Are Algorithms Biased in Education?
Exploring Racial Bias in Predicting Community College Student Success

Kelli A. Bird (University of Virginia)
Benjamin L. Castleman (University of Virginia)
Yifeng Song (University of Virginia)

Abstract

Predictive analytics are increasingly pervasive in higher education. However, algorithmic bias has the potential to reinforce racial inequities in postsecondary success. We provide a comprehensive and translational investigation of algorithmic bias in two separate prediction models--one predicting course completion, the second predicting degree completion. We show that if either model were used to target additional supports for "at-risk" students, then the algorithmic bias would lead to fewer marginal Black students receiving these resources. We also find the magnitude of algorithmic bias varies within the distribution of predicted success. With the degree completion model, the amount of bias is over five times higher when we define at-risk using the bottom *decile* than when we focus on students in the bottom *half* of predicted scores; in the course completion model, the reverse is true. These divergent patterns emphasize the contextual nature of algorithmic bias and attempts to mitigate it. Our results moreover suggest that algorithmic bias is due in part to currently-available administrative data being relatively less useful at predicting Black student success, particularly for new students; this suggests that additional data collection efforts have the potential to mitigate bias.

Acknowledgements

We are very grateful to Dr. Cat Finnegan and the team at the Virginia Community College System for their collaboration on this project. We also greatly appreciate feedback from Eric Taylor and two anonymous reviewers. Any remaining errors are our own.

INTRODUCTION

Predictive analytics are increasingly pervasive in higher education. By one estimate, nearly 40 percent of higher education institutions now use some form of predictive analytics (Barhsay & Aslanian, 2019; Ekowo & Palmer, 2016, Swaak, 2022). Institutions apply predictive analytics across a broad array of student services, from directing scholarships to students who are predicted to persist at the institution, to allocating additional academic supports or proactive advising to students predicted to struggled in individual courses or in college overall (Barhsay & Aslanian, 2019; Ekowo & Palmer, 2016; Paterson, 2019; Smith, Lange, and Huston, 2012; Treaster, 2017). The use of predictive analytics in early-alert systems, which flag potentially struggling students within a course, particularly swelled during the COVID-19 pandemic, with over 80 percent of public colleges using some form of this technology (Ogundana & Ositelu, 2022). However, concerns about potential racial bias in prediction algorithms raise the question as to whether predictive analytics could negatively impact colleges' and universities' broader efforts to promote greater racial equity.

In recent years, the data science research community has explored algorithmic bias in education contexts, and in nearly all instances researchers demonstrate the presence of algorithmic bias when predicting student success (for an extensive review, see Baker & Hawn, 2021).[1] For example, some papers find lower model accuracy for underrepresented minority (URM) groups; this lower accuracy leads to more URM students having a mis-classified risk status (Lee &

[1]Other recent research examines algorithmic bias in other public policy settings. Obermeyer et al. (2019) demonstrate that a commercial algorithm used to enroll patients in a high-risk healthcare management program is less likely to identify sick Black patients compared with equally sick White patients. Angwin et al (2016) and Corbett-Davies et al (2017) find that the COMPAS algorithm, which predicts future crime for court defendants, assigns higher risk to Black defendants who have the same actual reoffense rate as similar White defendants, resulting in undue harsher pre-trial or sentencing decisions for Black defendants. Arnold, Dobbie, and Hull (2021) find similar bias in an algorithm used in New York City courtrooms.

Kizilcec, 2020, Riazy, Simbeck, and Schreck, 2020, Sha et al, 2022). Other papers find that algorithms are more likely to predict URM students will struggle or fail when they in fact succeed, while the reverse is true of White and Asian students (Anderson, Boodhwani, & Baker, 2019, Jiang & Pardos, 2021, Jeong et al, 2021, Yu, Lee, & Kizilcec, 2021, Yu et al, 2020). This algorithmic bias has important implications for educational policy and practice, since it could result in inefficient targeting of resources. To date, however, this line of work has primarily targeted data science communities rather than policy makers or education researchers, and has focused more on investigating technical aspects of model development (e.g. exploring novel methods for bias mitigation or applying post-prediction adjustments to risk scores) than on translating policy- and practice-relevant insights.

In this paper, we build on the existing line of research on algorithmic bias in education by providing a comprehensive *and* translational investigation of two separate prediction algorithms--the first predicting course-level completion, and the second predicting overall degree completion. We developed these models using detailed student-level data from the Virginia Community College System (VCCS); as we detail in the next section, our models apply a similar data science methodology (random forests) and incorporate similarly rich student-level data as those currently used at many colleges and universities.[2] We conduct our investigation of algorithmic bias in the context of the most common use-case of predictive analytics: To identify students with a lower likelihood of completing a course or dropping out; label these students as "at-risk" and to target

[2] The codebase for our models is publicly available: https://github.com/nudge4/Bias-In-Predictive-Modeling/. We hope that an ancillary benefit of this analysis is that colleges and universities can reference this code to predict student success without having to develop their own prediction models from scratch or pay for commercial products, which can cost hundreds of thousands of dollars per year (Barshay and Aslanian, 2019). For instance, we are currently collaborating with Piedmont Virginia Community College to use the course completion model to guide which students the instructors will target for additional support in large "gateway" math and English courses.

additional resources (e.g. advising, tutoring) to these students.[3] Therefore, we intentionally focus our analysis and discussion on the types of algorithmic bias that would lead to fewer resources being allocated to students from historically-disadvantaged groups. Specifically, we focus on algorithmic bias along two dimensions: calibration and accuracy. Calibration bias occurs when different student subgroups have different actual success rates, conditional on predicted risk scores. To assess accuracy bias, we compare the models' c-statistics (also known as AUC) across subgroups. We then explore multiple hypotheses regarding why these forms of algorithmic bias would be present in our models.

Our analysis provides four main takeaways. First, and consistent with prior research, we find evidence of algorithmic bias (for both calibration and accuracy) in the models we investigate. In both the course completion model and degree completion model, we find that among students with the same low predicted success rate at the beginning of the term, Black students have lower actual success rates than White students. If scarce resources were allocated to students identified as "at-risk" based on the biased models, then Black students would receive fewer of these resources compared to an allocation that was based on actual (but unobservable) risk. We further show that this calibration bias is more pronounced at certain points in the distribution of predicted success. Specifically, the degree completion model exhibits substantially higher bias among the population of students with the lowest predicted likelihood of success: The amount of bias is over five times higher when we set a "risk threshold" using the bottom *decile* than when we set a risk threshold using the bottom *half* of predicted scores. However, we also show that this pattern differs quite

[3]An important nuance to consider is the difference between the latent risk level at the relevant time period when the at-risk label is assigned, versus the time of observing the eventual outcome. When colleges label students as at-risk, they implicitly assume that students in the historical sample who eventually did not complete a course or program were at-risk at the time when the predictors were measures (i.e. at the beginning of a particular term), and conversely that students who did complete were not at-risk. We discuss this point, and in particular how it may impact the interpretation of algorithmic bias, in the Results section below.

meaningfully between the two models. For the course completion model, we see the reverse pattern: The amount of bias is over five times higher for a risk threshold set at the bottom half versus the bottom decile.

Second, we find that making the course completion model more attuned to race (by either including racial predictors or estimating race-specific models) decreases algorithmic bias without any meaningful reduction in model performance, while the opposite is true for the degree completion model. This result again highlights the highly contextual nature of algorithmic bias: Despite the similarities of the two models (i.e. using the same student population, similar predictors constructed from the same data source, and developed by the same team of researchers), the inclusion or exclusion of race in the model has divergent implications for reducing bias.

Third, we do not find evidence that data underrepresentation or differential student sorting significantly contribute to the algorithmic bias in our models. However, our results *do* suggest that Black students having shorter enrollment histories is a contributing factor to algorithmic bias in the models. Specifically, the amount of algorithmic bias within the sample of first-time VCCS students is typically more than double that of the bias within the sample of returning students. This finding suggests that the additional predictors available for returning students partially mitigates algorithmic bias. Finally, because our models consistently have lower levels of accuracy and generally lower values of other goodness-of-fit metrics for the Black sample, our results suggest that the data currently collected in college administrative systems may be inherently less effective at predicting success for Black students.

Our paper makes several important contributions. First, our primary focus is on generating insights that are relevant for policy makers, practitioners, and researchers who are less engaged with the data science community. Our approach is in line with recent perspectives on algorithmic

bias in economics, which encourage researchers to focus on the *marginal* implications of the algorithmic "outputs" (i.e. the predicted scores and in what context they are used), instead of focusing attention on the technical inputs of the modeling process (Kleinberg et al, 2018; Cowgill & Tucker, 2019; Rambachan et al, 2020). In so doing, the implications of our results differ meaningfully from existing research among data scientists investigating algorithmic bias in education. Prior studies conclude that algorithms are *overall* more "pessimistic" about Black student performance and relatively underestimate their likelihood of success (Anderson, Boodhwani, & Baker, 2019, Jiang & Pardos, 2021, Jeong et al, 2021, Yu, Lee, & Kizilcec, 2021, Yu et al, 2020), which would lead to more Black students being targeted for additional resources.[4] However, our focus of quantifying bias *on the margin* of being labeled at-risk reverses this takeaway: We find that conditional on predicted score, our algorithms relatively overestimate Black students' likelihood of success.[5] Second, we investigate how algorithmic bias differs across the distribution of predicted scores, and in so doing, we show the amount of algorithmic bias can differ substantially based on which particular segment of the student population an educational institution may choose to focus on for intervention. Third, we reinforce the importance of context-dependent investigations of and efforts to mitigate bias. Even holding constant the same sample, data source, and predictor structure, we find different magnitudes of bias overall and at particular points in the distribution of predicted student success depending on which prediction model we investigate, as well as differential efficacy of strategies to mitigate bias.

[4] Specifically, these related prior studies focus on a different definition of algorithmic bias (also referred to as "notions of fairness" in this literature). These studies tend to focus on the "equalized opportunity" notion of fairness, in which an algorithm is considered biased if the true or false positive rates are different across groups. We discuss in more detail below how choosing how to define algorithmic bias in a specific context can significantly impact the interpretation of the practical implications of the bias.

[5] This alternative interpretation is not due to contextual differences between the algorithms we investigate versus the algorithms that are the subjects of prior research. As we show below, if we use the same bias metrics as the previous papers (namely, the false positive rates), then we would also conclude that the models were underestimating Black students' likelihood of success.

Finally, whereas prior work has focused mainly on investigating the technical aspects of algorithmic bias in education, we provide an in-depth exploration of the potential sources of algorithmic bias in our models. As Obermeyer et al. (2019) demonstrate in a health-care setting, understanding the source of algorithmic bias can be paramount to effectively and efficiently reducing bias from an existing algorithm, and to inform future data collection and modeling efforts.

HIGHER EDUCATION INSTITUTIONS' USE OF PREDICTIVE ANALYTICS

To contextualize our analysis of algorithmic bias in higher education and to motivate the specific models that we investigate, we provide a brief overview of how predictive analytics are used by college and universities--both how they develop the prediction models and then how colleges and universities typically leverage the resultant predictions in student outreach and support.

Roughly two-thirds of prediction algorithms used in higher education are "home-grown" models developed by individual institutions (Educause, 2018). The balance are offered by commercial vendors. Both algorithmic approaches typically follow a similar high-level format, while customizing to institution's available data, preferences, and context: They rely on historical student-level data including both predictors (e.g. prior GPA, enrollment intensity) and the ultimate outcome; in turn, the algorithms use the same set of predictors for a new sample of students to predict their likelihood of achieving the outcome. We highlight two notable applications that provide more detailed documentation on the prediction modeling process. First, the University of Oregon is currently predicting first-year student retention using XGBoost (Greenstein et al, 2023), which is a similar tree-based model to random forest and which yields predictions with similar rates of accuracy in the context of predicting college completion (Bird et al, 2021). Recent FOIA requests revealed that EAB's Student Success Predictive Model is "a combination of several

penalized logistic regression models applied to different subgroups," where colleges work with EAB to decide what student information can be introduced as model predictors (Feathers, 2021).

There are several recent case studies that illustrate how dozens of institutions have incorporated predictions from these algorithms into practice (APLU, 2016; Burke et al, 2017; Ekowo & Palmer, 2016; Paterson, 2019; Stark, 2015; Treaster, 2017). A common theme across these case studies is that institutions are looking to predictive analytics to more efficiently target scarce student resources. This incorporation of predictive analytics into resource allocation occurs across the student life cycle and across numerous functions. For instance, enrollment managers leverage predictions of which students are likely to apply to or choose to matriculate at an institution to guide their marketing and recruitment efforts; financial aid offices use predictive analytics to guide aid disbursements to students predicted to persist at the institution (Ekowo & Palmer, 2016; Treaster, 2017). The most prevalent use of predictive analytics reported in these case studies is to guide academic supports to students who might otherwise struggle to succeed, either in an individual course or in the overall program of study. One prominent institution using predictive analytics is Georgia State University GSU, which partnered with EAB to develop an algorithm predicting student academic performance. GSU targeted students predicted to struggle with proactive outreach from advisors and additional academic supports (Ekowo & Palmer, 2016). Other institutions, such as the University of North Carolina-Greensboro, differentiate the intensity of academic intervention based on students' predicted level of risk (Klempin, Grant, & Ramos, 2018).

