Journal of Postsecondary Student Success 4(3)
doi: 10.33009/fsop_jpss135687

The Classification Power: A Metric for Evaluating and Comparing Postsecondary Mathematics Placement Policies

Kristin M. Frank*, Alexei Kolesnikov, & Xiaoyin Wang

Towson University

Abstract

We introduce a new metric—the classification power—to examine the effectiveness of postsecondary mathematics placement policies. This metric addresses the methodological challenges of contextualizing the effectiveness of a single placement policy and comparing the effectiveness of multiple placement policies across different student populations. We leverage an information-theoretic approach to construct the classification power as a measure of the improvement of the implemented policy over a hypothetical policy that places students by a Bernoulli coin-flip process. We expect an effective policy to perform significantly better than this hypothetical weighted coin-flip placement. We illustrate the utility of this metric by applying this methodology to institutional data from our four-year institution. We find the classification power provides more information about the quality of placement compared to the previously defined severe error rate metric.

Keywords: mathematics placement, four-year institution, classification power, error rate, bootstrapping, information theory

* Contact: kfrank@towson.edu

The Classification Power: A Metric for Evaluating and Comparing Postsecondary Mathematics Placement Policies

Prior to matriculating in a postsecondary institution, most U.S. students take a mathematics placement exam to help match each student with an appropriate first math course (Fields & Parsad, 2012). Many four-year institutions use a student's placement exam score to determine whether the student is ready for college-level mathematics and, if so, for which college mathematics course the student is ready (e.g., Precalculus or Calculus I). It is common for students and faculty to express dissatisfaction with placement policies (e.g., Meuschke & Gribbons, 2003). Thus, it is essential to have robust tools for evaluating and comparing the effectiveness of placement policies to inform placement policy decisions.

One approach to evaluating the effectiveness of a placement policy is to treat the placement decision as a classification problem where the goal of placement is to identify, with as high accuracy as possible, the students who are ready to take the target course and those who need to take a prerequisite (see Belfield & Crosta, 2012; Scott-Clayton, 2012). The quality of the decision can be measured with the *error rate*, the proportion of students who have been misplaced by either being overplaced into a course they are not prepared for or unnecessarily underplaced into a prerequisite course.

While the error rate is one approach to measure the quality of the placement decision, we questioned whether counting the number of errors made by a placement policy was sufficient to evaluate and compare the placement policy's ability to determine students' readiness for a given course. More specifically, we identified three challenges with using the error rate to inform the evaluation of placement policies: (a) challenges associated with determining whether a given error rate is evidence of an effective placement policy; (b) challenges associated with comparing error rates across placement policies; and (c) interpretational challenges due to possible variation in the magnitude of the error rate due to chance. These challenges arise because the error rate is highly dependent upon characteristics of the sample. Thus, our work focused on developing a new metric that simultaneously contextualizes the error rate by attending to characteristics of the sample and also provides a benchmark that supports administrators in evaluating and comparing mathematics placement policies.

In this paper we use an information-theoretic approach to define a new metric—the classification power. This metric quantifies the amount of information provided by the placement policy about students' readiness for a given course by measuring the improvement in the error rate of an implemented placement policy and a hypothetical policy that provides no information about students' readiness. Since both the implemented and hypothetical policies place the same population of students and the classification power measures the improvement of the error rate over the hypothetical policy, one can directly compare the classification power of placement policies implemented with different student populations. We conclude this paper by using institutional data from our university to illustrate the limitations of the error rate and the utility of the

classification power in evaluating and comparing placement policies. We intend to show that the classification power provides more information than the error rate about a placement policy's ability to identify students ready for a given course and thus is useful for administrators making decisions around postsecondary placement.

Relevant Literature

The literature provides two main strategies to evaluate the effectiveness of a placement policy: comparative analysis of observable outcomes and classification error analysis. We briefly describe these approaches and key findings.

Many researchers examine the relationship between placement decisions and observable outcomes such as: students' grades (or pass rates) in their first credit-bearing math course, the percentage of students taking and completing college-level mathematics at the beginning of their study, and time to graduation (e.g., Ayele et al., 2023; Barnett et al., 2018; Chen, 2016; Rodríguez, 2014; Thornton et al., 2019). Frank et al. (2022) argued that comparative analysis of observable outcomes is not sufficient to meaningfully evaluate and compare placement policies since the placement decision has low explanatory power for both short-term and long-term outcome variables.

Around 2010, researchers at the Community College Research Center (CCRC) developed a new methodology, based on Sawyer's (1996) construct of accuracy rates, to measure the predictive validity of a placement test (see Belfield & Crosta, 2012; Scott-Clayton, 2012; Scott-Clayton et al., 2014). As Scott-Clayton (2012) explains, this approach treats placement as a classification problem and quantifies the error in the placement decision by estimating the number of students misclassified by the placement policy. The CCRC studies examined placement policies designed to identify whether a student was ready for college-level mathematics or if they should first take developmental math. In this context there are two ways students can be misclassified: students can be placed into a college-level mathematics while not being ready for that course (these are overplaced students) or students can be placed into developmental math despite being ready for college-level math (these are underplaced students). The error rate is the sum of the percentage of overplaced students and the percentage of underplaced students.

The challenge with this approach is that the percentage of misplaced students is not directly observable from the data. Researchers at CCRC (see Belfield & Crosta, 2012; Scott-Clayton, 2012) used a form of statistical matching between the students who took college-level mathematics without remediation and those who took a developmental math course before enrolling in college-level math to estimate the relationship between the placement decision and a student's readiness to succeed in college-level math. They constructed multiple models based upon different success criteria such as earning a grade of B or better in the target course, earning a grade of C or better in the target course, or earning a grade of D or better in the target course. They applied each

model to students who took the prerequisite course to estimate the percentage of underplaced students. Using a C as the cutoff for success, the researchers computed accuracy rates between 49% and 58%. Whereas, using a D as the cutoff for success produced accuracy rates between 40% and 55% (Belfield & Crosta, 2012; Scott-Clayton, 2012). These findings suggest that nearly half of the students were misplaced.

As Scott-Clayton (2012) described, there is a significant limitation to considering multiple success criteria; each success criteria lends itself to a different interpretation and recommendation for policy makers. To address this limitation, Scott-Clayton introduced the *severe error rate* (SER) metric to isolate particularly egregious misplacement error and provide a single metric for the accuracy rate, or in this case, the complementary error rate. She defined the severe error rate as the sum of the percentage of students who were placed into developmental math but would have been able to pass college-level mathematics with the grade of B or higher and the percentage of students who were placed into college-level math but earned a failing grade (F). The CCRC studies report severe error rates between 16% and 28% in the context of mathematics placement at two-year institutions (Belfield & Crosta, 2012; Scott-Clayton, 2012).

To contextualize the accuracy rates, the CCRC researchers also reported accuracy rates associated with hypothetical placement systems that assign all the students to a lower-level course or all the students to a college-level course; see Table 8 of Belfield & Crosta (2012) and Table 3 of Scott-Clayton (2012). As Scott-Clayton (2012) explains, "In most cases, similar or even higher accuracy rates could have been achieved without using the placement exams at all, but instead by assigning all students to either the developmental or college-level course" (p. 23). We return to this finding in the next section.

