduction

Making a decision to admit a student into a program is a binary prediction of how well the individual student will do in the program. By admitting the student, the decision makers believe there is enough evidence to predict that the student will graduate from the program. Conversely, by rejecting the student, the decision makers believe there is enough evidence to predict that the student will **not** graduate from the program.

In making such decisions, using a statistical (actuarial) method usually results in much more accurate decisions than clinical judgement (Æisdottir et al., 2006; Grove, Zald, Lebow, Snitz, & Nelson, 2000). On average, decisions based on statistical models are 10%-13% more accurate than clinical decisions. In addition, actuarial methods have the added benefit of being 100% reproducible since they follow defined set of procedures.

In academic settings, often the variables used to make statistical decisions about students are *test* scores, broadly defined. These can be scores from an actual test (e.g., SAT), but can also be a grade from a course, GPA, or some other quantified measure. When using a test to make a decision, the results usually will not show compelling evidence for or against the decision; instead, the evidence will be a matter of degree (Swets, 1992). That is, how much better can the program predict success (i.e., graduation) having the information about contained in the test scores than without having such information.

## 1.1. Research Design

The research designs for these projects will typically be made up of archival data, at least at the initial stages. That is, it will use data from students who were already been admitted into the program 4-6 years ago, and for whom there is already data on whether they graduated from the program or not. For example, if a program was looking to implement a test for admissions starting in the 2014-2015 academic year, they would need to to examine data from cohorts of students who started the program during the 2009-2010 academic year and before. This gives the already admitted students (for whom the proposed admission test was not used) the opportunity to have graduated.

Th research design assumes that the proposed admissions test was already administered to the students in these years. If the test was not administered, then the research design will have to *start* the data collection with a current cohort of students. While graduation will still likely be the outcome of interest, proxies for this outcome (e.g., retention, changing major) will likely have to be used for the first 4-6 years.

## 1.2. Decisions and Outcomes

The binary admissions decisions are *probabilistic*, made with more or less confidence depending on the values of the test scores. Thus, making the binary decision in a systematic fashion requires selecting a threshold along the continuum of test scores such that values above the threshold uniformly lead to one decision and values below it lead to the opposite decision.[2]

In the simplest scenario, there are only two decisions (i.e., admit, reject) and two outcomes (i.e., graduate, not-graduate). This leads to four possible outcomes for a particular student when making the admission decision. (Appendix A provides definitions of some of the new terminology.)

[2] The decision is dependent on how outcome is scaled, but I will assume that values above the threshold lead to the positive decision (i.e., admission), while values below the threshold lead to the negative decision (i.e., rejection).

1. If the test says to admit and the student graduates, it is counted as a *true positive* (TP).

2. If the test says to admit and the student does **not** graduate, it is counted as a *false positive* (FP).

3. If the test says to reject and the respondent does **not** graduate, it is counted as a *true negative* (TN).

4. If the test says to reject and the respondent graduates, it is counted as a *false negative* (FN).

These four outcomes can be shown using a 2 × 2 contingency table, an example of which is shown in Table 1. The numbers along the major diagonal represent the correct predictions, while numbers in the off diagonal represent prediction errors.

**Table 1:** *Example Contingency Table for Graduation Outcome.*

| | | True Graduation Status | |
|---|---|---|---|
| | | **Graduate** | **Not Graduate** |
| Predicted | **Graduate** | True Positive (TP) | False Positive (FP) |
| Graduation Status | **Not Graduate** | False Negative (FN) | True Negative (TN) |

The *true positive rate* is defined as

$$\text{True Positive Rate} = \frac{TP}{TP + FN} = \frac{\text{Correctly Predicted to Graduate}}{\text{Total Graduated}} \tag{1}$$

The *false positive rate* of a predictor is defined as

$$\text{False Positive Rate} = \frac{FP}{TN + FP} = \frac{\text{Incorrectly Predicted to Graduate}}{\text{Total } \textbf{not} \text{ Graduated}} \tag{2}$$

Another term for the true positive rate is *sensitivity*. It tells how many times you will admit a student who would graduate from the program. The converse of sensitivity is *specificity*, which is the proportion of non-graduating students who are correctly predicted to **not** graduate, which is defined as

$$\text{Specificity} = \frac{TN}{TN + FP} = \frac{\text{Correctly Predicted to not Graduate}}{\text{Total } \textbf{not} \text{ Graduated}} = 1 - \text{False Positive Rate} \tag{3}$$

Specificity tells you how many times you did **not** admit the student (i.e., reject) who did **not** graduate from the program.