DATA AND METHODS

The data for this study come from VCCS system-wide administrative records from the Summer 2007 through Fall 2019 academic terms. These records include detailed information about each term in which a student enrolled, including their program of study, courses taken, grades earned, credits accumulated, financial aid received, and degrees earned. The records also include basic demographic information, including gender, race, and parental education. Finally, we observe all degrees and certificates awarded by VCCS colleges beginning in 2007. In addition to VCCS administrative records, we also have access to National Student Clearinghouse graduation and enrollment records. National Student Clearinghouse data allow us to observe all enrollment periods and postsecondary credentials earned at non-VCCS institutions from 2004 onward.

We build two separate prediction algorithms using the VCCS administrative data. The first predicts course-level completion ("course completion model") and the second predicts completion of a degree or certificate ("degree completion model"). We use random forest for both prediction models. Random forest (RF) is a decision-tree based ensemble model commonly used in predictive analytics. In similar contexts of predicting degree completion (Bird et al, 2021) and course performance (Kung & Yu, 2020), recent studies have found very similar levels of overall model performance when comparing RF to other modeling strategies, such as logistic regression, support vector machines, or recursive neural networks.[6] The output from RF is a raw predicted score ranging from 0 to 1, with lower values corresponding to lower predicted likelihood of success.[7]

[6] We also built versions of our base models using logistic regression instead of random forest, and find that random forest has slightly higher overall performance: The RF course completion models' c-statistics are 3 percent higher than the logistic version, and the RF degree completion models' c-statistics are 1 percent higher than the logistic version. (see Appendix Table A1). We also find that the amount of calibration bias is typically lower for the RF models compared to logistic regression (see Appendix Figure A1).

[7] For the vast majority of our analysis, we convert the raw predicted scores percentiles. There is only one piece of our analysis where we need to convert the raw predicted scores to binary predictions: for calculating the True Negative Rate. For this analysis, we set a threshold such that the share of students with a "positive" prediction to be equal to the overall success rate of the training sample (75.4 percent for course completion, 34.0 percent for degree completion).

Using standard predictive analytics practices, we split the data into training and validation samples. We divide the full data such that the training sample includes observations from earlier historical cohorts, and the validation sample includes those from more recent historical cohorts.[8] We use the training sample to build the model and select the optimal parameters (number of decision trees, maximum depth, and number of random predictors to include at each node for splitting) using five-fold cross-validation. This process functions as feature selection, which reduces model overfitting (Ghojogh and Crowley, 2019; Breiman 2002). We provide the optimized values of these parameters for our base models in Appendix Table A2. Finally, we compute a feature importance (FI) score for each predictor using the mean decrease in impurity method (Breiman, 2002). A predictor with a higher FI score makes a larger overall contribution to the model generating the predictions. Because FI scores do not provide precise magnitude comparison of predictors to each other (i.e. a predictor with an FI score that is double another predictor is not necessarily making twice as large a contribution to model accuracy), we instead focus on the relative rankings of predictors based on FI scores.

While both models rely on the same VCCS administrative data, each has a unique outcome, sample, and set of predictors, which we describe below.

Course Completion Model

The binary outcome for the course completion model is equal to one if the student earned a grade of A, B, or C, and equal to zero for grades of D, F, or W.[9] Based on this definition, 75.6 percent

[8] This division best simulates how these models would be applied in practice, where predictions for current students are based on models built from recent historical cohorts. If we instead split the data randomly between the training and validation sample, then the evaluation results would be based on a hold-out sample of observations from the same time frame as the training sample, making the evaluation results less generalizable to the typical application of these models.

[9] While a grade of D earns the student credit for the course and is technically considered a passing grade, students cannot satisfy some VCCS program requirements with a D, and other colleges and universities typically do not accept transfer credit for D grades.

of student x course observations achieved the outcome of course completion. Separating the sample by race subgroups, 79.8 percent a White student x course observations achieved the outcome, compared to 69.4 percent of Black student x course observations.

The course completion sample consists of student x course observations from Spring 2014 to Fall 2019. We restrict the sample to focus on college-level coursework for regularly-enrolled students. Specifically, we exclude dual-enrollment observations (i.e. current high school students taking courses through an arrangement between their high school and a VCCS college). We also exclude observations outside the traditional A-F grading scale.[10] The resulting sample size is 5,168,903 student-by-course observations. We split the sample into training and validation sets based on term: The training sample consists of Spring 2014 to Fall 2018 (n = 4,414,694) and the validation sample consists of Spring 2019 to Fall 2019 (n = 754,209).

We construct 186 predictors from the VCCS administrative data. We describe them at high-level here, and include a full list in Appendix Table A3. The predictors include student demographic and socio-economic information (e.g. age during target term, gender, median income of households in zip code); student academic history at VCCS, prior to the target term, both generally (e.g. prior credits earned, cumulative GPA) and specific to the target course (e.g. the GPA in all the target course's prerequisites); student enrollment characteristics of the target term (e.g. program of study currently pursuing, enrollment intensity); characteristics of the target course (e.g. course enrollment, average grade from the most recent five years); and instructor characteristics (e.g. tenure, full-time versus adjunct). If the observation is for a student's first term at VCCS, we are only able to construct predictors for demographics and socio-economic status, student enrollment characteristics of the target term, characteristics of the target course, and

[10] The vast majority of these observations correspond to developmental courses, which are graded as pass or fail.

instructor characteristics. We handle missing values for predictors related to students' academic history (e.g. prior cumulative GPA is missing for students in their first term of enrollment) by replacing with the value of m-1, where m is the minimum value of the predictor in the study sample; in this way, the RF model is able to distinguish which observations were missing values for each predictor.[11]

Degree Completion Model

The binary outcome for the degree completion model is equal to one if the student completed a VCCS degree or credit-bearing certificate within 6 years. Based on this outcome definition, 34 percent of students in our full sample completed a degree (39.4 for White students, and 22.4 percent for Black students). Our sample consists of students who enrolled at a VCCS college as a degree-seeking (or certificate-seeking), non-dual enrollment student for at least one term, with an initial enrollment term between Summer 2007 and Spring 2013. For each student in our sample, we observe their information for the entire 6-year window after their initial enrollment term. While we use the full 6 years of data to construct the outcome measure, we construct the model predictors to resemble the population of students currently enrolled--many of whom have only been enrolled in one or two terms. Therefore, we randomly truncate the data in the full sample to resemble the distribution of enrollment lengths among Spring 2019, Summer 2019 and Fall 2019 enrollees.[12] We split the sample into training and validation sets based on the term of students' first enrollment: The training sample consists of all students who first enrolled at VCCS from Summer 2007 to Spring 2012 (n = 323,182), while the validation sample consists of students who first enrolled at VCCS from Summer 2012 to Spring 2013 (n = 62,618).

[11] Appendix Figure A2 shows the distribution of predicted scores from the validation sample for the course completion model (Panel A) and the degree completion model (Panel B), separately by racial subgroup.
[12] For a full discussion of the choice of outcome and this truncation method, see Bird et al (2021).

We construct 255 predictors from the VCCS administrative data. We describe them at high-level here, and include a full list in Appendix Table A4. The predictors include student demographic and socio-economic information (e.g. age at initial enrollment, gender, median income of households in zip code); student overall VCCS academic history, measured during the most recent (truncated) term (e.g. cumulative GPA, trend of term-level GPA); student overall non-VCCS academic history (e.g. ever enrolled in non-VCCS college before initial VCCS enrollment, number of non-VCCS terms enrolled); student financial aid receipt history at VCCS (e.g. average student loan borrowing across terms); and term-specific academic information for both VCCS and non-VCCS enrollment (e.g. share of credits a student withdrew from in their first Fall term, and separately the share of credits a student withdrew from in their second Spring term).[13] We handle missing values in the same manner as the course completion model described above.

Measuring Algorithmic Bias

The data science research community has a burgeoning line of work describing different types of algorithmic bias (see, for example, Chouldechova & Roth, 2018).[14] However, there is no consensus on which type of algorithmic bias is most important to address, and in fact, attempts to reduce a particular type of bias often comes at the expense of increasing a different form of bias (Kleinberg, Mullainathan, & Raghavan, 2016). Combining the need to assess algorithmic bias based on specific contextual factors (Paulus & Kent, 2020) and recent perspectives from the field of economics suggesting that examining bias based on the algorithm's outcomes instead of inputs or functional form is the most relevant and actionable (Cowgill & Tucker, 2019), we focus our

[13] Note that we do not include non-VCCS or financial aid predictors in the course completion model because during early model development phases, we found that these predictors did not contribute to the performance of the course completion model.

[14] Algorithmic bias is a term often used interchangeably with algorithmic fairness; this literature also often refers to "notions of fairness" when describing different types of algorithmic bias.

investigation on algorithmic biases that would result in less resource allocation to students from historically-disadvantaged groups who would otherwise not succeed.

Specifically, we focus on two related but distinct forms of algorithmic bias. First, holding constant the algorithm's predictions, do we observe the same actual success rates for Black students versus White students? This is referred to as "calibration bias." For an illustrative example of the consequences of calibration bias, suppose there is a group of students who were all assigned the same low predicted score of 0.2, indicating the students are at-risk for not completing the course or degree. Now suppose that among this group of students, the actual observed success rate of Black students is 10 percent while the actual observed success rate of White students is 30 percent. This pattern would indicate that conditional on predicted score, Black students are actually less likely to succeed academically than White students. This calibration bias would result in some Black students not receiving the additional resources, even though they are actually less likely to succeed than some White students who did receive the resources. Following Obermeyer et al. (2019), we quantify the amount of calibration bias following these steps: (1) Select all White and Black students whose predicted scores are below the "risk threshold" (e.g. a student whose predicted score is below the 30th percentile of the predicted scores); (2) if the actual success rate of the labeled at-risk Black students is lower than that of the labeled at-risk White students, change the label of the Black student whose predicted score is lowest among the Black students not labeled at-risk , and simultaneously drop the White student whose predicted score is highest among the White students labeled at-risk; (3) repeat step two until the actual success rates of the two selected groups become equal. We then compare the numbers of Black to White students who would be

labeled at-risk by the algorithm both before and after the simulated removal of the calibration bias.[15]

The second form of algorithmic bias we consider is whether algorithms are equally accurate at predicting success for Black students as they are for White students. To assess this "accuracy bias," we compare subgroup-specific c-statistics.[16] Lower levels of model accuracy imply that more students would be mis-labeled as at-risk or not.

The base versions of our models are built using the full training and validation sets, and do not include racial predictors.[17] In order to test how incorporating more information about race in the models impacts algorithmic bias, we also estimate versions of the model that include race as predictors, and separately we estimate race-specific models (i.e. we build a random forest model for the White subgroup, and a separate random forest model for the Black subgroup) to allow the determinants of success to differ between White and Black students.[18] For calculating the amount

[15] While this simulation exercise allows us to quantify the amount of calibration bias for a particular threshold of predicted scores, it is important to note that this technique cannot be used to mitigate bias in real-time, because it relies on knowing both the students' predicted scores and their eventual outcomes.

[16] C-statistic, also referred to as the area under the curve (AUC), is a "goodness of fit" measure that is equal to the probability that a randomly selected positive observation (i.e. a student who passed a particular course) has a higher predicted score than a randomly selected negative observation. A c-statistic of 0.5 corresponds to a model being no better than choosing at random, while a c-statistic of 1 corresponds to a model perfectly predicted the outcome. A c-statistic of 0.8 or higher is considered strong performance; and a c-statistic of 0.9 or higher is considered outstanding (Hosmer, Lemeshow, and Sturdivant, 2013). We compute the standard errors of the c-statistic based on the fact that the c-statistic in this case is equivalent to the statistic used in Wilcoxon rank-sum test based on predicted scores (Hanley and McNeal, 1982).

[17] One line of reasoning is that simply removing race as information the model can draw on should eliminate any potential biases. However, as we demonstrate below and as has been previously demonstrated in the data science literature, "race-blind" models may still contain relevant algorithmic bias (e.g. Yu, Lee, & Kizilcec, 2020). This often occurs because there are other predictors in the model which are highly correlated with race. Appendix Table A5 shows the differences between the White and Black subgroups for the top 10 predictors of each model; all of these differences are statistically significant with $p < 0.01$. For a specific example, Black students have 12-21 percent lower cumulative GPAs, compared to White students. As a result, even absent including race indicators in our model, we would still expect significant differences in the distribution of predicted scores between Black and White students. In Appendix Table A6, we provide further race-specific summary statistics separately for the training and validation sets. Overall, Black students are older, more female, have lower values on the socio-economic measures, and have worse academic outcomes. The differences between the training and validation samples reflect the temporal division of the sample on this line.