Leeds and Mokher (2020) extend the work done at CCRC and use SERs to examine misplacement in the context of multilevel placement policies. Like the CCRC studies, Leeds and Mokher studied placement in a large community college system. However, they considered a multi-level placement system: a student's score on the placement test is used to determine whether the student is placed in lower-level developmental mathematics, upper-level developmental mathematics, or college-level mathematics. They found that, (a) changing the cut score can improve placement accuracy and (b) eliminating placement into upper-level developmental mathematics so that students either take two developmental mathematics courses or enroll directly into college level mathematics would reduce the number of severe errors.

Theoretical Framework

We assume that the goal of mathematics placement at postsecondary institutions is to determine whether a student is ready (or not) for a given course (we call it the *target course*). We position the effectiveness of a placement policy as the amount of information the placement decision provides about a student's readiness for the target

course. This amount of information can then be compared across placement policies. We provide both an empirical and theoretical example to illustrate the distinction between the number of severe errors made by a placement policy and the amount of information the placement decision conveys about students' readiness.

For an empirical example, we return to Scott-Clayton's (2012) finding that one can often reduce the number of severe errors by allowing all students into the course. While this finding is likely to be true at many institutions, test-based placement policies and remedial math programs are still prominent. We argue placement policies are preferred because they provide some information about students' readiness whereas allowing all students into a given course provides no information. This example illustrates how reducing the SER might not result in a placement decision that provides more information about students' readiness.

For a theoretical example, consider two placement policies that are applied to two different populations of 100 students. Both policies overplace 20 not-ready[1] students and underplace 20 ready students so both policies have an error rate of 0.4. However, there is variation among the two populations; the population placed by Policy A has 20 not-ready students while the population placed by Policy B has 40 not-ready students. Thus, Policy A overplaced *all* the not-ready students and did not identify any not-ready students whereas Policy B overplaced half of the not-ready students. While both policies made the same number of errors, we argue Policy B provides more information about students' readiness.

This theoretical example highlights the need to consider separately the proportion of severe errors amongst the ready and not-ready students. More specifically, we recognize two types of severe errors:

- Severe Error for Ready Students (SER_R): Proportion of students who are placed into a prerequisite course but are ready for the target course without taking a prerequisite (i.e., underplaced students).
- Severe Error for Not-Ready Students ($SERN_R$): Proportion of students who are placed into the target course but are not ready and failed the target course on their first attempt (i.e., overplaced students).

We emphasize that these estimates are byproducts of computing the SER where the overall SER is the sum of SER_R and SER_{NR}. In the next section we explain how we utilize the estimates for both SER_R and SER_{NR} to quantify the amount of information a placement decision conveys about students' readiness.

1 We formally define *ready* and *not-ready* in the Methodology section.

Information Theory

The field of Information Theory has developed robust methods for measuring the amount of information one variable provides about another. One of the metrics is the Pointwise Mutual Information (PMI). First defined by Fano (1961), the PMI compares the empirical probability of simultaneous occurrence of two events to the probability of simultaneous occurrence if the events were independent. Formally, if x and y are two outcomes of random variables X and Y then the PMI of (x, y) is $\log_2 \frac{\mathbf{P}(x, y)}{\mathbf{P}(x)\mathbf{P}(y)}$. The PMI will be positive if x is more likely to occur given y and the PMI will be negative if x is less likely to occur given y (for more details, see Cover & Thomas, 1991).

In the case of placement, the two random variables of interest are a student's readiness for a given course (X) and the placement decision (Y). We consider two events for each variable: *ready* and *not ready*, and *placed* and *not placed*. Consistent with computing the SER, we focus on the amount of information lost through errors in the placement decision and consider the outcomes *ready* and *not placed* as well as *not ready* and *placed*.

Historically, applications of information theory dealt with transmission of bits of information, so the logarithm offered a natural interpretation of the metric. For our work in analyzing and comparing placement policies the ratio, $\frac{\mathbf{P}(x, y)}{\mathbf{P}(x)\mathbf{P}(y)}$, and more specifically the deviation of this ratio from 1, will provide a more intuitive interpretation.

First, we treat subsets of the target course population as probability-theoretic events to express the SER in terms of probabilities. The SER_R is the probability a student is *ready* and *not placed* (i.e., $SER_R = \boldsymbol{P}$ (Ready, Not Placed)) while the SER_{NR} is the probability a student is *not ready* and *placed* (i.e., $SER_{NR} = \boldsymbol{P}$ (Not Ready, Placed)). We can then interpret the numerator of the ratio $\frac{\mathbf{P}(x, y)}{\mathbf{P}(x)\mathbf{P}(y)}$ as the SER_R (or SER_{NR}) of an implemented placement policy. We interpret the denominator of the ratio as the SER_R (or SER_{NR}) of a hypothetical placement policy where the placement decision is independent of a student's readiness. More specifically, we conceptualize the hypothetical policy as a weighted coin-flip procedure so that the percentage of students placed into a given course is the same for both the implemented and hypothetical policy. We use ϱ_R (and ϱ_{NR}) to denote the SER for ready (and not-ready) students for this hypothetical weighted coin-flip policy.

The ratio $\frac{\mathbf{P}(x, y)}{\mathbf{P}(x)\mathbf{P}(y)}$ can then be interpreted as the relative number of severe errors made by the implemented policy to the number of severe errors made by a hypothetical policy using a weighted coin-flip to place each student. The deviation of this ratio from 1, a metric we define as *the classification power*, is the reduction in the number of severe errors from implementing the given policy instead of a weighted coin-flip policy. Since

the classification power measures the reduction in severe errors, the greater the classification power, the more effective the policy. More specifically, if the classification power is positive then the implemented policy made fewer severe errors than a hypothetical weighted coin-flip policy whereas if the classification power is negative then the implemented policy made more severe errors than a hypothetical weighted coin-flip policy.

We consider the classification power for both ready and not-ready students. We define the classification power of the placement policy for the ready students (CP_R) as

$$CP_R = 1 - \frac{\mathbf{P}(\text{Ready and Not Placed})}{\mathbf{P}(\text{Ready})\mathbf{P}(\text{Not Placed})} = 1 - \frac{SER_R}{\rho_R}$$

and the classification power of the placement policy for the not-ready students (CP_{NR}) as

$$CP_{NR} = 1 - \frac{\mathbf{P}(\text{Not Ready and Placed})}{\mathbf{P}(\text{Not Ready})\mathbf{P}(\text{Placed})} = 1 - \frac{SER_{NR}}{\rho_{NR}}.$$

We offer two interpretations of CP_R (and CP_{NR}). First, CP_R (CP_{NR}) students can be interpreted as the reduction in underplaced (or overplaced students) by the implemented model. For example, if CP_R is 0.20 then 20% of the ready students who would be underplaced by a hypothetical weighted coin-flip policy are no longer misplaced by the implemented policy. If CP_R is –0.20 then 20% of the ready students who would have been correctly placed by the weighted coin-flip policy are misplaced by the implemented policy. Thus, a negative classification power implies an *increase* in the number of severe errors relative to the hypothetical weighted coin-flip policy.

We offer another interpretation: a placement policy with classification power for ready students of CP_R is functionally equivalent to a policy that perfectly identifies the proportion CP_R of ready students and places the rest with a coin-flip where the coin is weighted by the placement rate, r. To illustrate this interpretation, suppose the group of ready students has been divided into two subgroups. For one subgroup, Group A, the placement decisions are made perfectly so that there is no error in placement. For the other subgroup, Group B, the ready students are placed with a weighted coin-flip where r percent of students placed into the target course. Then the error in placing ready students is given by $SER_R = \mathbf{P}(\text{Not Placed}) \cdot \mathbf{P}(\text{Group B})$ since we are placing everyone in Group B with the same placement rate as we place the entire population of students.