## 1.3. Base Rates

*Base Rate* (i.e., prevalence) is the frequency event of interest in the population.[3] For the purposes of program admissions, the BR is the proportion of students who graduate from the

[3] Here, *population* represents all the students within a program past, present, and future. Of course, data cannot be collected on future students in the present; thus the inference part of the data analysis is that data collected from the past/present students will apply to future students in the program.

program. Stated differently, what is the probability a randomly selected student in the program will graduate? It is defined as

$$\text{Base Rate} = \frac{(\text{True Positive}) + (\text{False Negative})}{\text{All Decisions}}. \tag{4}$$

Although they are typically ignored (Meehl & Rosen, 1955), BRs should play a *Large* part in these diagnostic types of decisions (Finn & Kamphuis, 1995; Treat & Viken, 2012). This is worth repeating: **BRs should play a *Large* part in program admissions decisions**. The more rare the outcome (i.e., smaller BR), the more difficult it is to predict the outcome will occur with accuracy. Likewise, the more common the outcome (i.e., larger BR), the more difficult it is to predict that the outcome will **not** occur with accuracy. Thus, when the BR is *low*, tests work best at *ruling out* an outcome; and when a BR is *high*, tests will work best at *ruling in* an outcome. From a different perspective, if the BR is low, then stronger evidence is needed to be able to predict that the event will occur than when the BR is high. Likewise, if the BR is high, then stronger evidence is needed to be able to predict that the event will **not** occur than when the BR is low.

When the BR for graduating from a program is *low*, an admissions test will do the best at determining students who will likely **not** graduate from the program. Likewise, it will require stronger evidence to predict that a student will graduate from the program than when the BR is high.

When the BR for graduating from a program is *high*, an admissions test will do the best at determining who will likely graduate from the program. Likewise, it will require stronger evidence to predict that a student will **not** graduate from the program than when the BR is low.

Related to the BR are the *odds*, which are

$$\text{Odds} = \frac{\text{Base Rate}}{1 - \text{Base Rate}} \tag{5}$$

## 1.4. Likelihood Ratios

The *Positive Likelihood Ratio* (LR+, $\frac{\text{Sensitivity}}{\text{1-Specificity}} = \frac{\text{TP}}{\text{TN}}$) represents the *odds* that a student admitted into the program, based on some test's value, will actually graduate from the program. It compares how often the test's threshold value occurs in those who graduate (i.e., sensitivity) versus its rate in those students who will **not** graduate (i.e., 1-specificity). For example, a LR+ of 3 can be interpreted as: Being admitted into the program based on the test (at a given threshold) is 3 times greater for those who will graduate than for those who will not graduate.

> ...risk factors or tests producing [Likelihood Ratios] of less than 2 are rarely worth adding to the evaluation process, whereas values around 5 are often helpful, and values greater than 10 frequently have decisive impact on an evaluation. (Youngstrom, 2012, p. 9)

The *Negative Likelihood Ratio* (LR-, $\frac{\text{Specificity}}{\text{1-Sensitivity}} = \frac{\text{TN}}{\text{TP}}$) represents the *odds* that a student rejected from the program would actually **not** graduate from the program. As values become

$> 1$, it indicates that the admission test is giving more information about **not** graduating than guessing (i.e., using the BR alone). For example, a LR- of 2.5 can be interpreted as: Being rejected from the program based on the test (at a given threshold) is two and one-half times greater for those who would **not** graduate from the program than for those who would graduate from the program.

## 1.5. Predictive Values

The *Positive Predictive Value* (PPV, $\frac{\mathrm{TP}}{(\mathrm{TP})+(\mathrm{FP})}$) is the probability of graduating when admitted to the program and *is BR dependent.*

Of all students who are admitted into a program based on the admission test's threshold, only the PPV proportion will actually graduate from the program. Another way of thinking of PPVs is: The probability a student will graduate from the program, given the graduation BR and the student's score on the admissions test. When the PPV values are high, it means that only a few tests indicating the students should be admitted is likely enough information to say the individual will likely graduate from the program.

The *Negative predictive value* (NPV, $\frac{\mathrm{TN}}{(\mathrm{TN})+(\mathrm{FN})}$) is the probability of **not** graduating when rejected from the program and *is BR dependent.*

Of all students who are rejected from the program based on the admission test's threshold, only NPV proportion will actually **not** graduate. Another way of thinking about NPV is: the probability a student will **not** graduating from the program, given the BR of **not** graduating (i.e., 1 - Base Rate) and the students's score on the test. With high NPV values, only a few tests indicating to reject the student from the program is likely enough information to say the student will likely **not** graduate from the program.

Because predictive values are BR dependent, BRs are used in their calculations. Specifically, the PPV weights the LR+ by the appropriate BR, while the NPV weights the LR- by the appropriate BR. This is shown in Equation (6) and Equation (7), respectively.

$$\mathrm{PPV} = \frac{\text{Base Rate} \times \text{Sensitivity}}{(\text{Base Rate}) \times \text{Sensitivity} + (1 - \text{Base Rate}) \times (1 - \text{Specificity})} \tag{6}$$

$$\mathrm{NPV} = \frac{1 - \text{Base Rate} \times \text{Specificity}}{(1 - \text{Base Rate}) \times \text{Specificity} + (\text{Base Rate}) \times (1 - \text{Sensitivity})} \tag{7}$$

What makes PVs nice is that not only do they incorporate the BR, but they can be combined across different tests. I show an example of this in Section 2.