[18] When fitting RF models on the full training sample of students, the tree-growing procedure seeks to identify the optimal predictor and the corresponding threshold to make each node splitting in order to separate the observations with success from nonsuccess. Because a larger share of the sample are White students, then these node-splitting

of calibration bias in the race-specific models, we first combine the raw predicted scores from the race-specific models to set a common distribution of predicted scores. In other words, the 10th percentile in the Black-specific model corresponds to the same raw predicted score as the 10th percentile in the White-specific model.

Comparing Complex and Simple Algorithms

We recognize that many colleges may be interested in the data-driven approaches to targeting resources, but do not have the means to employ advanced prediction models. These colleges may choose a very simple algorithm instead, such as labeling students as at-risk if they have a cumulative GPA below a certain threshold (e.g. 2.0, which often is used to determine whether students are making Satisfactory Academic Progress to maintain financial aid eligibility). To supplement our results for the random forest models, we also show the potential racial inequities that could result from using a simple targeting strategy.

For predicting both course and degree completion, we focus on two separate predictors: cumulative GPA and enrollment intensity. As we show in Appendix Table A5, cumulative GPA is the most important predictor in the degree completion model, and the second most important predictor in the course completion model. However, this predictor is only available for students with prior VCCS enrollment history. Enrollment intensity is the most important student-specific academic predictor in both models that is populated for first-time students--it is the overall most important predictor for the course completion model (as seen in Appendix Table A5), and the

decisions might be more reflective of the determinants of success of White students, which could be different from the patterns that are specific to Black students. When we estimate race-specific versions of the RF models, we allow the node-splitting decisions to differ across racial lines. Furthermore, we explore whether the determinants of success differ between Black and White students by comparing the top 10 predictors from the race-specific models, as determined by the FI score. Appendix Table A7 shows that two race-specific course completion models share 9 of the top 10 predictors; similarly, the two race-specific degree completion models share 8 of the top 10 predictors. However, the ordering of the predictors differs slightly across race-specific models indicating that there are subtle differences in the best way to predict success between the subgroups.

eleventh most important predictor for the degree completion model. We compute the calibration and accuracy biases for the simple versions of the algorithms where targeting is based solely on one predictor.

RESULTS

Demonstrating Calibration Bias in Random Forest Models

We first illustrate the presence of calibration bias in Figure 1. The small "x" points correspond to bins of two percentiles of the predicted score distribution (e.g. the leftmost "x" corresponds to the average actual success rate for students in the first and second percentiles of predicted scores), whereas the larger dots correspond to deciles of predicted scores.[19] In both plots, we observe clear evidence of calibration bias (as seen by the gaps between the blue and green lines), although the pattern of bias differs between our two models. Panel A shows that in the course completion model, White students have significantly higher actual success rates compared to Black students at nearly all points in the distribution of predicted scores.

For example, suppose that a college used a risk threshold equal to the 30th percentile. In this case, the actual success rate among selected White students would be 50.3 percent, while the actual success rate among selected Black students would be 45.8 percent; this difference is statistically significant with a p-value of less than 0.01. The imbalance between the White and Black success rate around the risk threshold means that the course completion model is overestimating the success of Black students and underestimating the success of White students. If resources were allocated to students based on this model's predictions, then Black students

[19] The decile points also include vertical 95 percent confidence interval bars, although in many instances the confidence intervals are so tight that the bars are difficult to see due to the large size of our sample.

would receive relatively fewer additional supports (and in turn White students would receive more) than if the allocation was based on students' actual risk levels.

In Panel B of Figure 1, we observe a different pattern for the degree completion model. The predicted scores appear better calibrated between the White and Black students for the lower and middle portions of the predicted score distribution. At a risk threshold at the 30th percentile, the observed success rates are quite similar: 3.4 percent for White students, and 3.3 percent for Black students; this difference is not statistically significant (p-value of 0.682). We do observe gaps between the White and Black success rates for higher points in the distribution of predicted success (i.e. above the 60th percentile), although these are less relevant to the use case we are considering.[20]

Next, we quantify the amount of calibration bias by simulating an unbiased distribution of students around a given risk threshold (see above for details). Figure 2 shows the percentage change in the number of Black students labeled as at-risk students after we simulate the removal of the calibration bias, for a variety of risk thresholds.[21] Figure 2 further illustrates the different patterns of calibration bias across our two models. For the course completion model (blue line), the amount of bias is five times greater when labeling students in the bottom half as at-risk compared to labeling the students in the bottom decile as at-risk. In the degree completion model (green line), the reverse is true such that the amount of bias is five times higher when using the bottom decile as the risk threshold versus the bottom half.[22]

[20] If instead the degree completion model was used to target the most promising students with some form of positive outcome (e.g. admission into an honors program, additional financial aid), then the calibration bias would also be relevant to consider, particularly given the large visible gaps in the higher points of the predicted score distribution. In this instance, the calibration bias could actually be beneficial to Black students, since they would be more likely to receive the positive treatment despite having lower success rates.

[21] Appendix Figure A3 shows the corresponding percentage change in the number of White students labeled as at-risk at the different risk thresholds shown in Figure 2.

[22] There are established procedures for achieving better overall calibration in prediction models. We test whether applying two separate procedures (sigmoid and isotonic) mitigate the calibration bias we observe in Figure 2. In these

To further contextualize the bias metric depicted in Figure 2, consider the 30th percentile risk threshold for the course completion model. As shown in the plot, the "unbiased simulation" would label 14.2 percent more Black students than the actual model. This translates to the share of all Black students who are targeted as at-risk being 25 percent in the actual model and 28.5 percent in the unbiased simulation. In other words, the calibration bias present in this model would cause Black students to receive fewer resources than they would receive if there was no bias present.[23]

Demonstrating Accuracy Bias in Random Forest Models

Table 1 describes model accuracy separately for the White and Black samples. In both the course completion and degree completion models, the Black c-statistic is slightly but statistically significantly lower than the White c-statistic (3 percent and 1.2 percent lower, respectively). These racial differences in c-statistics are statistically significant at the $p < 0.01$ level.

Other data science papers investigating algorithmic bias in education contexts consider alternative notions of fairness. Among the most popular are equalized opportunity and equalized odds, which require equality across groups of one or both the true and false positive rates, which are also considered accuracy metrics.[24] In our context, the most relevant corollary is the true negative rate (among students who do not succeed, what share is labeled as at-risk?). Consistent

approaches, the training sample is split into two parts: the first to fit the model, and the second to generate a calibration correction (Menon et al, 2012; Niculescu-Mizil and Caruana, 2005; Platt, 1999). However, as we show in Appendix Figure A4, these approaches do not meaningfully change the amount of calibration bias present in the models.

[23] We do not display distributional points above the 70th percentile for two reasons. First, we do not view these as relevant points for labeling at-risk students. Second, above a certain threshold (due to the intrinsic gap between Black and White students in terms of actual success rate), it is not possible to perform the calibration bias simulation because there is an insufficient number of Black students who are not flagged as at-risk by the model to make up for the gap in success rate, despite the fact that most of them achieved the outcome of interest.. For the course completion model, this occurs by the 75th percentile. For the degree completion model, this occurs by the 90th percentile.

[24] Similarly, while there is a growing strand of data science literature for developing and testing approaches to mitigate bias in models, these papers focus on improving equalized opportunity, equalized odds, which requires that the TPR and TNR are the same across subgroups, or demographic parity, which requires the predicted score distribution be the same across subgroups (see, for examples, Pleiss et al, 2017; Wadsworth, Vera, and Piech, 2018; Zhang, Lemoine, and Mitchell, 2018).

with prior studies, we find higher true negative rates for Black students than White students, as shown in Table 1 (Anderson, Boodhwani, & Baker, 2019, Jiang & Pardos, 2021, Jeong et al, 2021, Yu, Lee, & Kizilcec, 2021, Yu et al, 2020). Based on this metric alone, one may conclude that algorithms are underestimating Black student success, and therefore Black students would be more likely to be targeted for additional resources--which is the opposite of what we conclude from the calibration bias results we present above. While TNR measures *overall* accuracy and efficiency of labeling at-risk students, the calibration bias measures the *marginal* accuracy and efficiency of labeling at-risk students. Based on our interpretation of which is most important to consider in the context of labeling at-risk college students, and following the example of Obermeyer et al (2019), we choose to focus our bias investigation on the calibration bias. Our results highlight the fact that focusing on a single metric of algorithmic bias may misrepresent the equity implications of using a prediction model.[25]

Bias in Simple Algorithms

We now compare the results from our two random forest models to the "simple" algorithms: using only GPA to predict student success, and separately, using only enrollment intensity (specifically, number of credits currently attempting) to predict student success. Table 2 compares the c-statistics of the simple models across subgroups. As expected, these simple models have substantially lower accuracy than the more complex random forest model (as shown in Table 1). The accuracy is typically lower for the Black subgroup, with the exception of using cumulative GPA to predict degree completion.

[25] Indeed, prior research has shown that there are inherent tensions between different metrics of algorithmic bias, and that often mitigating one form of bias requires increasing another (Lee & Kizilcec, 2020; Mitchell et al, 2021; Quy et al, 2022).

Figure 3 presents the calibration plots for the simple models. We find the same directional but now exaggerated pattern from Figure 1: In Figure 3, we see that within percentile of prior cumulative GPA or enrollment intensity, the White student success rate is significantly higher than the Black student success rate. This is particularly true when using enrollment intensity, where conditional on credits attempted, the White student course or degree completion rate is typically at least 10 percentage points higher compared to Black students. Figure 4 (analogous to Figure 2) confirms that there is substantial calibration bias for these simple models at most points in the distribution of cumulative GPA and enrollment intensity. This analysis shows that the same concerns about calibration bias exist even if colleges were to use these very simple targeting mechanisms.

Making Models More Attuned to Race Impacts Algorithmic Bias

Next, we compare algorithmic bias from the base models (Figure 2, Table 1) to models including racial predictors and race-specific models. Figure 5 plots the amount of calibration bias across these three specifications. Again, we find meaningfully different patterns in bias between the course completion model (Panel A) and the degree completion model (Panel B). For course completion, we see that making the base model more attuned to race significantly reduces the amount of calibration bias; this is especially true for risk thresholds focused on the students with the lowest predicted likelihood of success. This finding indicates that, for the course completion model, there is some useful information contained in race to allow for better calibration.[26] This finding aligns with other recent research that finds reduction in bias when race predictors are included (e.g. Yu, Lee, & Kizilcec, 2020). The reduction in calibration bias we see in Figure 5 for

[26] For instance, if there are unobserved factors that are highly predictive of student success and also highly correlated with race, then excluding the race predictors could result in underestimation of White student's predicted scores, and overestimation of Black student's predicted scores.

the course completion model does not come at the expense of overall model accuracy: Table 3, Panel A shows that the c-statistics for both the course and degree completion models including race information are nearly identical to the c-statistics of the base model.

However, including race information does not reduce calibration bias in the degree completion model (Figure 5, Panel B). With the exception of the bottom decile risk threshold for the model including racial predictors, the bias metric is the same or higher for the models more attuned to race. The overall pattern of results we observe in Figure 5 is noteworthy for two reasons. First, it further emphasizes the highly contextual nature of algorithmic bias. While the course completion and degree completion models predict separate measures of success, the models are overall quite similar in that they draw from the same administrative data set; include the same population of students in the samples; include similar types of predictors; and were built by the same team of researchers. The fact that introducing race into the models produced nearly opposite effects on the level of calibration bias demonstrates just how idiosyncratic algorithmic bias can be across models.[27] Second, most prior research shows that including race in the model (either through including race predictors or estimating race-specific models) does not increase algorithmic bias (e.g. Yu, Lee, & Kizilcec, 2020); our results for the degree completion model go against this general finding.[28]

[27] One explanation as to why we observe these divergent patterns is that including race may proxy for other unobservable factors that are related to both race and the outcome. We find that this is more likely to be the case when predicting course versus degree completion: The feature importance of the Black indicator is significantly higher in the course completion model (ranked 37th out of 191) compared to the degree completion model (ranked 116th out of 260). So, including race significantly improves the course completion model's ability to provide better marginal estimation across racial groups (even though the overall model performance measured by the c-statistic remains the same). A separate explanation is that the sample size of the course completion model is ten times as large as that of the degree completion model. It's possible that the increase in calibration bias in the race-specific degree completion models is due (at least in part) to the meaningful reduction in sample size. With a smaller sample, machine learning algorithms can be less stable and more prone to the idiosyncrasies of the training sample.