This gives the following equality:

$$CP_R = 1 - \frac{SER_R}{\rho_R} = 1 - \frac{\mathbf{P}(\text{Not Placed}) \cdot \mathbf{P}(\text{Group B})}{\mathbf{P}(\text{Not Placed}) \cdot \mathbf{P}(\text{Ready})}$$

$$CP_R = 1 - \frac{\mathbf{P}(\text{Group B})}{\mathbf{P}(\text{Ready})} = \frac{\mathbf{P}(\text{Ready}) - \mathbf{P}(\text{Group B})}{\mathbf{P}(\text{Ready})} = \frac{\mathbf{P}(\text{Group A})}{\mathbf{P}(\text{Ready})}$$

This gives that CP_R is the proportion of ready students in Group A, or the proportion of students ready for the target course who we confidently place at the start of the

placement procedure. A similar argument shows that a placement policy with classification power of CP_{NR} is equivalent to a policy that perfectly identifies the proportion CP_{NR} of students not ready for the target course and uses a weighted coin-flip to place the rest of the not-ready students with the placement rate used to place the entire population of students.

Finally, we define the classification for the overall policy, CP, as a weighted average of CP_R and CP_{NR} where each metric is weighted by the relative proportion of severe errors made by the weighted coin-flip policy. This gives the overall classification power as

$$CP = CP_R \cdot \frac{\rho_R}{\rho} + CP_{NR} \cdot \frac{\rho_{NR}}{\rho} = 1 - \frac{SER}{\rho}.$$

Methodology

The Severe Error Rate

Since the classification power is computed from byproducts of estimating the SER, we begin with an overview of our assumptions and approach for estimating the SER (see Scott-Clayton, 2012 for more details).

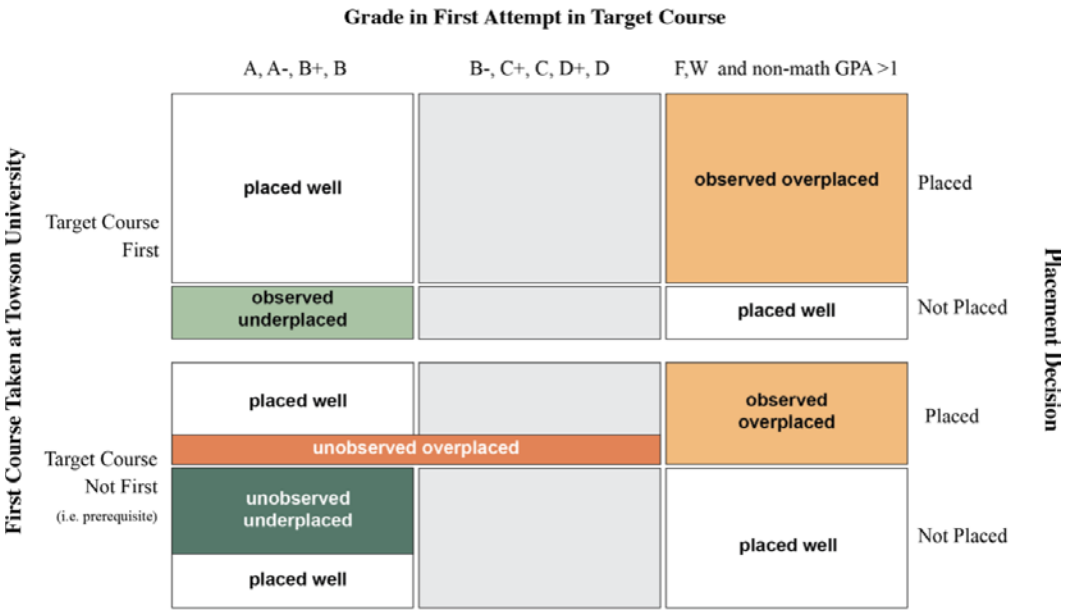
To evaluate and compare placement policies that place students into multiple college-level courses (e.g., Precalculus, Calculus I, etc.), we determine the SER for a single course at a time: we call this the *target course*. Each student who eventually takes the target course as part of their program of study—the *target course population*—was either placed or not placed into the target course when matriculating into the four-year institution. A student is placed into the target course when they have satisfied the criteria for the placement policy—often a qualifying score on a placement test. A student is said to be "not placed" when they do not satisfy these criteria and thus were placed into a lower-level course. Our data consists of institutional records for students in the target course population, including measures of academic preparation (high school GPA, SAT/ACT) as well as demographic factors (race, gender, age).

Aligned with Scott-Clayton's (2012) methodology, we perform a statistical matching between the students who took the target course[2] without taking prerequisites and those who took the target course after a prerequisite course. This allows us to estimate the unobserved severe errors. Figure 1 illustrates the subsets of students severely misplaced, both observed and unobserved, for each target course. We note that at our institution some students do not follow the placement decision resulting in observed underplaced students and unobserved overplaced students.

We explicitly make two theoretical assumptions when estimating the unobserved errors:

2 We are interested in students' first attempt in the target course and do not consider subsequent attempts in the analysis.

Figure 1. Key Subsets of the Target Course Population



1. Consistent with Scott-Clayton's (2012) approach for estimating *severe* errors, we assume a student is *ready* for the target course if they earned (or are predicted to earn) a B or higher on their first attempt in the target course without taking a prerequisite course. We say a student is *not ready* if the student withdrew from the target course or earned (or are predicted to earn) an F on their first attempt in the target course. We do not include students who earned (or are predicted to earn) less than a B but higher than an F in either group. As a result, the proportion of students identified as either *ready* or *not ready* is not 100% of the target course population and these proportions vary across cohorts of students.
2. We assume there is a subgroup of students who fail the target course for nonacademic reasons and that this should not be considered an error in placement. We hypothesize that these students are also likely to fail (or do poorly) in non-mathematics courses during the relevant semester. Thus, we exclude from our study students with a term GPA less than or equal to 1.0 in the semester they took the target course. In each course we studied, approximately 2% of students were removed from the analysis due to their term GPA. This consistency across courses supports our hypothesis that there are factors outside the mathematics course which contribute to students doing poorly in, or withdrawing from, the majority of their classes.

To estimate the probability that a student is severely underplaced we fit a regression model on the group of students who took the target course without prerequisites. The model estimates the probability of a student earning higher than a B (i.e., ready) given a vector of academic and demographic data for the student.

$$\log \frac{p_R}{1 - p_R} = \beta_0 + \beta_1 \text{SAT} + \beta_2 \text{Highschool GPA} + \overrightarrow{\beta_3} \cdot \text{Demographics}$$

We apply this model to the set of students who earned at least a B in the target course after taking a prerequisite to predict whether a student who took a prerequisite course would have been able to earn higher than a B in the target course without the prerequisite. Students with a predicted probability of passing over 50% are classified as underplaced. Together with the students observed to be underplaced, this gives an estimate for SER_R.