The relationship between BRs and predictive values is illustrated in Table 2. Table 2a is the best case scenario because the program admissions test has both high sensitivity and high specificity. Except for the extremely low BRs (i.e., $< .05$) having a test score that meets the criterion indicates that the student has a non-negligible probability of graduating from the program. Likewise, except for the extremely high BRs (i.e., $> .95$) having a test score that does **not** meet the criterion indicates that the student has a non-negligible probability of **not** graduating from the program.

Table 2b provides a more realistic scenario for higher education. Here, having a test score meet the criterion increases the student's probability of graduating from the program, but only small amount. Likewise, having a test score **not** meet the criterion increases the student's probability of **not** graduating from the program, but only small amount.

**Table 2:** *Relationship between base rates and predictive values for certain levels of sensitivity and specificity of a test's score.*

| Base Rate | + PV | -PV |
|---|---|---|
| 0.99 | 1.00 | 0.16 |
| 0.95 | 1.00 | 0.50 |
| 0.90 | 0.99 | 0.68 |
| 0.50 | 0.95 | 0.95 |
| 0.10 | 0.68 | 0.99 |
| 0.05 | 0.50 | 1.00 |
| 0.01 | 0.16 | 1.00 |

(a) *Sensitivity=.95 and Specificity = .95.*

| Base Rate | + PV | -PV |
|---|---|---|
| 0.99 | 1.00 | 0.02 |
| 0.95 | 0.98 | 0.11 |
| 0.90 | 0.95 | 0.21 |
| 0.50 | 0.70 | 0.70 |
| 0.10 | 0.21 | 0.95 |
| 0.05 | 0.11 | 0.98 |
| 0.01 | 0.02 | 1.00 |

(b) *For Sensitivity=.70 and Specificity = .70*

| Base Rate | + PV | -PV |
|---|---|---|
| 0.99 | 0.99 | 0.01 |
| 0.95 | 0.95 | 0.05 |
| 0.90 | 0.90 | 0.10 |
| 0.50 | 0.50 | 0.50 |
| 0.10 | 0.10 | 0.90 |
| 0.05 | 0.05 | 0.95 |
| 0.01 | 0.01 | 0.99 |

(c) *For Sensitivity=.50 and Specificity = .50*

Table 2c shows what happens when using tests whose scores do not give much information about program admissions. Irrespective of whether the test score met the criterion or not, the student's probability of graduating, or **not** graduating, from the program is equivalent to the BR.

As an aside, test scores can have strong reliability and validity evidence, but still not be useful in making decisions (Hayes, Nelson, & Jarrett, 1987; Wasserman & Bracken, 2003).

Tests whose scores do not give much information (i.e., have low sensitivity and specificity) do very little to aid in the decision making process above and beyond what is known from the base rate.

## 1.6. Receiver Operating Characteristic Curves

One way to examine if a test is able to predict graduation is to use receiver operating characteristic (ROC; Zweig & Campbell, 1993) curves. ROC curves are two-dimensional graphs of the TP rate ($Y$ axis) against the FP rate ($X$ axis). It depicts relative trade-offs between benefits (TPs) and costs (FP) of using the test. An example is given in Figure 1. The lower left point (0, 0) in an ROC curve represents the situation of never issuing a positive prediction (i.e., no student will graduate). While this situation commits no FP errors, it also gains no TPs. The opposite situation (i.e., unconditionally predicting that all the students will graduate) is represented by the upper right point (1, 1). The point (0, 1) represents perfect prediction by the test.

Usually, a point on an ROC curve graph is better than another if it is to the northwest (TP rate is higher, FP rate is lower, or both) of the other point. Points appearing on the left hand-side of an ROC graph, near the $X$ axis, are conservative because a positive prediction is made only with strong evidence. While there are few FPs at this cutoff of the predictor, they often have low TP rates as well. Points on the upper right-hand side of an ROC graph are more liberal as they make positive predictions with weak evidence; thus, they classify nearly all positives correctly, but they often have a high FP rate.

The diagonal line $y = x$ represents the strategy of randomly guessing membership, (i.e., it can be expected to correctly predict half the students who graduate and half the students who do **not** graduate). Thus a test that does not do any better than random will produce a ROC point that hovers around the diagonal. Any test that appears in the lower right triangle performs worse than random guessing, and is usually empty in ROC graphs.

One way to reduce ROC performance to a single value, for comparisons across predictors, is to calculate the area under the ROC curve, abbreviated AUC. Its value will always be between 0 and 1.0. Random guessing produces lines along the diagonal line between (0, 0) and (1, 1), which has an AUC of 0.5; thus, no AUC values should be less than 0.5.

## 1.7. Setting the Test's Threshold

The optimal value on a test to use as the threshold (i.e., cut score) varies from one situation to another, and will depend on the BRs as well as the costs and benefits of the TPs and FPs.

With a *Liberal* threshold, it : (a) is *easier* to get admitted to the program; (b) lowers specificity; and (c) raises sensitivity. Liberal test thresholds cast a wide net. Thus, they make it more likely to admit everyone who would graduate from the program (TP), but also make it more likely to admit more students who will **not** graduate (FP).