[28] We also test how excluding socio-economic information from the model impacts the calibration bias in the models. In Figure A5, we compare the base models (which excludes race but includes other demographic and socio-economic characteristics) with versions that exclude all demographic and socio-economic characteristics (except for

Exploring Why Algorithmic Bias Exists

One potential reason algorithmic bias may arise is if the outcome used within the model does not translate to the same true outcome across different subgroups. Obermeyer et al. (2019) provide an example: The algorithm they analyze uses health care expenditures as a proxy for actual health, but expenditure behavior differs in meaningful ways across races.[29] In our case, however, we are able to observe the actual outcome of interest, not just a proxy. Still, if Black students systematically choose to enroll in different courses or degree programs compared to White students, then the outcome of "success" could mean something different across racial lines. For example, consider two degree programs offered at VCCS: a transfer-oriented associate degree in Liberal Arts, and a certificate in welding. These programs differ substantially in the types of courses required, the time commitment needed to complete the program, and the types of skills necessary to succeed. Because of these programmatic differences, we would expect there to be meaningful differences in the types of students who choose to pursue either program, which may include differences on the dimension of race.[30]

To explore whether the differential sorting of students contributes to the models' calibration bias, we estimate course-specific and degree-specific models for the five most popular courses and degree programs offered by VCCS. We also estimate separate models according to program level: AA&S (transfer oriented associate degree); AAS (applied associate degree); CERT

student age). We find that including the additional socio-economic information slightly reduces the amount of bias in the model at most points in the distribution of predicted scores.

[29] Conditional on health, Black patients have lower health expenditures (due to being less likely to seek out medical care when ill), and are more likely to have costs associated with emergency care instead of preventative care.

[30] We test the hypothesis as to whether Black student enrollment choices are meaningfully different by regressing the share of Black students enrolled in a particular course or degree program on the success rate among White students in that course or program. We find that Black students are relatively more likely to enroll in courses with higher success rates. This pattern could result from Black students being more likely to enroll in courses that attract more high-performing students, or that Black students are more likely to enroll in "easy" courses. Conversely, we find that Black students are relatively more likely to enroll in degree programs with lower success rates--either programs that attract lower-performing students or more "difficult" programs. See Appendix Table A8 for these results.

(certificate program); and CSC (shorter "career studies certificate"). Appendix Figure A6 compares the amount of calibration bias in the full models versus course-specific models (Panel A, course completion model only); degree-specific models (Panel B, degree completion model only); or program level-specific models (Panel C, degree completion model only).[31] Overall, while we see that the levels of calibration bias tend to be more variable across courses or academic programs (e.g. less bias in ENG112 compared to ENG111; less bias in CERT programs compared to AA&S programs), we do not observe any meaningful pattern of bias reduction using the program-specific models instead of the full model.[32] This finding suggests that the calibration bias is not due to differential sorting of students across courses or programs of study.[33]

A separate reason why algorithmic bias may arise is due to underrepresentation of the minority group in the training data (Jiang & Pardos, 2021; Sha et al, 2022). Prediction models are typically trained to maximize overall accuracy; this means that successful prediction of students in the majority subgroups is given more weight in determining how predictions are made. We test this hypothesis explicitly by downsampling the non-Black subgroup to be the same size as the Black subgroup. We compare the amount of calibration bias in the base models to the downsampled models in Appendix Figure A8. For both the course completion model (blue lines) and degree completion model (green lines), the downsampled versions have very similar levels of

[31] To create a relevant comparison of the results, the "Full base model" refers to the model built using the full training sample but applied to just the validation sample from the relevant course, program of study, or program level.

[32] We find a similar pattern of results when we estimate college-specific models for the five largest VCCS colleges, as shown in Appendix Figure A7: While the amount of bias can vary across colleges, there is no meaningful bias reduction when estimating the college-specific models compared to the full model applied to the college-specific validation set. Again, this pattern of result supports two of our main conclusions: (1) The calibration bias is not due to differential sorting of students; and (2) algorithmic bias is highly contextual, and can differ substantially across fairly similar contexts.

[33] This conclusion is further supported by an additional series of test, where we regress the success outcomes on race only, and compare the coefficient estimates on the Black student indicator to separate regressions of the success outcomes on race and program of study fixed effects (for degree completion outcome) or course fixed effects (for course completion outcome). We find that the coefficient estimates are very similar regardless of whether program of study or course fixed effects are included in the model. This pattern of results indicates that the differences in success rates between Black and White students are not driven by the differential selection into program of study or course.

calibration bias at nearly all risk thresholds. These results suggest that data underrepresentation is not a meaningful source of calibration bias in our models.[34]

Still, there is another indirect type of data underrepresentation that occurs in our sample. Because Black students are less likely to persist in college, Black students in our sample are more likely to have only been enrolled for one term[35]; this lack of enrollment history limits the amount of information we can include in the models to predict their success outcomes. We test whether this type of data underrepresentation is a source of the algorithmic bias explicitly by separating our sample into two subgroups: first-term student observations and returning student observations, and estimating subgroup-specific models. We present the amount of calibration bias for the subgroup-specific models in Figure 6.[36] For both the course and degree completion models, we find that the amount of calibration bias is substantially higher for first-term students compared to returning students (i.e. comparing the blue lines to the green lines). When we compare the levels of bias between the full model and the subgroup-specific models (dashed lines versus solid lines of the same color), we find that the levels of bias in the subgroup-specific models are very similar for the course completion model (Panel A). However, we do observe some reduction in bias for first-term students at more stringent at-risk thresholds for the degree completion model (Panel B). This pattern of results supports the hypothesis that Black students having shorter enrollment histories is a contributing factor to some (but not all) of the calibration bias of our models, and suggests that the additional predictors available for returning students, such as cumulative GPA and credits completed from prior terms, partially mitigates the calibration bias. This finding also

[34] We also performed other tests of the data underrepresentation hypothesis, including upsampling and upweighting Black observations. We find very similar results to the downsampling tests.
[35] Specifically, within the validation samples, Black students are 17 percent (course completion) and 25 percent (degree completion) more likely to have enrolled for only one term compared to White students.
[36] Again, the comparison "full" models in Figure 6 are the models trained on the full training set but applied to the subgroup-specific validation set.

suggests that in some settings (though not necessarily universally), estimating separate models for new and returning students could reduce the level of algorithmic bias.

One potential confound in interpreting the calibration bias is that, while we observe the ultimate outcomes of course and degree completion, we do not observe students' latent risk level for time points prior to when the outcomes are revealed. Specifically, suppose that a Black student and a White student begin a term with the same true (but unobservable to the researcher) risk level. If, due to implicit biases on the part of faculty and staff, the White student labeled as at-risk by faculty/staff receives more positive attention (e.g. offers for additional tutoring resources), or the Black student labeled as at-risk receives more negative attention (e.g. discouragement to continue on the course of study), then by the end of the term the White student would be more likely to succeed than the Black student, even though they began with the same true risk level. This scenario could explain the differences in actual completion rates conditional on predicted score that we see in Figure 1. We may be particularly concerned about misinterpreting the patterns in Figure 1 as algorithmic bias in cases where colleges are actively using predictive analytics to target at-risk students. However, to the best of our knowledge and based on conversations with several system and college administrators, VCCS was not using predictive analytics or any other systematic risk assessment tools during the time spanning our data.[37] What's more, if there were systematic differences in the treatment of at-risk students by race, and assuming these differences were true during the time periods of the training and validation sample, then including racial predictors or

[37] A separate potential issue related to algorithmic bias is the "selective labels problem", whereby the outcome of interest is not observed for all observations in the training sample. This issue is particularly prominent in criminal justice applications that inform judges' bail decisions (i.e. whether to detain the defendant or grant them bail). In these instances, data scientists can only observe the relevant outcome (i.e. whether the defendant skipped bail) for those defendants who were granted bail. Because historically Black defendants are less likely to be granted bail, this selective labels problem can introduce algorithmic bias (Kleinberg et al, 2018). In our context, however, we fully observe the outcomes of interest for all observations in our sample. However, if colleges wished to use these algorithms to assess the likelihood of course or degree completion for potential students (instead of enrolled students), then the selective labels problem would be a relevant issue.

estimating race-specific models would account for these differences. Because we still see some amount of calibration bias for these model specifications (see Figure 5), we are confident in our interpretation of our results as being driven (at least in large part) by algorithmic bias.

Persistent differential accuracy by race

Throughout our investigation, we continue to find similar racial differences in overall accuracy as measured by the c-statistic, regardless of the amount of calibration bias present in a particular model. The c-statistics for the various models represented in Figures 6, A6, A7, and A8 are in Appendix Tables A9 through A12. This pattern of results suggests that predicting success for Black students may be inherently more difficult compared to White students. We explore this hypothesis by comparing two additional goodness of fit metrics from the training sample: Efron's R-squared (Efron, 1978) and McFadden's Adjusted R-squared (McFadden, 1973). We compute these statistics for the race-specific models, so that the determinants of success are allowed to differ across groups. Table 4 shows that with the exception of the Efron's R-squared for the course completion model, there are meaningfully higher goodness of fit metrics for White students compared to Black students. These results indicate that the information contained in our set of predictors is not as related to Black student success. As we tried to be as inclusive as possible in constructing our predictors--that is, we tried to incorporate as much information about students as we could given what we observe in the administrative data--any additional information to improve accuracy of Black student success would need to come from outside the existing VCCS administrative data. These new data could come from linked student high school records or student intake surveys.

DISCUSSION

Given rapid expansion of predictive analytics in higher education and persistent racial gaps in student success, algorithmic bias is an important and policy-relevant topic. However, current literature investigating algorithmic bias focuses primarily on technical aspects of model development; comparatively little work has provided translational insights about the presence and implications of algorithmic bias for policy- and practice-oriented audiences. In the two models we consider (course completion and degree completion), we find meaningful algorithmic bias on two dimensions: (1) Conditional on predicted score, Black students have worse observed outcomes than White students, which would lead to some Black students being less likely to receive additional resources than White students who are comparatively more likely to succeed; and (2) the models have slightly to moderately worse accuracy for Black students, which could lead to higher misclassification. The first dimension of bias (calibration) translates to the models relatively overestimating Black student performance at the threshold of being labeled as at-risk--this is despite alternative metrics (e.g. true negative rate) showing that model underestimates Black student performance *overall*. This finding highlights the importance of choosing how to measure algorithmic bias, as overall measures can mask the implications of bias at the margin.

However, comparing the two models, we find significant differences in both the amount of bias (both overall and at different points in the distribution of predicted success) and its practical implications (e.g. whether including race information mitigates or exacerbates bias). These findings are somewhat surprising given that the two models draw on the same data and were built by the same team of researchers, and emphasizes the highly contextual nature of algorithmic bias. These differences in algorithmic bias across highly-similar models reinforces the importance of researchers and policy-makers investigating and mitigating bias in the specific context in which predictive algorithms are being used.

Our findings suggest that algorithmic bias in our models is driven (at least in part) by available administrative data being less useful at predicting Black student success compared with White student success. This is especially true for first-time students, where the amount of information that can be used for prediction is extremely limited in the community college context. The comparatively lower value of existing administrative data in predicting Black students' outcomes may reflect historical inequities in the extent to which colleges and universities have focused their data collection efforts on measures relevant to the success of students from diverse backgrounds. Incorporating additional data sources--such as high school transcripts, student surveys, or engagement on learning management system platforms--may reduce the algorithmic bias in models predicting college student success for a more diverse array of students.

Given that many of the private vendors offering predictive analytics tools in higher education treat their models as proprietary, it is important to address what colleges can do to address algorithmic bias when they do not have direct access to the models.[38] As Ekowo and Palmer (2016) also emphasize, choosing a vendor that is willing to be transparent about their product and being knowledgeable of the underlying models is the first step to success. Colleges can insist that vendors provide documentation of the presence of and mitigation efforts to address algorithmic bias within the same (or closely-similar) contexts to where the institution plans to use predictive analytics. Colleges could also request raw predicted scores to perform their own algorithmic bias investigation--most of the results we present in the paper do not require having access to the underlying model, only the students' observed outcomes and their predicted scores.

Particularly as broad-access colleges and universities continue to grapple with declining enrollments and in turn revenues, they are likely to be in the position of even scarcer resources,

[38] Numerous reports, including a recent analysis by the Government Accountability Office, raise concerns about this lack of transparency and its implications for the accuracy or fairness of commercial prediction models (GAO, 2022).

while still serving many students who may need support to earn their credential or degree. Predictive analytics have the potential to enhance institutions' ability to target these resources to students most in need of assistance, yet as our analyses show, algorithmic bias may result in at-risk Black students receiving less support than similar White students. Identifying and mitigating algorithmic bias will therefore be an important component of colleges' and universities' broader efforts to work towards greater racial equity.