To estimate the probability that a student is severely overplaced, we fit a second regression model to predict whether a student who was placed into the target course but chose to take a prerequisite would not have been ready for the target course. Here, we fit the regression model on the group of students who took the target course without prerequisites; the model estimates the probability of a student earning an F in the target course given a vector of academic and demographic data. This regression model is applied to the set of students who earned at least a D in the target course after taking a prerequisite. Students with a predicted probability of failing over 50% are classified as overplaced, allowing us to estimate the number of unobserved overplaced students. Together with the students observed to be overplaced this gives an estimate for SER_{NR}. The overall SER is the sum of SER_R and SER_{NR}.

Estimating the Variation Due to Chance

The methodology described in the previous section produces a point estimate of the SER and byproducts of this computation produce a point estimate for the classification power. We questioned how much variation we would expect to see in these estimates due to chance. In other words, if the characteristics of our student body were slightly different—including small differences in both demographics and academic preparation—would we expect a lot of variation in these measures of the quality of the placement decision? Understanding the variation due to chance would, for example, help policy makers determine whether a difference between point estimates of the classification power is evidence of a meaningful difference in the effectiveness of the placement policy. To quantify the variation due to chance we use bootstrap resampling techniques.

First developed by Efron (1979), bootstrapping is a method to estimate the probability distribution of a statistic from a single sample. Assuming the sample is representative of the population, one generates hypothetical samples from the original data by sampling with replacement so that an individual that is selected for the new sample is returned to the initial sample so they can be selected again. This process is repeated until the resulting sample, *the replicate data set*, has the same length as the initial data set. We note that this methodology is equivalent to conceptualizing the population as a number of copies of the initial representative sample and then sampling from that constructed

population. This methodology provides reliable estimates of the population from a representative sample.

We implemented the bootstrapping methodology by first partitioning the target course population by the placement policy when the student was admitted. This gives, for each target course, four samples where each sample contains all the students who eventually took the target course and were placed under the same placement policy. For each sample, we resampled with replacement to obtain a new replicate data set. For each replicate data set we apply the regression procedure described above: we fit two regression models on the students who took the target course first to predict whether a student who took a prerequisite course was likely prepared (or likely unprepared) for the target course without taking a prerequisite. We computed all relevant quantities (i.e., SER, the percentage of students placed, the percentage of students likely prepared, and percentage of students likely unprepared) for the replicate data set.

We considered both 3,000 and 5,000 replicate data sets. We saw minimal difference between the resulting sampling distributions which suggests that producing 5,000 replicate data sets is large enough to create a stable distribution giving evidence that this distribution is a good approximation of the theoretical sampling distribution. We used the outcomes for each of the 5,000 replicate data sets to compute the probability distributions and 95% confidence intervals of the quantities of interest.

Statistical Tests for Significance

To evaluate the effectiveness of an implemented placement policy, we examine the confidence intervals for the classification power. We deem a policy ineffective if the 95% confidence interval for its classification power includes 0 since a classification power of 0 indicates that the implemented policy provides no improvement compared to a hypothetical weighted coin-flip policy.

To compare the relative effectiveness of multiple placement policies, including different tests and placement procedures (described in the next section), for a given target course, we conduct a two-population t-test between the classification powers for two placement policies being compared. We considered there to be a statistically significant difference between the effectiveness of the two implemented policies if we are able to reject the null hypothesis that the difference between their classification powers is zero.

Institutional Context and Data

To illustrate the utility of the classification power metric in evaluating and comparing placement policies, we evaluate and compare four placement policies implemented at our four-year institution between 2011 and 2020. While we do not intend the results from our institution to be generalizable across institutions, we use our institutional

context to illustrate the relevance of the classification power metric and broader methodology described in this paper. In this section, we briefly describe these placement policies.

Until 2014, mathematics placement at our institution was intended to identify students who needed a developmental math course prior to enrolling in college-level mathematics. Students with an SAT math score above 500 (or equivalent ACT score) were deemed ready for college-level mathematics and identified an appropriate course with their academic advisor. Students with an SAT math score below 500 (or equivalent ACT math score) took a proctored ACCUPLACER exam. If the student earned below 109 on the ACCUPLACER test, then they were placed into a developmental mathematics course. If they earned above 109 on the ACCUPLACER test, then they could enroll in a college-level mathematics course recommended by their advisor. We call this the Developmental Placement Model (DPM).

In Fall 2014 our institution implemented a multilevel placement model intended to simultaneously assess a student's readiness for college-level mathematics and differentiate between a student's readiness for different levels of college mathematics. Under this placement policy all incoming students without college math credit took the Mathematical Association of America's Basic Algebra Test (BAT)—an unproctored online placement test. A cut score was established for each course. For example, students wishing to enroll in Precalculus needed to earn at least a 17 on the BAT. In 2016, in an attempt to improve student's success in their mathematics courses, leaders in our mathematics department increased the cut scores: placement into Precalculus now required a minimum score of 20 on the BAT. We refer to the 2014–2015 placement policy as BATv1, and to the 2016–2017 placement policy as BATv2.

In Fall 2018, our institution adopted the Assessment and Learning in Knowledge Spaces Placement, Preparation, and Learning Test (ALEKS PPL). This is an adaptive test consisting of 25 questions designed to assess students' knowledge on 314 topics in order to place them into courses from Basic Math to Calculus (McGraw Hill, 2024). Cut scores for each course were set based on recommendations from McGraw Hill. Consistent with the previous placement policies, all incoming students without college math credit take the ALEKS placement test in an online and unproctored environment. Students are eligible to retake the ALEKS placement test after spending at least five hours in the associated learning modules. Our data includes students' highest score since that was the score used in the placement decision[3].

3 We recognize that the option to retake the ALEKS test could impact the effectiveness of the policy; we are interested in comparing the effectiveness of each implemented policy and so we do not consider this difference to impact the merits of the analysis.

Data

Our data consists of the academic records for all first-time freshmen who enrolled in Towson University, a large public university in the mid-Atlantic region, between Fall 2011 and Fall 2018. We exclude students who enrolled in/after Fall 2019 due to confounding effects of the COVID pandemic and online teaching. Additionally, we restrict our analysis to first-time freshmen because, for a given year, the placement policy is uniform for this population.

We evaluated and compared the effectiveness of the four placement policies for multiple courses including College Algebra, Precalculus, and Calculus. In the body of the paper, we focus our analysis on evaluating and comparing placement into Precalculus[4], a college-level course at our institution. We restrict our analysis to students who took Precalculus as either their first mathematics course or after taking a prerequisite during their first two years[5] at Towson University.

We focus our discussion on placement into Precalculus for two reasons. First, we recognize that many prior studies evaluating placement focused on placement into college-level and remedial mathematics. At our four-year institution, the placement policy also determines which college-level course students should take. Thus, we select Precalculus to highlight challenges of evaluating placement into a course that has a credit bearing prerequisite (College Algebra). Also, during the data collection period, our institution implemented multiple pathways to satisfy the University's math requirement. As a result, fewer students took Precalculus over time. Thus, we see shifts in demographics and placement rates over time. These differences between cohorts of students will help illuminate the utility of the classification power in comparing placement policies.

Between Fall 2011 and Fall 2018 there were 3,462 students who first took Precalculus in their first two years of study. We partition these students into cohorts based on the placement policy at the time of enrollment for sample size considerations. We have complete data for 3,174 students (about 8% attrition). Table 1 summarizes the demographic information for the Precalculus target course population and Table 2 summarizes measures of academic preparation for this population. While we see some shifts in the demographics of our population—in terms of gender, race, and academic preparation—we do not see these shifts as problematic since our methodology partitions the data into cohorts based on year of enrollment and the characteristics of the sample are more consistent within a cohort[6].