With a *Conservative* criterion, it: (a) is *harder* to get admitted to the program; (b) raises specificity; and (c) lowers sensitivity. Conservative test thresholds cast a narrow net. Thus, they make it is more likely to reject everyone who would **not** graduate (TN), but also make it more likely to reject more individuals who would have graduated (FN).

### 1.7.1. Incorporating Costs and Benefits in Determining Test Thresholds

The decision of where to set the test threshold for decision making needs to take into account both the BRs *and* the costs and benefits of correct and incorrect predictions (Swets, 1992).[4] One method of incorporating costs and benefits is to make a ratio of the benefits related to admission vs. rejection. One such ratio is

$$\text{ROC Optimal Threshold} = \frac{1 - BR}{BR} \times \frac{B_{TN} - C_{FP}}{B_{TP} - C_{FN}} \tag{8}$$

[4] This is not the same as a cost-benefit analysis as is typically done with program evaluations (Yates, 1994; Yates & Taub, 2003). Instead, this section refers to the cost and benefit associated with setting a test's threshold.

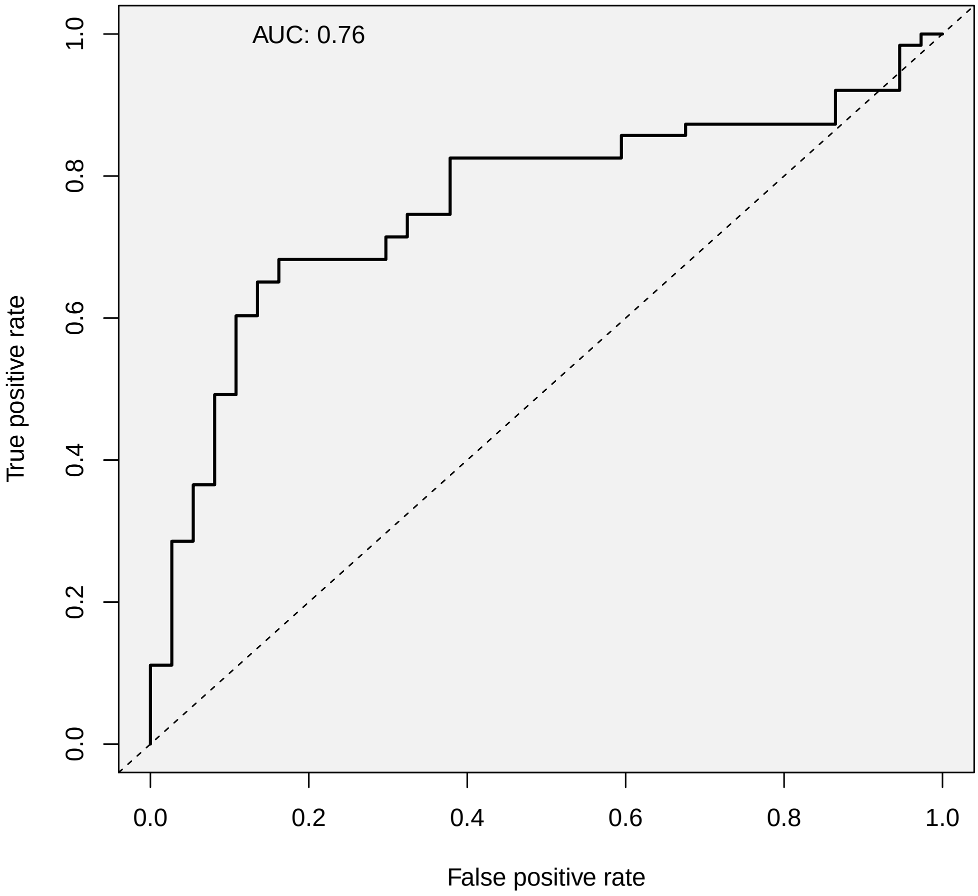


**Figure 1:** *Example receiver operating characteristic (ROC) curve.*

where $B_{TN}$ and $B_{TP}$ are the benefits of correct decisions and $C_{FP}$ and $C_{FN}$ are the costs of incorrect decisions (Metz, 1978). The ROC optimal threshold ($\mathrm{OT}_{\mathrm{ROC}}$) is the value on the ROC curve that best balances the weight of the costs and benefits. To use it, one needs to find the slope of the ROC curve (Somoza, Steer, Beck, & Clark, 1994) that matches the $\mathrm{OT}_{\mathrm{ROC}}$ and then map that value onto the test threshold that gives the appropriate FP and TP rates.

For a fixed set of benefits and costs, the $\mathrm{OT}_{\mathrm{ROC}}$ will be relatively large when $1-BR$ (the base rate for the event [i.e., graduation] **not** occurring) is appreciably greater than the $BR$. In other words, one needs very strong evidence to predict a student will graduate from a program when the probability of **not** graduating from the program is high. With a higher $BR$, the $\mathrm{OT}_{\mathrm{ROC}}$ will be smaller and more lenient. This allows for a higher FP rate, but this can be tolerated because the number of students not graduating is relatively small.