REFERENCES

Anderson, H., Boodhwani, A., & Baker, R. (2019). Assessing the Fairness of Graduation Predictions. *Educational Data Mining*.

Angwin, Julia, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine Bias. ProPublica, May 23, 2016: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Association of Public & Land-Grant Universities. (2016). Congress should lift the ban on student level data in HEA reauthorization (Policy Brief). https://www.aplu.org/library/congress-should-lift-the-ban-on-student-level-data-in-hea-reauthorization/File

Arnold, David, Will S. Dobbie, and Peter Hull. "Measuring Racial Discrimination in Algorithms." NBER Working Paper 28222, January 2021. https://www.nber.org/system/files/working_papers/w28222/w28222.pdf

Baker, R.S., Hawn, A. Algorithmic Bias in Education. *Int J Artif Intell Educ* 32, 1052–1092 (2022). https://doi.org/10.1007/s40593-021-00285-9

Barshay J., Aslanian S. (2019). *Under a watchful eye: Colleges are using big data to track students in an effort to boost graduation rates, but it comes at a cost* (APM Reports). https://www.apmreports.org/story/2019/08/06/college-data-tracking-students-graduation

Bird, K. A., Castleman, B. L., Mabel, Z., & Song, Y. (2021). Bringing Transparency to Predictive Analytics: A Systematic Comparison of Predictive Modeling Methods in Higher Education. *AERA Open*, *7*. https://doi.org/10.1177/23328584211037630

Breiman, L. (2002). Manual on setting up, using, and understanding random forests v3. 1. Statistics Department University of California Berkeley, CA, USA 1 58.

Burke M., Parnell A., Wesaw A., Kruger K. (2017). Predictive analysis of student data: A focus on engagement and behavior. The National Association of Student Personnel Administrators. https://www.naspa.org/images/uploads/main/PREDICTIVE_FULL_4-7-17_DOWNLOAD.pdf

Chouldechova, Alexandra and Aaron Roth (2018). "The Frontiers of Fairness in Machine Learning." https://doi.org/10.48550/arXiv.1810.08810

Corbett-Davies, Sam, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. 2017. Algorithmic decision making and the cost of fairness. In Proceedings of KDD '17, Halifax, NS, Canada, August 13-17, 2017, 10 pages. DOI: 10.1145/3097983.3098095

Cowgill, Bo and Catherine Tucker. "Economics, Fairness, and Algorithmic Bias." Working paper, May 11, 2019: http://conference.nber.org/confer/2019/YSAIf19/SSRN-id3361280.pdf

Efron, Bradley (1978) Regression and ANOVA with Zero-One Data: Measures of Residual Variation, Journal of the American Statistical Association, 73:361, 113-121, DOI: 10.1080/01621459.1978.10480013

Ekowo, Manuela and Iris Palmer. "The Promise and Peril of Predictive Analytics in Higher Education: A Landscape Analysis." New America: Policy Paper, Oct. 24, 2016. https://www.newamerica.org/education-policy/policy-papers/promise-and-peril-predictive-analytics-higher-education/#:~:text=In%20a%20new%20paper%2C%20The,well%20in%2C%20and%20provide%20digital

GAO (United States Government Accountability Office). Consumer Protection: Congress Should Consider Enhancing Protections around Score Used to Rank Consumers. Report to the Chairwoman, Subcommittee on Research and Technology, Committee on Science, Space, and Technology, House of Representatives. May 2022. https://www.gao.gov/assets/gao-22-104527.pdf

Ghojogh, B. and M. Crowley. The Theory Behind Overfitting, Cross Validation, Regularization, Bagging, and Boosting: Tutorial. https://arxiv.org/abs/1905.12787:arXiv.

Greenstein, Nathan, Grand Crider-Phillips, Claire Matese, and Sung-Woo Choo. Predicting Risk Earlier: MAching Learning to Support Success and Combat Inequity. Academic Data Analytics Briefing Document, February 2023. University of Oregon, Office of the Provost: https://provost.uoregon.edu/analytics/student-success

Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*. 1982 Apr;143(1):29-36. doi: 10.1148/radiology.143.1.7063747. PMID: 7063747.

Hosmer, David W. Jr., Stanley Lemeshow, and Rodney X. Sturdivant (2013). Applied Logistic Regression, Third Edition. John Wiley & Sons, Inc., Hoboken New Jersey. ISBN 978-0-470-58247-3.

Jeong, Haewon, Michael D. Wu, Nilanjana Dasgupta, Muriel Medard, Flavio P. Calmon. Who Gets the Benefit of the Doubt? Racial Bias in Machine Learning Algorithms Applied to Secondary School Math Education. 35th Conference on Neural INformation Processing Systems (NeurlIPS 2021).

Jiang, Weijie and Zachary A. Pardos. 2021. Towards Equity and Algorithmic Fairness in Student Grade Prediction. In Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (AIES '21), May 19–21, 2021, Virtual Event, USA. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/ 3461702.3462623

Kleinberg, Jon, Jens Ludwig, Sendhil Mullainathan, and Ashesh Rambachan. Algorithmic Fairness. *AEA Papers and Proceedings 2018*, 108: 22-27 https://pubs.aeaweb.org/doi/pdfplus/10.1257/pandp.20181018

Kleinberg, Jon, Sendhil Mullainathan, and Manish Raghavan (2016). "Inherent Trade-Offs in the Fair Determination of Risk Scores." https://doi.org/10.48550/arXiv.1609.05807

Kung, Catherine and Renzhe Yu. "Interpretable Models Do Not Compromise Accuracy of Fairness in Predicting College Success." *L@S'20*, August 12-14, 2020, Virtual Event, USA.

Lakkaraju, Himabindu, Jon Kleinberg, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. "The Selective Labels Problem: Evaluating Algorithmic Predictions in the Presence of Unobservables." KDD'17, August 13-17, 2017, Halifax, NS, Canada. https://dl.acm.org/doi/10.1145/3097983.3098066

Lee, Hansol and Rene F. Kizilcec (2020). "Evaluation of Fairness Trade-offs in Predicting Student Success." https://doi.org/10.48550/arXiv.2007.00088

McFadden, D. (1973) Conditional Logit Analysis of Qualitative Choice Behavior. In: Zarembka, P., Ed., Frontiers in Econometrics, Academic Press, 105-142.

Menon AK, Jiang XJ, Vembu S, Elkan C, Ohno-Machado L. Predicting accurate probabilities with a ranking loss. Proc Int Conf Mach Learn. 2012;2012:703-710. PMID: 25285328; PMCID: PMC4180410. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4180410/

Niculescu-Mizil, Alexandru and Rich Caruana. Predicting Good Probabilities with Supervised Learning. *Proceedings of the 22nd International Conference on Machine Learning*, Bonn, Germany, 2005. https://www.cs.cornell.edu/~alexn/papers/calibration.icml05.crc.rev3.pdf

Obermeyer, Ziad, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 25 Oct 2019, Vol 366, Issue 6464, pp. 447-453. DOI: 10.1126/science.aax234

Ogundana, Ewaoluwa and Monique O. Ositelu. "Early Alert Systems: Why the Personal Touch is Key." New America: Blog Post, March 15, 2022. https://www.newamerica.org/education-policy/edcentral/early-alert-systems-why-the-personal-touch-is-key/

Paterson, James. "What a predictive analytics experiment taught 11 colleges about sharing data." Higher Ed Dive, April 18, 2019: https://www.highereddive.com/news/what-a-predictive-analytics-experiment-taught-11-colleges-about-sharing-dat/552986/

Paulus, J.K., Kent, D.M. Predictably unequal: understanding and addressing concerns that algorithmic clinical prediction may increase health disparities. *npj Digit. Med.* 3, 99 (2020). https://doi.org/10.1038/s41746-020-0304-9

Platt, John C. Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods. In *Advanced in Large Margin Classifiers*, Alexander J. Smola, Peter Bartlett, Bernard Scholkopf, Dave Schuurmans, eds, MIT Press, (1999). https://home.cs.colorado.edu/~mozer/Teaching/syllabi/6622/papers/Platt1999.pdf

Pleiss, Geoff, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q. Weinberger. On Fairness and Calibration. 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA. https://papers.nips.cc/paper_files/paper/2017/file/b8b9c74ac526fffbeb2d39ab038d1cd7-Paper.pdf

Pope, Devin G., and Justin R. Sydnor. 2011. "Implementing Anti-discrimination Policies in Statistical Profiling Models." *American Economic Journal: Economic Policy*, 3 (3): 206-31. DOI: 10.1257/pol.3.3.206

Rambachan, Ashesh, Jon Kleinberg, Jens Ludwig, and Sendhil Mullainathan. An Economic Perspective on Algorithmic Fairness. *AEA Papers and Proceedings 2020*, 110: 91-95 https://pubs.aeaweb.org/doi/pdfplus/10.1257/pandp.20201036

Riazy, S.; Simbeck, K. and Schreck, V. (2020). Fairness in Learning Analytics: Student At-risk Prediction in Virtual Learning Environments. In *Proceedings of the 12th International Conference on Computer Supported Education - Volume 1: CSEDU*, ISBN 978-989-758-417-6; ISSN 2184-5026, pages 15-25. DOI: 10.5220/0009324100150025

Sha, L., M. Raković, A. Das, D. Gašević and G. Chen, "Leveraging Class Balancing Techniques to Alleviate Algorithmic Bias for Predictive Tasks in Education," in *IEEE Transactions on Learning Technologies*, vol. 15, no. 4, pp. 481-492, 1 Aug. 2022, doi: 10.1109/TLT.2022.3196278.

Smith, Vernon C., Adam Lange, and Daniel R. Huston. Predictive Modeling to Forecast Student Outcomes and Drive Effective Interventions in Online Community College Courses. *Journal of Asynchronous Learning Networks*, Volume 16: Issue 3 (2012). https://files.eric.ed.gov/fulltext/EJ982673.pdf

Stark T. (2015, September 14). Leveraging analytics in community colleges. EDUCAUSE Review. https://er.educause.edu/articles/2015/9/leveraging-analytics-in-community-colleges

Swaak, Taylor. "How higher ed is trying to improve student performance with data." PBS News Hour, Aug 26, 2022: https://www.pbs.org/newshour/education/how-higher-ed-is-trying-to-improve-student-performance-with-data

Treaster, Joseph B. "Will You Graduate? Ask Big Data." The New York Times, Feb. 2, 2017, https://www.nytimes.com/2017/02/02/education/edlife/will-you-graduate-ask-big-data.html?_r=1

U.S. Department of Education (2022). Department of Education Releases Equity Action Plan as Part of Biden-Harris Administration's Efforts to Advance Racial Equity and Support Underserved Communities. Press release, April 14, 2022. https://www.ed.gov/news/press-releases/department-education-releases-equity-action-plan-part-biden-harris-administrations-efforts-advance-racial-equity-and-support-underserved-communities

Wadsworth, Christina, Francesca Vera, and Chris Piech. Achieving Fairness through Adversarial Learning: an Application to Recidivism Prediction. FAT/ML Workshop, July 2018, Stockholm, Sweden. https://www.fatml.org/media/documents/achieving_fairness_through_adversearial_learning.pdf

Yu, Renzhe, Hansol Lee, and Rene F. Kizilcec. Should College Dropout Prediction Models Include Protected Attributes? In Proceedings of the ACM Conference on Learning at Scale (L@S) 2021. https://doi.org/10.48550/arXiv.2103.15237

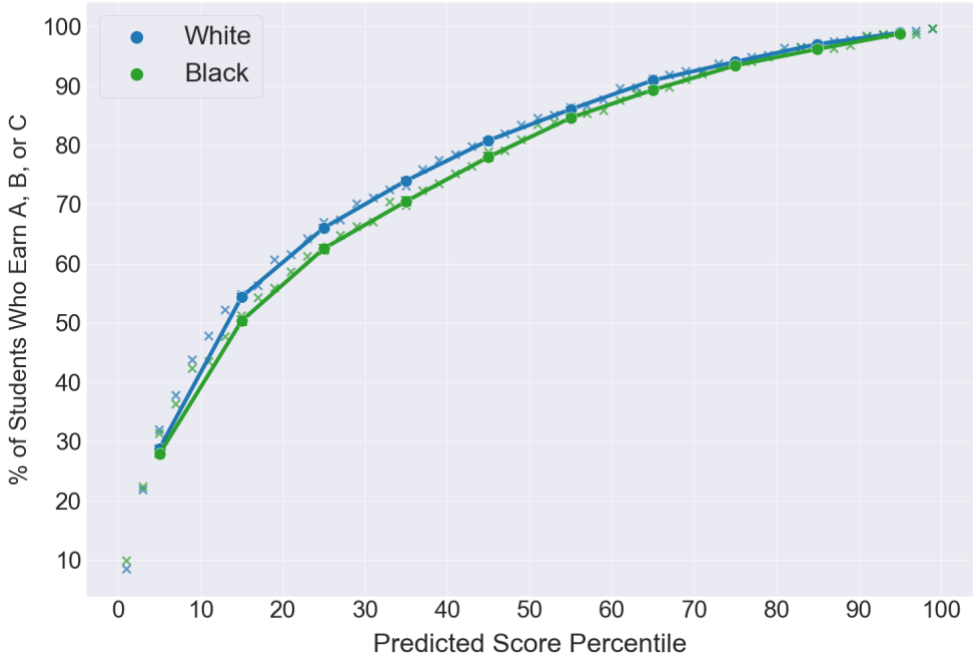
Yu, Renzhe, Qiujie Li, Christian Fischer, Shayan Doroudi and Di Xu "Towards Accurate and Fair Prediction of College Success: Evaluating Different Sources of Student Data" In: Proceedings of The 13th International Conference on Educational Data Mining (EDM 2020),