4 We provide results for two additional courses—College Algebra and Calculus—in Appendix B.

5 More than 97% of students who eventually take Precalculus do so in their first two years of study.

6 We include a more detailed demographic summary, organized by year, in Appendix A.

Table 1. Demographic Summary for Precalculus Target Course Population – Portioned by Placement Period

	DPM		BATv1		BATv2		ALEKS	
	F11-S14		F14-S16		F16-S18		F18-S20	
	n	%	*n*	%	*n*	%	*n*	%
Sample Size	1,523		782		577		580	
Female	821	53.91	399	51.02	273	47.31	264	45.52
White	925	60.74	434	55.50	226	39.17	218	37.59
Black	242	15.89	161	20.59	186	32.24	213	36.72
Hispanic	93	6.11	61	7.80	56	9.71	44	7.59
Asian	115	7.55	61	7.80	49	8.49	57	9.83
Other	148	9.72	65	8.31	60	10.40	48	8.28

Note. DPM refers to the Developmental Placement Model and BAT refers to the cohorts of students placed using the Mathematical Association of America's Basic Algebra Test.

Table 2. Summary of Academic Preparation and Placement Rate for Precalculus Target Course Population – Portioned by Placement Period

	DPM		BATv1		BATv2		ALEKS	
	F11-S14		F14-S16		F16-S18		F18-S20	
	n	Avg.	*n*	Avg.	*n*	Avg.	*n*	Avg.
SAT	1425	581.42	711	572.38	512	556.45	534	563.67
ACT	298	23.25	199	23.07	141	21.89	107	21.66
High School GPA	1499	3.60	779	3.57	573	3.53	580	3.64
	n	%	*n*	%	*n*	%	*n*	%
Placed into Precalculus	1371	90	649	83	346	60	348	60

Note. DPM refers to the Developmental Placement Model and BAT refers to the cohorts of students placed using the Mathematical Association of America's Basic Algebra Test.

Results

In this section we provide both the SER and the classification power for each placement policy to illustrate affordances and limitations of analysis based upon the SER and the classification power.

We first present the SER for each of the four placement policies (see Table 3). We see point estimates for the SER between 8% and 16%: estimates comparable to those previously reported in the literature. We interpret these point estimates as the percentage of severe errors made by each placement policy (e.g., 15% of students who need to take Precalculus are severely misplaced by ALEKS). From these results we see that BATv1 has the smallest SER of 8% and we see that two policies have nearly identical SERs (BATv2 and ALEKS). Additionally, we note that 8 percentage points separate the largest and smallest SER.

Table 3. Point Estimates for SER for Placement Into Precalculus

Placement	DPM	BATv1	BATv2	ALEKS
SER	11	8	16	15

Note. DPM refers to the Developmental Placement Model, BAT refers to the cohorts of students placed using the Mathematical Association of America's Basic Algebra Test, and SER refers to the Severe Error Rate Metric.

For these results to inform policy makers, we would need to be able to answer whether (1) an SER of 8% is small enough to provide evidence of an effective policy. In other words, is there a benchmark to compare 8% against? And (2) Can we directly compare the SER across different placement policies that place different cohorts of students? And, if so, (3) is the 8 percentage points that separate these four estimates due to chance or evidence of significant differences between the policies? We provide additional context about the types of errors made by each policy (see Table 4) in order to answer these questions.

In Table 4 we partition the severe errors for ready and not-ready students and we provide the results of using bootstrapping to compute the 95% confidence interval for each metric. We also provide byproducts of estimating the SER: the percentage of students in the target course population who were placed into Precalculus and estimates for the percentage of ready and not-ready students.

Looking at Row 1 of Table 4, we see 95% confidence intervals of the SER with widths between 4 and 9 percentage points. These confidence intervals suggest a single digit improvement in the SER, with other conditions fixed (such as percentage of students placed into the target course), would provide only minimal evidence of an improvement in the quality of placement decision.

Rows 2 and 3 provide point estimates for the number of errors made among ready and not ready students, respectively. We see that DPM made more severe errors among not ready students. This is not surprising since, under this placement model, students who scored above 500 on the SAT (or equivalent on ACT) had no restriction on their first math course. That is, ready students had no restriction on their math course and thus, the policy did not make many errors among the ready students. In contrast, the test-based policies utilizing BATv2 and ALEKS made more severe errors among ready

Table 4. Estimates for SER Metrics for Placement Into Precalculus

Placement	**DPM**		**BATv1**		**BATv2**		**ALEKS**	
Parameter	Est. (%)	CI (%)	Est. (%)	CI (%)	Est. (%)	CI (%)	Est. (%)	CI(%)
SER	11	(9,13)	8	(6.10)	16	(13,20)	15	(11,20)
SER.R	1	(1,2)	3	(2,5)	11	(7,14)	10	(6,15)
SER.NR	10	(8,11)	4	(3,6)	5	(4,8)	5	(2,8)
Placed	90	(89,92)	83	(80,85)	60	(56,65)	60	(54,66)
Ready	45	(42,47)	48	(44,52)	41	(36,46)	34	(28,40)
Not Ready	11	(10,13)	7	(5,9)	12	(9,15)	7	(4,10)

Note. DPM refers to the Developmental Placement Model and BAT refers to the cohorts of students placed using the Mathematical Association of America's Basic Algebra Test. SER refers to the Severe Error Rate Metric while SER.R and SER.NR partition the Severe Error Rate for Ready and Not Ready students, respectively.

students than not-ready students. This suggests that test-based placement might serve as a gatekeeper for ready students.

The last two rows of Table 4 reveal that, at our institution, there is variation in the percentage of students ready (and not ready) for Precalculus across different placement cohorts. This is significant because the total proportion of ready and not-ready students gives an upper bound for the SER and thus, the variation in these estimates impacts our interpretation of the SER and our ability to compare SERs across policies. For example, BATv2 and ALEKS had similar SERs (16% and 15%, respectively). However, under BATv2 53% of the target course population is estimated to be either ready or not ready whereas under ALEKS only 41% of the target course population is estimated to be either ready or not ready. Thus, BATv2 severely misplaced 30% of students estimated to be either ready or not ready while ALEKS severely misplaced over 36% of students estimated to be ready or not ready students. This distinction between the SER and the proportion of misplaced students is even more apparent when considering the subset of not-ready students: BATv2 severely misplaced 41.7% of not-ready students and ALEKS severely misplaced 71.4% of not-ready students. This illustrates the need to consider the number of severe errors made by a policy relative to the characteristics of the sample, including the proportion of ready and not-ready students.

While considering the proportion of severe errors to the number of ready/not-ready students allows us to compare the number of severe errors across placement policies, it does not provide a benchmark for comparison. For example, is severely misplacing 30% of students identified as either ready or not ready evidence of an effective placement policy? Without a benchmark to evaluate against, it is difficult for policy makers to justify policy decisions.

Analysis Based on Classification Power

In the previous section we used empirical data from our institution to highlight limitations of using the SER to evaluate and compare placement policies. In this section we illustrate how using the classification power metric addresses these limitations and provides a more robust tool to evaluate and compare the effectiveness of mathematics placement policies. Table 5 provides point estimates and 95% confidence intervals for the overall classification power (*CP*), the classification power for ready students (CP_R), and the classification power for not-ready students (CP_{NR}).