If the two kinds of wrong decisions are equally bad and the two kinds of correct decision are equally good (i.e., $B_{TN} = B_{TP}$ and $C_{FP} = C_{FN}$), then the $\mathrm{OT}_{\mathrm{ROC}}$ will be determined solely by the BRs and will be the place on the ROC curve where the slope equals $\frac{1-BR}{BR}$.

When the BRs are constant, the $\mathrm{OT}_{\mathrm{ROC}}$ is large when the numerator of the benefit-cost part of Equation (8) ($B_{TN} - C_{FP}$) is large compared to the denominator ($B_{TP} - C_{FN}$), which is the case when more importance is attached to the accuracy of predicting who will **not** graduate. Conversely, the $\mathrm{OT}_{\mathrm{ROC}}$ is small and lenient when the benefit-cost denominator is large relative to its numerator, which is the case when it is more important to make correct decisions about predicting who will graduate than predicting who will **not** graduate.

Treat and Viken (2012) suggests two other ways to incorporate costs and benefits when determining a test's threshold. One is utility and the other is information.

### 1.7.1.1. Utility

This approach requires specifying the differential utility (i.e., place different values) of TPs, FPs, TNs, and FNs. As with $\text{OT}_{\text{ROC}}$, this approach takes into account the BRs and assigns a value to each of the four possible outcomes. The *overall utility* ($U_{Overall}$) weights the probability of a outcome (i.e., BR × rate of occurrence) by the outcome's utility, and then sums over the four outcomes, as shown in Equation (9).

$$\begin{aligned}\text{Overall Utility} = {} & (BR)(TP_R)(TP_U) + (BR)(FN_R)(FN_U) \\ & + (1 - BR)(FP_R)(FP_U) + (1 - BR)(TN_R)(TN_U)\,, \end{aligned} \tag{9}$$

where the $U$ subscripts represents the utility value and the $R$ subscript represents the rate. As $FN_R = 1 - TP_R$ and $TN_R = 1 - FP_R$, Equation (9) can be simplified to

$$\begin{aligned}\text{Overall Utility} = {} & (BR)(TP_R)(TP_U) + (BR)(1 - TP_R)(FN_U) \\ & + (1 - BR)(FP_R)(FP_U) + (1 - BR)(1 - FP_R)(TN_U). \end{aligned} \tag{9b}$$

To make things easier, utility values are typically set to range between 0 (least desired) and 1 (most desired). Thus, correct decisions (i.e., TPs and TNs) are assigned utilities $\geq .5$ while incorrect decisions (i.e., FPs and FNs) are assigned utilities $\leq .5$.

To maximize the proportion of correct decisions using the $U_{Overall}$ criterion, set $TP_U = TN_U = 1$ and $FN_U = FP_U = 0$. Doing this caries the major assumptions that correctly admitting and correctly rejecting students is of equal benefit. Likewise, it also assumes that incorrectly admitting and incorrectly rejecting students is of equal cost. Neither of these assumptions might actually be the case, however, so the values assigned to the utilities should be done with care.

### 1.7.1.2. Information Gain

A third criterion that can be used in setting the test's threshold is the *information gain* ($I_{Gain}$) function (Metz, Goodenough, & Rossmann, 1973), shown in Equation (10).[5] Here, the information gained refers to the reduction of uncertainty about the student's true graduation status of a student that results from knowing the student's score on the test. As with $\text{OT}_{\text{ROC}}$ and $U_{Overall}$, $I_{Gain}$ takes into account the BRs and assigns a value to each of the four outcomes. With $I_{Gain}$, however, there are no subjective values set for the outcomes.

[5] This is also referred to as the Kullback–Leibler divergence function (Kullback & Leibler, 1951).

$$\begin{aligned}\text{Information Gain} = (BR)(TP_R)(\log_2\left[\frac{TP_R}{SR}\right]) \\ + (BR)(1-TP_R)(\log_2\left[\frac{(1-TP_R)}{(1-SR)}\right]) \\ + (1-BR)(FP_R)(\log_2\left[\frac{FP_R}{SR}\right]) \\ + (1-BR)(1-FP_R)(\log_2\left[\frac{(1-FP_R)}{(1-SR)}\right]), \end{aligned} \tag{10}$$

where the $R$ subscript represents the rate, $\log_2$ is the base 2 logarithm, and $SR$ is the *Selection Ratio*: the proportion of a sample predicted to have the characteristic of interest based on the test's threshold.

## 2. Example

As an example of implementing the procedures outlined in this document, I simulated a dataset that has the following variables: (a) SAT-Math, (b) Course grade (A=4, B=3, C=2, D=1, F=0), (c) First-year GPA, and (d) Graduation status (yes=1, no=0). I sampled $n = 1000$ observations (students). The correlations of the sample data are shown in Table 3.

The BR for the sample is 0.45. A contingency table for the data before using any admissions test is shown in Table 4. If I had no test data, then I would predict 45 out of every 100 students in the program would graduate, and 55 would not. Alternatively, the odds of graduating from this program are: $\frac{\text{Base Rate}}{1-\text{Base Rate}} = \frac{0.45}{1-0.45} = 0.83$.