Anna N. Rafferty, Jacob Whitehill, Violetta Cavalli-Sforza, and Cristobal Romero (eds.) 2020, pp. 292 - 301

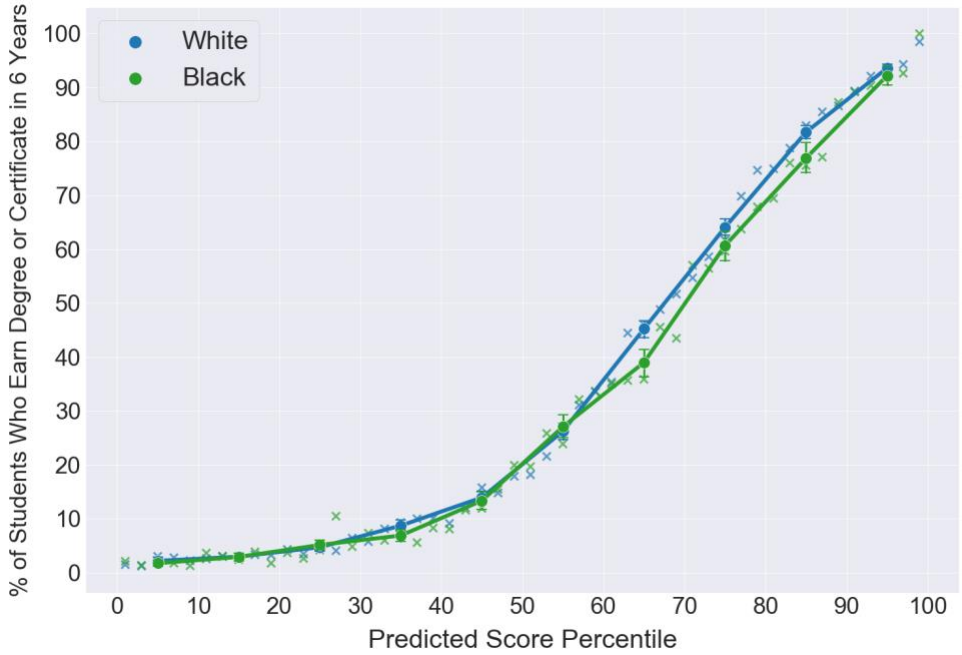
Zhang, Brian Hu, Blake Lemoine, and Margaret Mithcell. 2018. Mitigating Unwanted Biases with Adversarial Learning. In *2018 AAAI/ACM Conference on AI, Ethics, and Society (AIES '18), February 2-3, 2018, New Orleans LA, USA.*
https://dl.acm.org/doi/pdf/10.1145/3278721.3278779

Figure 1. Calibration of Base Random Forest Models, by Race.

Panel A: Course completion model

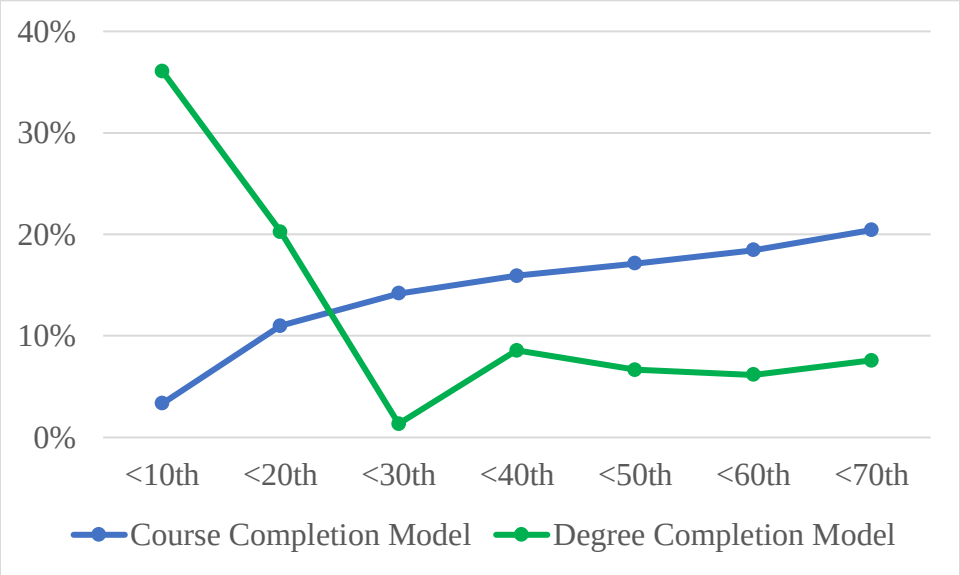


Panel B: Degree Completion Model



Notes: Each plot shows the average success rate of observations in the validation sample, conditional on predicted score percentile, separately for White and Black students. The small "x" points correspond to bins of two percentiles the predicted score distribution (e.g. the leftmost "x" corresponds shows the average actual success rate for students in the first and second percentiles of predicted scores), whereas the larger filled circles correspond to deciles of predicted scores. The decile points also include vertical 95 percent confidence interval bars.

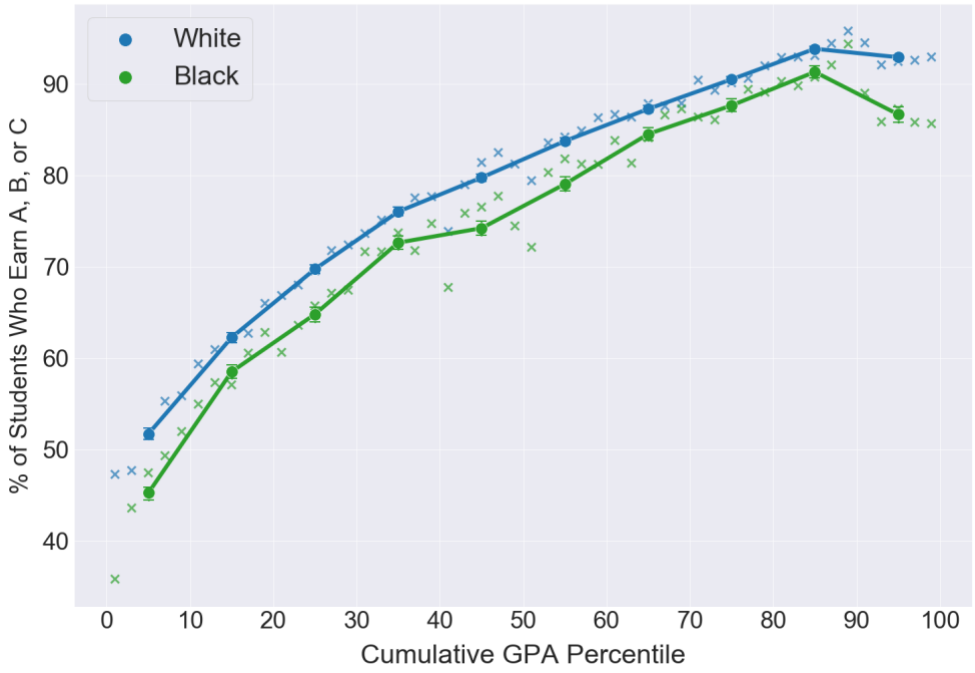
Figure 2. The Percentage Change in Black Students Labeled At-Risk After Simulating Calibration Bias Removal from the Random Forest Models.



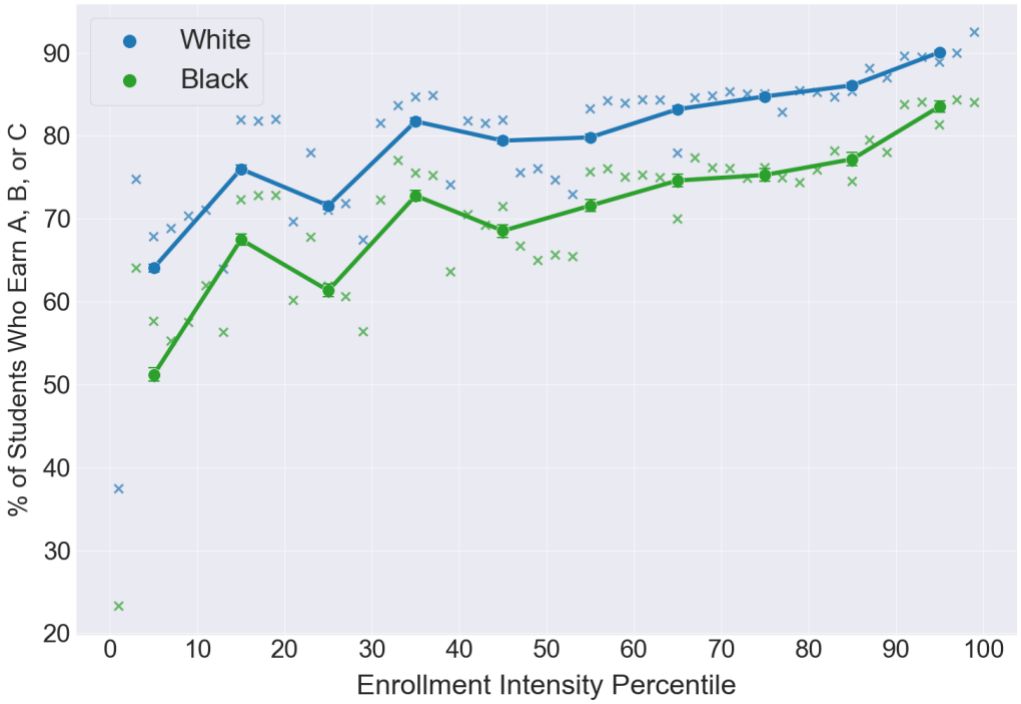
Notes: This plot shows, for a given risk threshold (e.g. < 30th percentile of predicted risk scores), the percentage increase in the number of Black students who would be labeled as at-risk after we simulate the calibration bias removal. Specifically, we calculate the amount of calibration by following these steps: (1) Select all White and Black students whose predicted scores are below the risk threshold; (2) if the actual success rate of the labeled at-risk Black students is lower than that of the labeled at-risk White students, change the label of the Black student whose predicted score is lowest among the Black students not labeled at-risk , and simultaneously drop the White student whose predicted score is highest among the White students labeled at-risk; (3) repeat step two until the actual success rates of the two selected groups become equal. See Appendix Figure A3 for the corresponding percentage decrease in the number of White students who would be labeled as at-risk.

Figure 3. Calibration of "Simple" Models Using Single Predictor of Student Success, by Race.