Table 5. Estimates for Classification Power Metrics for Placement Into Precalculus

Placement	DPM		BATv1		BATv2		ALEKS	
Parameter	Est.(%)	CI(%)	Est.(%)	CI(%)	Est.(%)	CI(%)	Est.(%)	CI(%)
CP	24	(18,31)	46	(34,58)	31	(18,43)	14	(–6,33)
CP.R	75	(61,88)	60	(45,74)	34	(20,49)	24	(3,45)
CP.NR	4	(–2,9)	27	(12,44)	25	(3,43)	–21	(–55,16)

Note. DPM refers to the Developmental Placement Model and BAT refers to the cohorts of students placed using the Mathematical Association of America's Basic Algebra Test. CP refers to the Classification Power Metric while CP.R and CP.NR refer to the Classification Power for Ready and Not Ready students, respectively.

We first focus on Rows 2 and 3 labeled "CP.R" and "CP.NR"; these values can be interpreted as the percentage of students, ready (not ready) for the target course, that were correctly placed (see Theoretical Perspective for more details). For example, for placement into Precalculus using BATv1 CP_R is 60%. We interpret this positive estimate as evidence that the implemented policy is more effective than our benchmark—a hypothetical policy based on a weighted coin flip. More specifically, the placement policy is equivalent to one that correctly places 60% of the students ready for Precalculus into the course and makes the placement decision for the remaining 40% of ready students with a weighted coin-flip that matches the placement rate of the target course population (here, 83%). Similarly, CP_{NR} was 27% meaning that for not ready students, the policy is equivalent to one that appropriately does not place 27% of not-ready students into Precalculus, while the remaining 73% of students not ready for Precalculus are placed with a weighted coin-flip.

We note that in one instance—ALEKS-based placement into Precalculus—the point estimate for CP_{NR} is negative and the associated 95% confidence interval includes 0. We interpret the point estimate of –21% as meaning that the ALEKS-based policy is equivalent to a policy that *incorrectly* placed 21% of not-ready students into Precalculus while the remaining 79% were essentially placed with a weighted coin-flip. This provides evidence that the ALEKS based policy is *worse* than using a weighted coin-flip to place not-ready students.

It is worth noting that for all placement periods, CP_R is greater than CP_{NR}. This suggests that placement policies at our institution are better at identifying ready students than not-ready students even though some policies (BATv2 and ALEKS) make fewer severe errors among not-ready students. This difference between the relative sizes of the SER and the classification power can be attributed to characteristics of the population and the percentage of students ready (and not ready) for Precalculus.

Finally, we consider the overall classification power for each placement period (see Row 1 of Table 5). Across the four placement periods we see classification powers between 14% and 46% suggesting each implemented policy was an improvement over a weighted coin-flip placement. However, under ALEKS-based placement, the 95% confidence interval for the classification power includes 0 which provides evidence that there is no statistically significant difference between ALEKS-based placement into Precalculus and a hypothetical weighted coin-flip placement policy.

Comparing Policies Based on Classification Power

Since both the implemented and hypothetical policies place the same population of students and the classification power measures the improvement of the error rate over the hypothetical policy, one can directly compare the classification power of placement policies implemented with different student populations. In practice, this means that the classification power can be compared across placement policies where a larger classification power indicates a placement policy that provides more information about students' readiness—that is, a better placement policy. We conducted *t*-tests between the classification power for each placement policy and the associated *p*-values for these *t*-tests are provided in Table 6.

Table 6. Comparison of Classification Power for Precalculus Under Different Placement Policies

Placement 1	Placement 2	Difference in CP	*p*-value	Difference in CP.R	*p*-value	Difference in CP.NR	*p*-value
DPM	BATv1	–21.61	0.00	14.87	0.20	–23.79	0.00
DPM	BATv2	–6.35	0.39	40.93	0.00	–20.96	0.01
DPM	ALEKS	10.58	0.29	50.22	0.00	24.12	0.00
BATv1	Batv2	15.26	0.09	26.06	0.01	2.83	0.84
BATv1	ALEKS	32.20	0.00	35.35	0.01	47.90	0.00
BATv2	ALEKS	16.93	0.13	9.29	0.46	45.07	0.02

Note. DPM refers to the Developmental Placement Model and BAT refers to the cohorts of students placed using the Mathematical Association of America's Basic Algebra Test. CP refers to the Classification Power Metric while CP.R and CP.NR refer to the Classification Power for Ready and Not Ready students, respectively.

First, let's return to the comparison of placement based on BATv2 and ALEKS. Recall that these two policies had similar SER, SER_R and SER_{NR}. When comparing the classification powers between these policies (Row 6 in Table 6) we report no statistical difference in the information provided for ready students, but a large difference and a p-value less than 0.05 for CP_{NR} in the amount of information provided about not-ready students. We interpret this as evidence that BATv2 was better able to identify not-ready students than ALEKS. This provides empirical evidence that two policies can make the same number of severe errors while providing different information about students' readiness for a course.

These t-tests also allow us to determine the impact of specific policy changes. For example, between BATv1 and BATv2 our institution increased the cut score required to place into each target course. In Row 4 of Table 6 we report a positive difference and p-values less than 0.1 for both CP and CP_R. We see no statistical difference between CP_{NR}. We interpret this as evidence that increasing the cut score resulted in a policy that provided less information about students' readiness for Precalculus while resulting in no difference in the effectiveness of identifying not-ready students. This supports Leeds and Mokher's (2020) finding that changing the cut score can improve placement accuracy. Additionally, we interpret this as compelling evidence that increasing the cut score served as a gatekeeper for ready students while not improving the policy's ability to identify not-ready students.

Additionally, we can compare across all policies in order to confidently recommend BATv1 over all other implemented policies for placement into Precalculus since the associated p-values for the corresponding t-tests are less than 0.1 (Rows 1, 4, and 5). In this empirical example, BATv1 made fewer severe errors than the other policies (Table 3) while also providing significantly more information about students' readiness (Table 6).

Discussion

We conclude the paper with a discussion of broader implications of using the classification power, instead of SER, for evaluating and comparing placement policies.

Evaluating a Given Placement Policy

The SER has an intuitive interpretation, the proportion of severe errors made by a given policy. And while one expects an effective policy to make few errors, some error should be expected. However, it is unclear how many placement errors should be expected from an effective policy: is severely misplacing 8% of the target course population too many errors? This issue is exacerbated by potential variation we expect in the SER due to chance, as revealed by the widths of the 95% confidence interval for the SER. Our empirical results suggest that the SER may vary by up to 9% due to chance and the specific characteristics

of the enrolled student population. That is, even a small point estimate of the SER might not be representative of the true SER for a given placement policy.

In contrast, the classification power has an inherent benchmark for comparison: a hypothetical weighted coin-flip policy. We expect an effective placement policy to provide more information about students' readiness than this hypothetical policy. In practice, this means if the classification power is not statistically different than 0 (i.e., the 95% confidence interval contains 0), we know we are not better (or worse) than a hypothetical policy using a weighted coin-flip and we say the placement policy is ineffective at identifying students ready (or not ready) for a given course.