I plotted the ROC as well as sensitivity and specificity for multiple SAT scores in Figure 2. I'll use a SAT score of 600 as the initial threshold, as that appears to be where sensitivity and specificity are both about equal. The results of using a SAT (Math) score of 600 as the threshold produces the results shown in Table 5 and Figure 3.

The PPV of 0.52 means that in a random sample of students in the program that were predicted to graduate based on having a SAT score $\geq 600$, 52% will actually graduate (TPs). Conversely, 48% of those predicted to graduate will not (FPs). As the graduation BR for this program is 0.45, using the SAT score of 600 as the admission threshold is doing better than chance at predicting graduation. Stated differently, before knowing students' scores on the admissions test, their probability of graduating from the program was 0.45. After knowing the students have a score on the SAT $\geq 600$, their probability of graduating from the program increases to 0.52.

**Table 3:** *Correlations for Simulated Student Data from a Single Program (n=1000).*

| | GPA | SAT-M | Course Grade | Graduate |
|---:|---:|---:|---:|---:|
| GPA | 1.00 | 0.38 | 0.25 | 0.51 |
| SAT-M | 0.38 | 1.00 | 0.32 | 0.20 |
| Course Grade | 0.25 | 0.32 | 1.00 | 0.22 |
| Graduate | 0.51 | 0.20 | 0.22 | 1.00 |

**Table 4:** *Program's Known Graduation Data Before using an Admissions Test.*

| | | Reality | | Predicted |
|---|---|---|---|---|
| | | Graduate | Not Graduate | Totals |
| Prediction from Data | Graduate | True + | False + | |
| | Not Graduate | False - | True - | |
| Reality Totals | | 453 | 547 | |

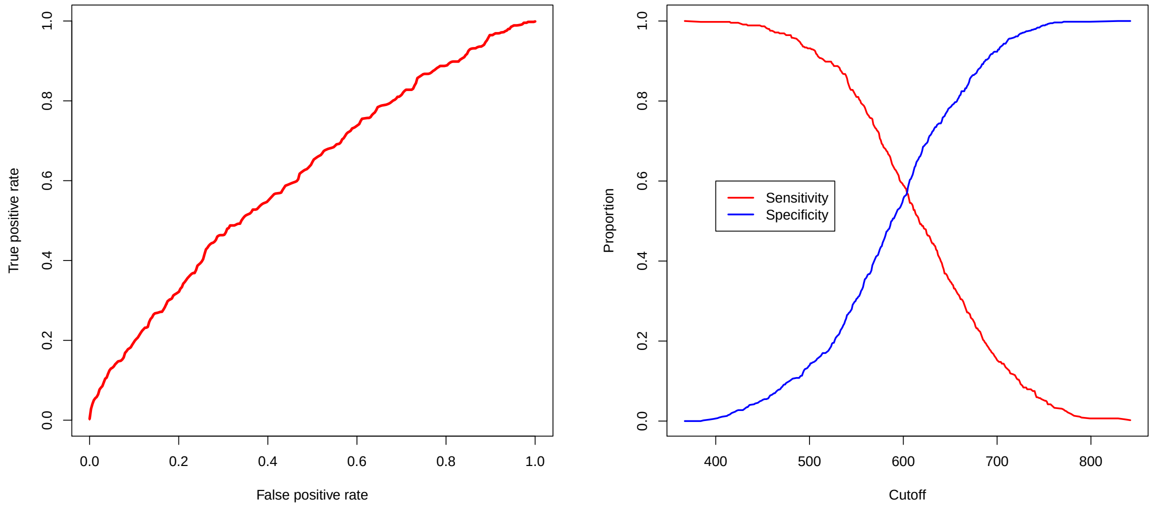
## 2.1. Combining Test Results

The PPVs can be combined across multiple tests. In the current example, the original BR is 0.45. For students whose SAT scores are $\geq$ 600, their PPV is 0.52. We can use this PPV as their *new* BR for graduation. That is, the BR of graduating form the program for students with SAT scores $\geq$ 600 is 0.52.

In addition to SAT scores, the data also contains the students' grades in an introductory course in the major. Arbitrarily, I set the threshold for this test to earning a grade of B- or higher. The sensitivity and specificity for this new test at the the B- threshold are 0.65 and 0.52, respectively. This makes the LR+ 1.37. Using a BR of 0.52, the PPV for students who met the thresholds for both the SAT and course grade tests is $\frac{0.52 \times 0.65}{0.52 \times 0.65 + 0.48 \times 0.48} = 0.6$. Thus, for students with SAT scores $\geq$ 600 *and* who earned grades of B- or higher in the introductory course, their probability of graduating from the program is 0.6.

## 2.2. Determining the Optimal Test Threshold

In Table 6 I show the optimal test thresholds under the different criteria I discussed in Section 1.7.1. The table makes it obvious that the optimal value to use for the SAT's threshold is dependent on both the BR and the importance of different costs and benefits of the admission/rejection decisions.