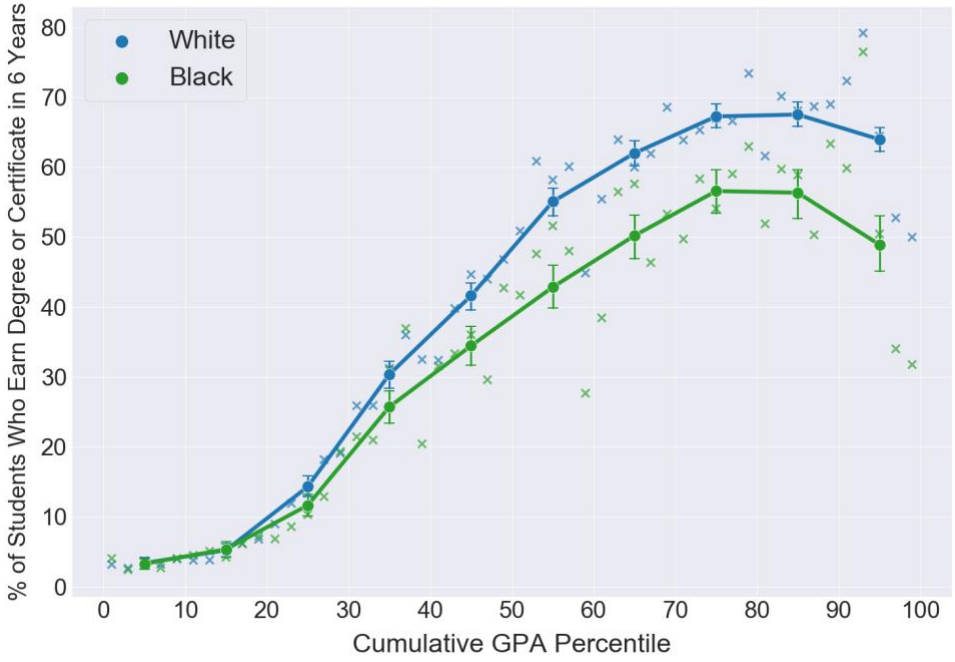
Panel A: Course completion model, GPA



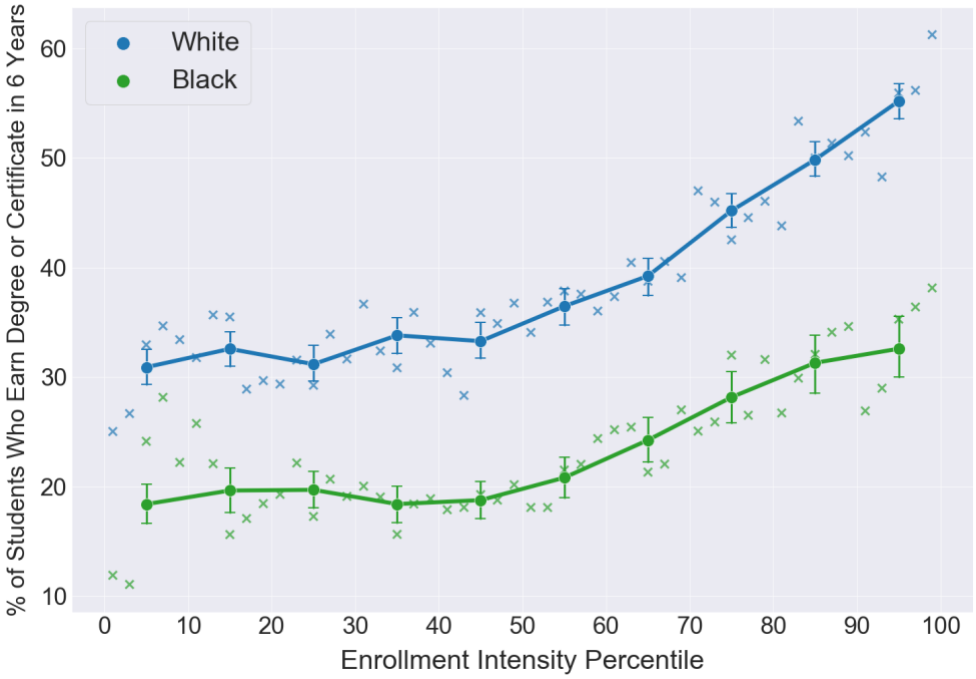
Panel B: Course completion model, enrollment intensity



Panel C: Degree Completion Model, GPA

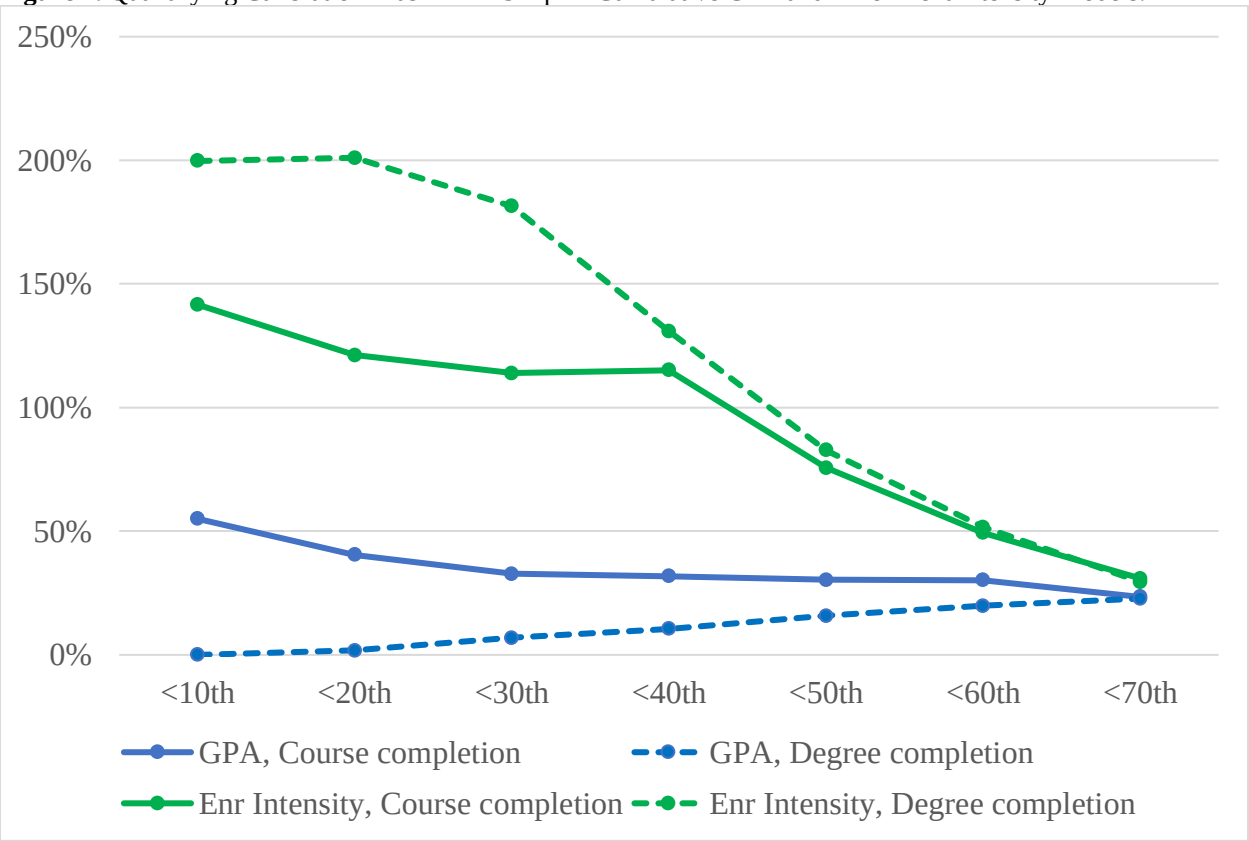


Panel D: Degree Completion Model, enrollment intensity



Notes: Each plot shows the average success rate of observations in the validation sample, conditional on the percentile of the single predictor used (cumulative GPA for Panels A and C; current enrollment intensity, measured by number of credits attempted, for Panels B and D), separately for White and Black students. The small "x" points correspond to bins of two percentiles the predicted score distribution (e.g. the leftmost "x" corresponds shows the average actual success rate for students in the first and second percentiles of predicted scores), whereas the larger filled circles correspond to deciles of predicted scores. The decile points also include vertical 95 percent confidence interval bars.

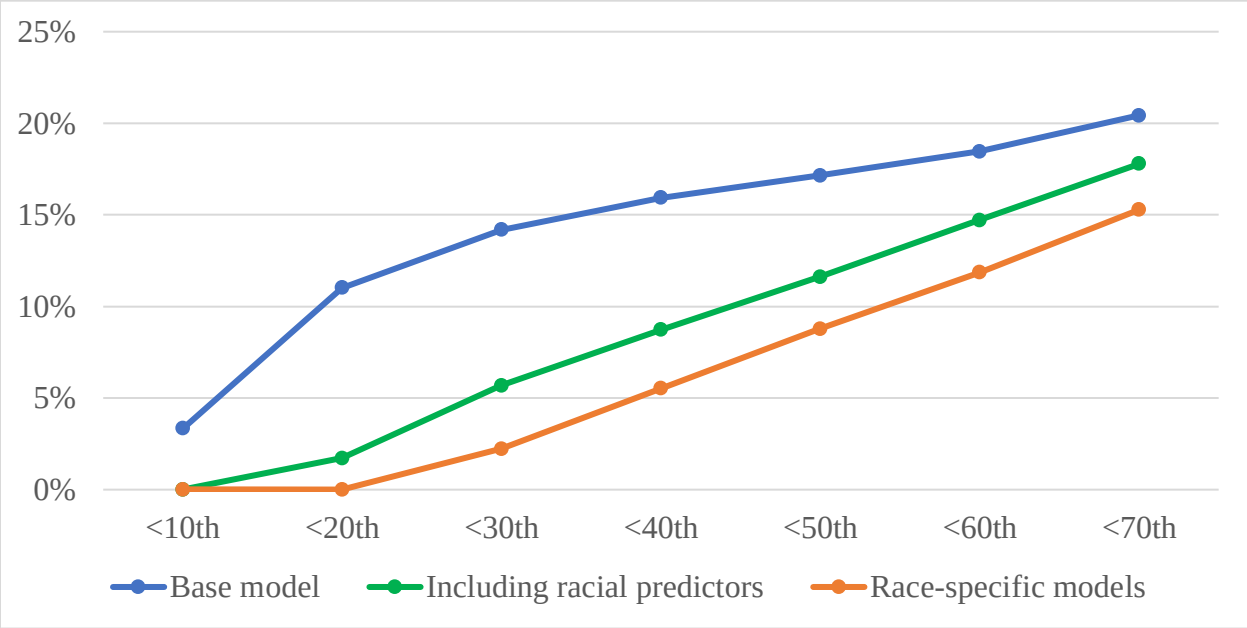
Figure 4. Quantifying Calibration Bias From "Simple" Cumulative GPA and Enrollment Intensity Models.



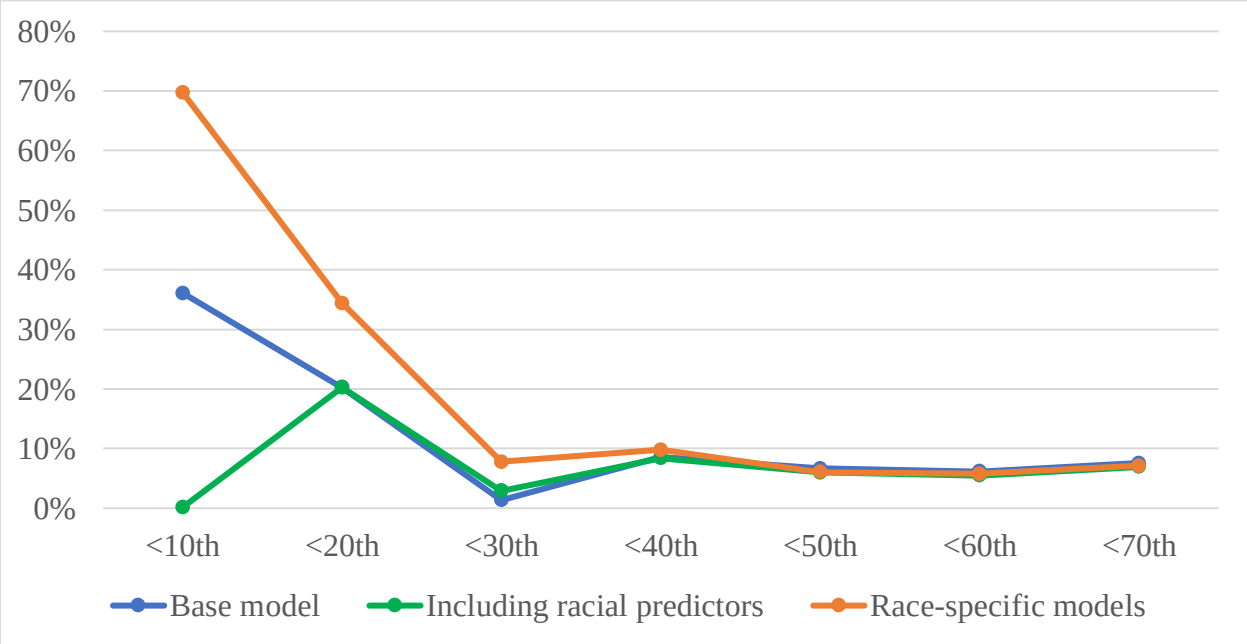
Notes: we generate these figures using the same methods as described in Figure 2, except in this case we use the percentile of cumulative GPA or percentile of enrollment intensity (i.e. number of credits attempted in the current term) from the training sample instead of the percentile of predicted score.

Figure 7: Comparing Calibration Bias in Base Random Forest Models Versus Making Models More Attuned to Race.