Comparing Placement Policies

One might expect that placement policies with similar SERs would also have similar classification powers. Our results provide empirical evidence that this is not necessarily true. The SER is highly dependent on characteristics of the sample, including the percentage of students placed into the target course and the proportion of students ready/not ready for the target course. Thus, directly comparing SERs does not account for potential differences in the cohorts of students. This makes it difficult for policy makers to compare policies at their own institution and near impossible to compare policies across institutions.

In contrast, the classification power takes into account characteristics of the sample by comparing the number of errors made by the implemented policy to a hypothetical policy that places the *same* cohort of students. The resulting metric can then be compared across cohorts of students. This allows policy makers to understand the impact of policy changes at their institution (such as implementing a new placement test or changing the cut score). Attending to characteristics of the sample is also essential in order to compare effectiveness of a given instrument, such as ALEKS PPL, across institutions.

Isolating the Impact of Policy Changes From Variation Due to Chance

A key motivator for our work is to compare the effectiveness of placement policies. Comparing point estimates of the classification power does not attend to variation in the classification power due to chance. Thus, we used bootstrapping techniques to quantify this variation, compute 95% confidence intervals for the relevant metrics, and perform statistical tests comparing the classification powers associated with two placement policies. The results of these *t*-tests allow us to determine whether differences in the classification power is evidence of a significant difference between the policies or due to chance. In other words, through bootstrapping we can isolate the impact of policy changes from variation due to chance.

Our empirical data suggests that even with relatively small sample sizes, we see significant differences between policies' ability to classify ready students and

not-ready students. By attending to the nuances in the placement decision, such as considering separately the placement of ready and not-ready students, policy makers can identify characteristics of policies that might be contributing to its effective (or ineffective) placement of groups of students. For example, at our institution we report that increasing the cut score on the BAT resulted in a policy that was less effective at identifying ready students while resulted in no difference in the effectiveness of identifying not-ready students. We interpret this as evidence that increasing the cut score served as a gatekeeper and restricted ready students' access to Precalculus. This type of analysis can inform policy changes that could result in a more effective policy.

Conclusion

In this paper we present a new approach to quantify the effectiveness of a placement policy—the classification power. While we do not intend the results from our institution to be generalizable across institutions, we see this methodology as relevant for all postsecondary institutions utilizing placement policies to identify students' readiness for coursework be it in mathematics or other disciplines, such as writing. The classification power can be used to (a) quantify the effectiveness of a single-implemented policy, (b) compare the effectiveness of policies that place different cohorts of students, and (c) understand potential impact of policy changes.

The classification power can be used to examine possible improvements in the classification power provided by proctoring of the exam, as well as the improvement of using other measures, such as high school GPA, either in addition to or instead of a placement test score. Since the classification power is independent of the placement rate, this metric allows comparison across courses, across placement policies, and even across institutions. While further study is necessary to validate this claim, in theory, institutions can use classification powers reported by other institutions to help inform their own policy decisions. Finally, we recognize that placement policies have the potential to have inequitable impact on different demographic groups. While the classification power metric is designed to evaluate an implemented (or proposed) policy it can also be used to measure the fairness of a policy for different demographic groups and we intend to pursue this in future studies.

Statements of Declaration

The research presented in this paper was funded by a U.S. Department of Education's First in the World Grant (FITW) awarded to the University System of Maryland. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the U.S. Department of Education or the University System of Maryland.

The data and policy information were provided by Towson University Office of Institutional Research, Academic Achievement Center, Academic Advising Center, and the Department of Mathematics.

All authors certify that they have no affiliations with or involvement in any organization or entity with any financial interest or non-financial interest in the subject matter or materials discussed in this manuscript.

Data Sharing and Documentation

Data files will be shared with other researchers upon request and approval from the Institutional Review Board. R-code will be shared with other researchers upon request.

References

Ayele, A. D., Carson, A., & Tameze, C. (2023). An efficacy study of ALEKS-based placement in entry-level college math courses. *PRIMUS: Problems, Resources, and Issues in Mathematics Undergraduate Studies*, *33*(4), 414–430. https://doi.org/10.1080/10511970.2022.2073623

Barnett, E., Bergman, P., Kopko, E. M., Reddy, V. T., Belfield, C., & Roy, S. (2018). *Multiple measures placement using data analytics: An implementation and early impacts report*. Center for the Analysis of Postsecondary Readiness. https://postsecondaryreadiness.org/wp-content/uploads/2018/09/multiple-measures-placement-using-data-analytics.pdf

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. Community College Research Center, Columbia University. https://ccrc.tc.columbia.edu/publications/predicting-success-placement-tests-transcripts.html

Chen, X. (2016). *Remedial coursetaking at US public 2-and 4-year institutions: Scope, experiences, and outcomes. Statistical analysis report*. NCES 2016-405. National Center for Education Statistics. https://nces.ed.gov/pubs2016/2016405.pdf

Cover, T. M., & Thomas, J. A. (1991). *Elements of information theory*. John Wiley & Sons.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, *7*(1), 1–26. https://doi.org/10.1214/aos/1176344552

Fano, R. M. (1961). *Transmission of information: A statistical theory of communications*. MIT Press.

Fields, R., & Parsad, B. (2012). *Test and cut scores used for student placement in postsecondary education: Fall 2011*. National Assessment Governing Board. https://files.eric.ed.gov/fulltext/ED539918.pdf

Frank, K., Kolesnikov, A., & Wang, X. (2022). Comparing placement policies at a four-year institution. *PRIMUS: Problems, Resources, and Issues in Mathematics Undergraduate Studies*, *32*(10), 1076–1091. https://doi.org/10.1080/10511970.2021.1983903

Leeds, D. M., & Mokher, C. G. (2020). Improving indicators of college readiness: Methods for optimally placing students into multiple levels of postsecondary coursework. *Educational Evaluation and Policy Analysis*, *42*(1), 87–109. https://doi.org/10.3102/0162373719885648

McGraw Hill. (2024). *ALEKS PPL: Placement, preparation and learning*. ALEKS Placement, Preparation and Learning. https://www.mheducation.com/highered/aleksppl.html#accurate

Meuschke, D. M., & Gribbons, B. C. (2003). *English and math placement surveys: Students and instructors–Fall 2002* (Report No. R-130). College of the Canyons, CA. https://files.eric.ed.gov/fulltext/ED480779.pdf

Rodríguez, O. (2014). *Increasing access to college-level math: Early outcomes using the Virginia placement test*. CCRC Research Brief, Number 58. https://ccrc.tc.columbia.edu/publications/increasing-access-to-college-level-math.html

Sawyer, R. (1996). Decision theory models for validating course placement tests. *Journal of Educational Measurement*, *33*(3), 271–290. https://doi.org/10.1111/j.1745-3984.1996.tb00493.x

Scott-Clayton, J. (2012). *Do high-stakes placement exams predict college success?* CCRC working paper no. 41. Community College Research Center, Columbia University. https://files.eric.ed.gov/fulltext/ED529866.pdf

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. https://doi.org/10.3102/0162373713517935

Thornton, D., Case, J., & Peppers, C. (2019). Low-stakes mathematics placement and preparation using EdReady. *Journal of the National College Testing Association*, *3*(1), 1–9. https://www.ncta-testing.org/assets/docs/JNCTA/JNCTA_EdReady_Final.pdf

Appendix A

Table A1. Demographic Data for Precalculus Target Course Population by Year

Placement Cohort	Year	Female		White		Black		Hispanic		Asian		Other	
		n	%	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%
DPM	2011	260	56.28	297	64.29	72	15.58	27	5.84	32	6.93	34	7.36
	2012	293	58.25	304	60.44	76	15.11	35	6.96	36	7.16	52	10.34
	2013	268	48.03	324	58.06	94	16.85	31	5.56	47	8.42	62	11.11
BATv1	2014	231	54.61	233	55.08	90	21.28	36	8.51	29	6.86	35	8.27
	2015	168	46.8	201	55.99	71	19.78	25	6.96	32	8.91	30	8.36
BATv2	2016	142	52.01	126	46.15	80	29.3	23	8.42	17	6.23	27	9.89
	2017	131	43.09	100	32.89	106	34.87	33	10.86	32	10.53	33	10.86
ALEKS	2018	148	48.52	115	37.7	117	38.36	22	7.21	25	8.2	26	8.52
	2019	116	42.18	103	37.45	96	34.91	22	8	32	11.64	22	8

Note. DPM refers to the Developmental Placement Model and BAT refers to the cohorts of students placed using the Mathematical Association of America's Basic Algebra Test.