(a) *ROC Curve.*

(b) *Sensitivity and Specificity.*

**Figure 2:** *Plots for SAT scores.*

**Table 5:** *Program's Graduation Data After using a SAT Score of 600 for the Admission Threshold.*

| | | Reality | | Predicted |
|---|---|---|---|---|
| | | Graduate | Not Graduate | Totals |
| Prediction from Data | Graduate | 267 | 243 | 510 |
| | Not Graduate | 186 | 304 | 490 |
| Reality Totals | | 453 | 547 | 1000 |

- Sensitivity: $\frac{267}{453} = 0.59$
- Specificity: $\frac{304}{547} = 0.56$
- LR+: $\frac{\text{Sensitivity}}{1-\text{Specificity}} = \frac{0.59}{1-0.56} = 1.33$
- LR- : $\frac{\text{Specificity}}{1-\text{Sensitivity}} = \frac{0.56}{1-0.59} = 1.35.$
- PPV: $\frac{267}{510} = \frac{\text{Base Rate}\times\text{Sensitivity}}{\text{Base Rate}\times\text{Sensitivity}+(1-\text{Base Rate})\times(1-\text{Specificity})} = \frac{0.45\times0.59}{0.45\times0.59+0.55\times0.44} = 0.52.$
- NPV: $\frac{304}{490} = \frac{(1-\text{Base Rate})\times\text{Specificity}}{(1-\text{Base Rate})\times\text{Specificity}+\text{Base Rate}\times(1-\text{Sensitivity})} = \frac{0.55\times0.56}{0.55\times0.56+0.45\times0.41} = 0.62$

**Figure 3:** *Results from using a SAT score of 600 as the test threshold.*

**Table 6:** *SAT Threshold Values as a Function of Base Rates and Optimization Function.*

| | Base Rate (Graduation) | | | | | | |
|---|---|---|---|---|---|---|---|
| Optimizing Function | .30 | .40 | **.45** | .50 | .60 | .70 | .80 |
| $\mathrm{OT}_{\mathrm{ROC}}$, with admission and rejection of equal importance | 752 | 713 | 634 | 628 | 538 | 452 | 424 |
| $\mathrm{OT}_{\mathrm{ROC}}$, with rejection twice the importance of admission | 771 | 757 | 757 | 713 | 634 | 567 | 480 |
| $\mathrm{OT}_{\mathrm{ROC}}$, with admission twice the importance of rejection | 634 | 538 | 480 | 480 | 447 | 424 | 367 |
| $U_{Overall}$, with admission and rejection of equal importance | 752 | 713 | 634 | 628 | 538 | 452 | 424 |
| $U_{Overall}$, with rejection twice the importance of admission | 771 | 757 | 757 | 713 | 634 | 567 | 480 |
| $U_{Overall}$, with admission twice the importance of rejection | 634 | 538 | 480 | 480 | 447 | 424 | 367 |
| $I_{Gain}$ | 634 | 634 | 634 | 634 | 628 | 628 | 628 |

*Note*. The base rate for the current sample is 0.45. $\mathrm{OT}_{\mathrm{ROC}}$: ROC optimal threshold. $U_{Overall}$: Overall utility. $I_{Gain}$: Information gain.

## 3. Recommendations

- Admission steps and criteria need to be documented in great detail so that the same procedures are applied to every applicant. While this does not remove the possibility of clinically-informed decisions (e.g., making a decision based on an interview), it should make their occurrences rare.

- Before a program's faculty can determine admissions criteria, they need to figure out (a) the current graduation BR for their program, and (b) the relative importance of acceptance vs. rejection decisions. This information should be incorporated into their decision-making process for the thresholds used for their admissions tests. Perhaps better that using a single BR, however, is to allow the BR to vary across a range of plausible values (Treat & Viken, 2012), as I did in Table 6.

- As BRs tend to fluctuate, the BR used to develop the admission test's threshold, as well as the test threshold itself, should be examined at least every third year.

- Admission test thresholds should be based on *at least* three cohorts of students; this number will likely need to increase (sometimes greatly) for programs with small cohorts. If possible, the cohort data should be be analyzed both separately and combined to see if the optimal test thresholds differ (i.e., sensitivity analysis). This should minimize the influence of flukes in the data.

- In addition to the methods described here, program faculty should include a projected cost-benefit analysis of implementing the admissions decisions. Using the criteria outlined in this document, programs can predict the number of students admitted and number of students who will graduate from a given cohort using the proposed test. This information can then be used as part of a more traditional cost-benefit analysis to predict the impact using the proposed test will have on the program. Yates (2012) provides a gentle introduction to this topic.

## References

Æisdottir, S., White, M. J., Spengler, P. M., Maugherman, A. S., Anderson, L. A., Cook, R. S., . . . Rush, J. D. (2006). The meta-analysis of clinical judgment project: Fifty-six years of accumulated research on clinical versus statistical prediction. *The Counseling Psychologist*, *34*, 341-382. doi: 10.1177/0011000005285875

Finn, S. E., & Kamphuis, J. H. (1995). What a clinician needs to know about base rates. In J. N. Butcher (Ed.), *Clinical personality assessment: Practical approaches* (p. 224-235). New York, NY: Oxford University Press.