Table A2. Academic Preparation for Precalculus Target Course Population by Year

Placement Cohort	Year	SAT		ACT		High School GPA	
		n	Avg.	*n*	Avg.	*n*	Avg.
DPM	2011	442	582.08	81	23.15	453	3.62
	2012	464	581.53	101	23.22	497	3.61
	2013	519	580.75	116	23.34	549	3.56
BATv1	2014	392	570.87	94	23.12	420	3.57
	2015	319	574.23	105	23.03	359	3.58
BATv2	2016	238	564.79	72	22.31	271	3.54
	2017	274	549.2	69	21.45	302	3.53
ALEKS	2018	278	563.85	55	21.58	305	3.62
	2019	256	563.48	52	21.75	275	3.67

Note. DPM refers to the Developmental Placement Model and BAT refers to the cohorts of students placed using the Mathematical Association of America's Basic Algebra Test.

Appendix B

Table B1. Estimates for SER and Classification Power for Placement Into Calculus

Placement	DPM		BATv1[1]		BATv2		ALEKS	
Parameter	Est.(%)	CI(%)	Est.(%)	CI(%)	Est.(%)	CI(%)	Est.(%)	CI(%)
SER	24	(21,29)	21	(17,25)	17	(14,20)	20	(15,26)
SER.R	0	(0,1)	4	(3,7)	3	(2,5)	6	(3,10)
SER.NR	24	(21,29)	16	(13,20)	13	(10,17)	14	(10,19)
CP	7	(2,11)	21	(9,30)	33	(22,44)	15	(–4,32)
CP.R	89	(25,100)	47	(27,66)	61	(43,76)	36	(9,61)
CP.NR	3	(0,7)	9	(–2,19)	18	(6,30)	1	(–21,21)
Placed	94	(92,95)	66	(62,70)	63	(59,67)	47	(41,53)
Ready	19	(16,22)	24	(20,29)	24	(20,28)	18	(13,24)
Not Ready	26	(23,31)	27	(23,31)	26	(22,30)	29	(24,36)

Note. DPM refers to the Developmental Placement Model and BAT refers to the cohorts of students placed using the Mathematical Association of America's Basic Algebra Test. SER refers to the Severe Error Rate Metric while SER.R and SER.NR partition the Severe Error Rate for Ready and Not Ready students, respectively. CP refers to the Classification Power Metric while CP.R and CP.NR refer to the Classification Power for Ready and Not Ready students, respectively.

1 Under BATv1 and BATv2, students who wanted to take Calculus 1 needed to earn a qualifying score of 17 on the Basic Algebra Test (BAT) and a qualifying score on the Calculus Readiness Test (CRT). Under BATv1 the CRT cut score for Calculus 1 was 11 and the CRT cut score increased to 15 under BATv2.

Table B2. Comparison of Classification Power for Calculus 1 Under Different Placement Policies

Placement 1	Placement 2	Difference in CP	*p*-value	Difference in CP.R	*p*-value	Difference in CP.NR	*p*-value
DPM	BATv1	–13.29	0.02	42.68	0.04	–5.57	0.24
DPM	BATv2	–25.94	0	28.26	0.11	–15.03	0
DPM	ALEKS	–7.79	0.15	53.43	0.08	2.49	0.82
BATv1	BATv2	–12.64	0.11	–14.42	0.24	–9.46	0.25
BATv1	ALEKS	5.5	0.66	10.75	0.57	8.06	0.49
BATv2	ALEKS	18.14	0.09	25.17	0.11	17.52	0.14

Note. DPM refers to the Developmental Placement Model and BAT refers to the cohorts of students placed using the Mathematical Association of America's Basic Algebra Test. CP refers to the Classification Power Metric while CP.R and CP.NR refer to the Classification Power for Ready and Not Ready students, respectively.

Table B3. Estimates for Placement Metrics for Placement Into College Algebra, a College-Level Math Course at Towson Institution

Placement	DPM		BATv1		BATv2		ALEKS	
Parameter	Est.(%)	CI(%)	Est.(%)	CI(%)	Est.(%)	CI(%)	Est.(%)	CI(%)
SER	12	(10,14)	7	(6,9)	10	(9,12)	13	(11,15)
SER.R	8	(6,10)	4	(3,6)	6	(5,8)	6	(5,8)
SER.NR	4	(3,5)	3	(2,4)	4	(3,5)	6	(5,8)
CP	24	(13,35)	35	(23,45)	26	(17,34)	31	(20,40)
CP.R	31	(18,45)	40	(27,53)	29	(18,39)	39	(26,52)
CP.NR	5	(−6,18)	23	(9,39)	20	(9,32)	19	(6,31)
Placed	81	(79,83)	89	(87,91)	83	(81,85)	75	(72,78)
Ready	60	(57,63)	66	(63,69)	54	(51,56)	42	(39,46)
Not Ready	6	(4,7)	4	(3,5)	6	(5,7)	10	(8,13)

Note. DPM refers to the Developmental Placement Model and BAT refers to the cohorts of students placed using the Mathematical Association of America's Basic Algebra Test. SER refers to the Severe Error Rate Metric while SER.R and SER.NR partition the Severe Error Rate for Ready and Not Ready students, respectively. CP refers to the Classification Power Metric while CP.R and CP.NR refer to the Classification Power for Ready and Not Ready students, respectively.

Table B4. Comparison of Classification Power for College Algebra Under Different Placement Policies

Placement 1	Placement 2	Difference in CP	*p*-value	Difference in CP.R	*p*-value	Difference in CP.NR	*p*-value
DPM	BATv1	−10.95	0.12	−9.37	0.27	−17.77	0.06
DPM	BATv2	−2.21	0.73	1.82	0.79	−14.39	0.09
DPM	ALEKS	−7	0.34	−8.49	0.35	−13.29	0.14
BATv1	BATv2	8.74	0.19	11.19	0.16	3.38	0.69
BATv1	ALEKS	3.95	0.55	0.88	0.89	4.48	0.63
BATv2	ALEKS	−4.79	0.46	−10.32	0.21	1.09	0.95

Note. DPM refers to the Developmental Placement Model and BAT refers to the cohorts of students placed using the Mathematical Association of America's Basic Algebra Test. CP refers to the Classification Power Metric while CP.R and CP.NR refer to the Classification Power for Ready and Not Ready students, respectively.