Grove, W. M., Zald, D. H., Lebow, B. S., Snitz, B. E., & Nelson, C. (2000). Clinical versus mechanical prediction: A meta-analysis. *Psychological Assessment*, *12*, 19-30. doi: 10.1037/1040-3590.12.1.19

Hayes, S. C., Nelson, R. O., & Jarrett, R. B. (1987). The treatment utility of assessment: A functional approach to evaluating assessment quality. *American Psychologist*, *42*, 963-974. doi: 10.1037/0003-066X.42.11.963

Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, *22*, 79-86.

Meehl, P. E., & Rosen, A. (1955). Antecedent probability and the efficiency of psychometric signs, patterns, or cutting scores. *Psychological Bulletin*, *52*, 194-216. doi: 10.1037/h0048070

Metz, C. E. (1978). Basic principles of ROC analysis. *Seminars in Nuclear Medicine*, *8*, 283-298. doi: 10.1016/S0001-2998(78)80014-2

Metz, C. E., Goodenough, D. J., & Rossmann, K. (1973). Evaluation of receiver operating characteristic curve data in terms of information theory, with applications in radiography. *Radiology*, *109*, 297-303. doi: doi:10.1148/109.2.297

Somoza, E., Steer, R. A., Beck, A. T., & Clark, D. A. (1994). Differentiating major depression and panic disorders by self-report and clinical rating scales: Roc analysis and information theory. *Behaviour Research and Therapy*, *32*, 771-782. doi: http://dx.doi.org/10.1016/0005-7967(94)90035-3

Swets, J. A. (1992). The science of choosing the right decision threshold in high-stakes diagnostics. *American Psychologist*, *47*, 522-532. doi: 10.1037/0003-066X.47.4.522

Treat, T. A., & Viken, R. J. (2012). Measuring test performance with signal detection theory techniques. In H. Cooper (Ed.), *APA handbook of research methods in psychology* (Vol. 1, p. 723-744). Washington, DC: American Psychological Association.

Wasserman, J. D., & Bracken, B. A. (2003). Psychometric characteristics of assessment procedures. In J. R. Graham & J. A. Naglieri (Eds.), *Handbook of psychology (Vol 10): Assessment psychology* (pp. 43–66). Hoboken, NJ: John Wiley and Sons. Retrieved from http://dx.doi.org/10.1002/0471264385.wei1003.

Yates, B. T. (1994). Toward the incorporation of costs, cost-effectiveness analysis, and cost-benefit analysis into clinical research. *Journal of Consulting and Clinical Psychology*, *62*, 729-736. doi: 10.1037/0022-006X.62.4.729

Yates, B. T. (2012). Program evaluation: Outcomes and costs of putting psychology to work. In H. Cooper, P. M. Camic, D. L. Long, A. T. Panter, D. Rindskopf, & K. J. Sher (Eds.), *APA handbook of research methods in psychology, Vol 2: Research designs: Quantitative, qualitative, neuropsychological, and biological* (pp. 569–586). Washington, DC: American Psychological Association.

Yates, B. T., & Taub, J. (2003). Assessing the costs, benefits, cost-effectiveness, and cost-benefit of psychological assessment: We should, we can, and here's how. *Psychological Assessment*, *15*, 478-495. doi: 10.1037/1040-3590.15.4.478

Youngstrom, E. A. (2012). Future directions in psychological assessment: Combining evidence-based medicine innovations with psychology's historical strengths to enhance utility. *Jour-*

*nal of Clinical Child and Adolescent Psychology*, 1-21. doi: 10.1080/15374416.2012.736358

Zweig, M. H., & Campbell, G. (1993). Receiver-operating characteristic (ROC) plots: A fundamental evaluation tool in clinical medicine. *Clinical Chemistry*, *39*, 561-77.

# A. Terminology

| Term | Definition |
|---|---|
| Area under the curve (AUC) | The probability that a randomly selected pair of positive and negative cases will be ranked correctly by the test. |
| Base rate (BR, Prevalence) | Proportion of cases that have the characteristic of interest. |
| False negatives (FN, Miss) | Cases that have that have the characteristic of interest that the test classified as not having the characteristic. |
| False positives (FP, False alarms) | Cases that do not have the characteristic of interest that the test classified as having the characteristic. |
| Negative predictive value (NPV) | Proportion of cases classified as not having the characteristic who actually do not have the characteristic. |
| Positive predictive value (PPV) | Proportion of cases classified as having the characteristic who actually have the characteristic. |
| Sensitivity | Proportion of cases that actually have the characteristic that the test classifies as having the characteristic. |
| Specificity | Proportion of cases that actually do not have the characteristic that the test classifies as not having the characteristic. |
| Test threshold | Value of test that distinguishes cases classified as having and not having the characteristic of interest. |
| True positive (TP, Hit) | Cases that have the characteristic of interest that the test classified as having the characteristic. |
| True negative (TN, Correct rejection) | Cases that do not have the characteristic of interest that the test classified as not having the characteristic. |

# Index