Panel A: Course Completion Model



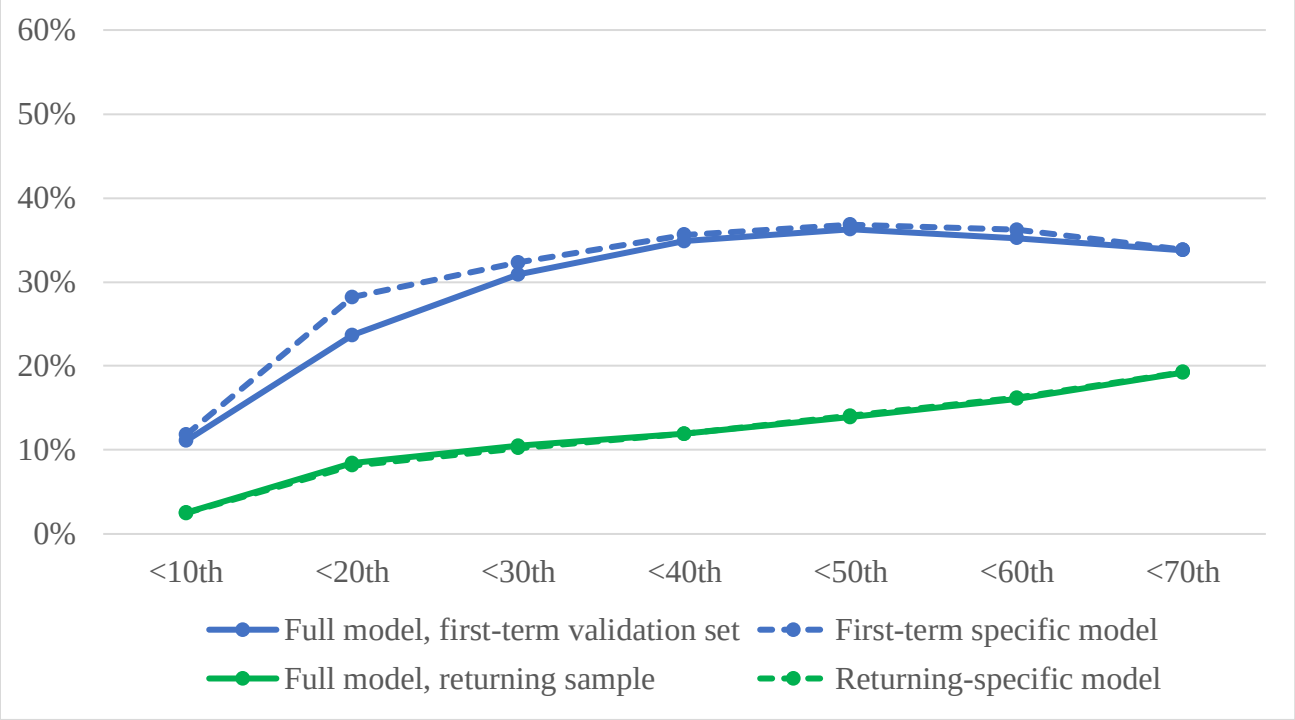
Panel B: Degree Completion Model



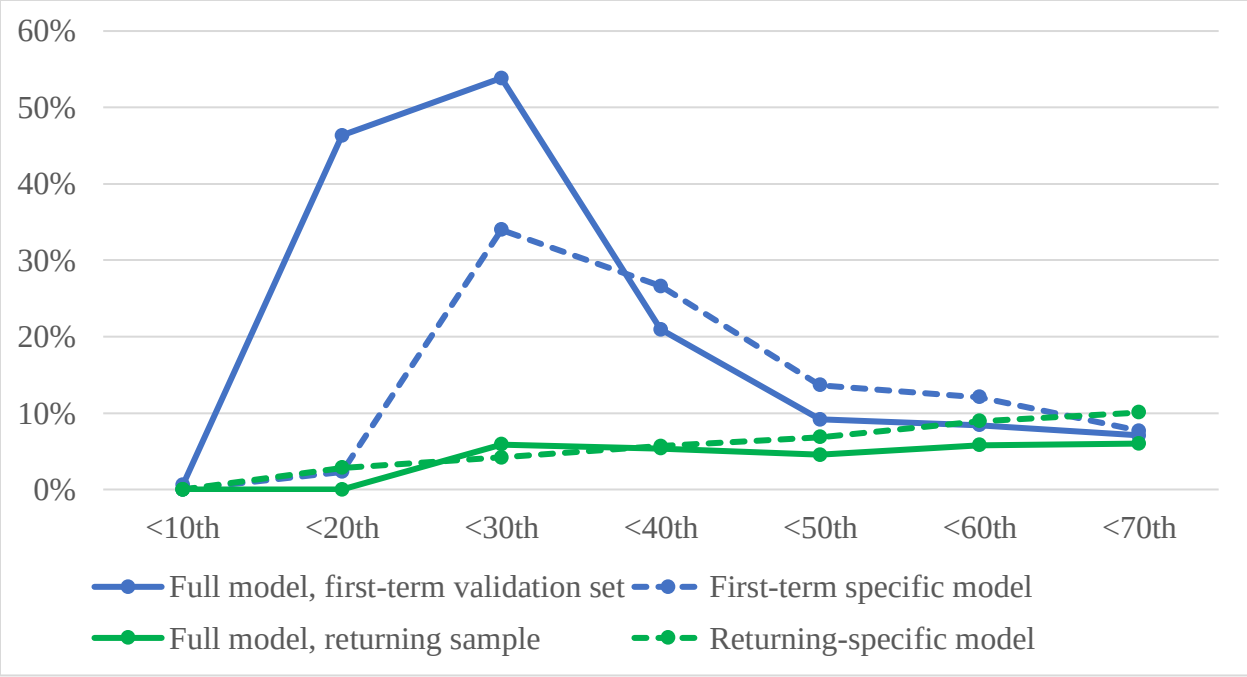
Notes: This figure plots the amount of calibration bias (measured by the increase in Black students targeted as at-risk in the simulated model, as in Figure 2) for the base RF model, the RF model including racial predictors, and race-specific RF models.

Figure 6: Comparing Calibration Bias for First Term and Returning Students.

Panel A: Course completion model



Panel B: Degree completion model



Notes: This figure plots the amount of calibration bias (measured by the increase in Black students targeted as at-risk in the simulated model, see Figure 2) for the base RF model to the RF models built using a training sample restricted to either first-term or returning student observations. For both the base models and the student-type-specific models, the amount of calibration bias is calculated using the validation sample restricted to observations of students who are either first-term or returning, accordingly.

Table 1. Accuracy of base random forest models.

Panel A: Course completion model			
	White	Black	% diff
C-statistic	0.8286	0.8037	-3.01%
	(0.0007)	(0.0012)	
True Negative Rate	0.4138	0.5006	20.98%
	(0.0017)	(0.0024)	
Panel B: Degree completion model			
	White	Black	% diff
C-statistic	0.8981	0.8878	-1.15%
	(0.0020)	(0.0036)	
True Negative Rate	0.849	0.9171	8.02%
	(0.0026)	(0.0024)	

Notes: Standard errors in parentheses. All differences between White and Black metrics are statistically significant at $p < 0.01$.

Table 2. C-statistics of "simple" models using single predictor of student success.

Panel A: course completion outcome

	White	Black	% diff
Cumulative GPA	0.7218	0.6993	-3.12%
	(0.0010)	(0.0016)	
Enrollment Intensity	0.6199	0.6022	-2.86%
	(0.0010)	(0.0016)	

Panel B: degree completion outcome

	White	Black	% diff
Cumulative GPA	0.7551	0.7878	4.33%
	(0.0031)	(0.0051)	
Enrollment Intensity	0.5946	0.5693	-4.25%
	(0.0033)	(0.0054)	

Notes: Standard errors in parentheses. All differences between White and Black metrics are statistically significant at $p < 0.01$.

Table 3. C-statistics of models including race information.

Panel A: course completion model

Base model (no race information)			*Full model including race predictors*			*Race-specific models*		
White	Black	%diff	White	Black	%diff	White	Black	%diff
0.8286	0.8037	-3.01%	0.8297	0.8029	-3.23%	0.8277	0.797	-3.71%
(0.0007)	(0.0012)		(0.0007)	(0.0012)		(0.0007)	(0.0012)	

Panel B: degree completion model

Base model (no race information)			*Full model including race predictors*			*Race-specific models*		
White	Black	%diff	White	Black	%diff	White	Black	%diff
0.8981	0.8878	-1.15%	0.8982	0.8879	-1.15%	0.8984	0.8834	-1.67%
(0.0020)	(0.0036)		(0.0020)	(0.0036)		(0.0019)	(0.0037)	

Notes: standard errors in parentheses. All differences between White and Black are significant at the $p < 0.01$ level

Table 4. Goodness of fit metrics for race-specific random forest models.

Panel A: Course completion model

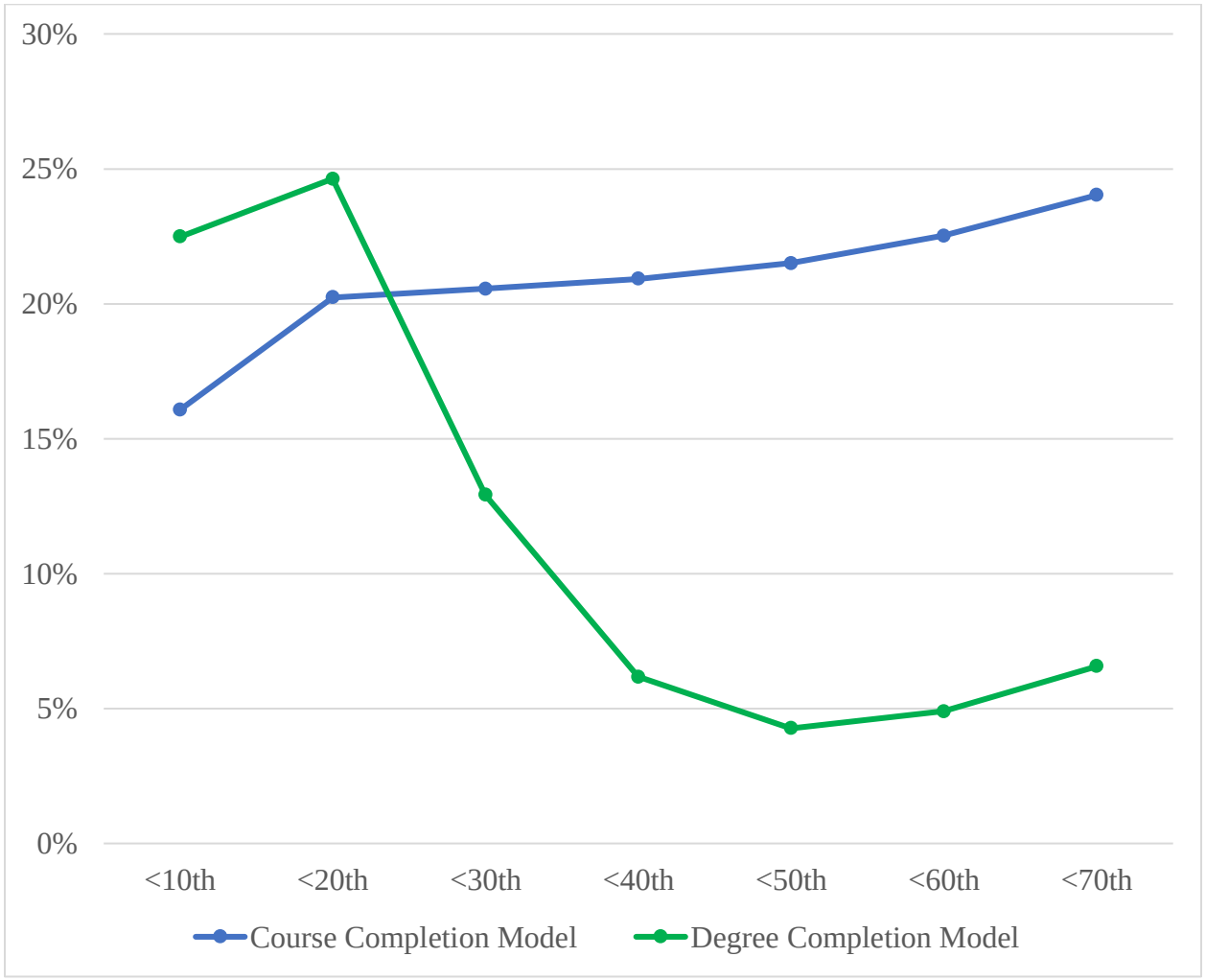
	White	Black
Efron's R-squared	0.7799	0.7847
McFadden's Adjusted R-squared	0.6736	0.6675

Panel B: Degree completion model

	White	Black
Efron's R-squared	0.5248	0.473
McFadden's Adjusted R-squared	0.4491	0.407

Notes: We calculate each goodness of fit metric using the model fitting from the race-specific models, and applied to the race-specific training sets.

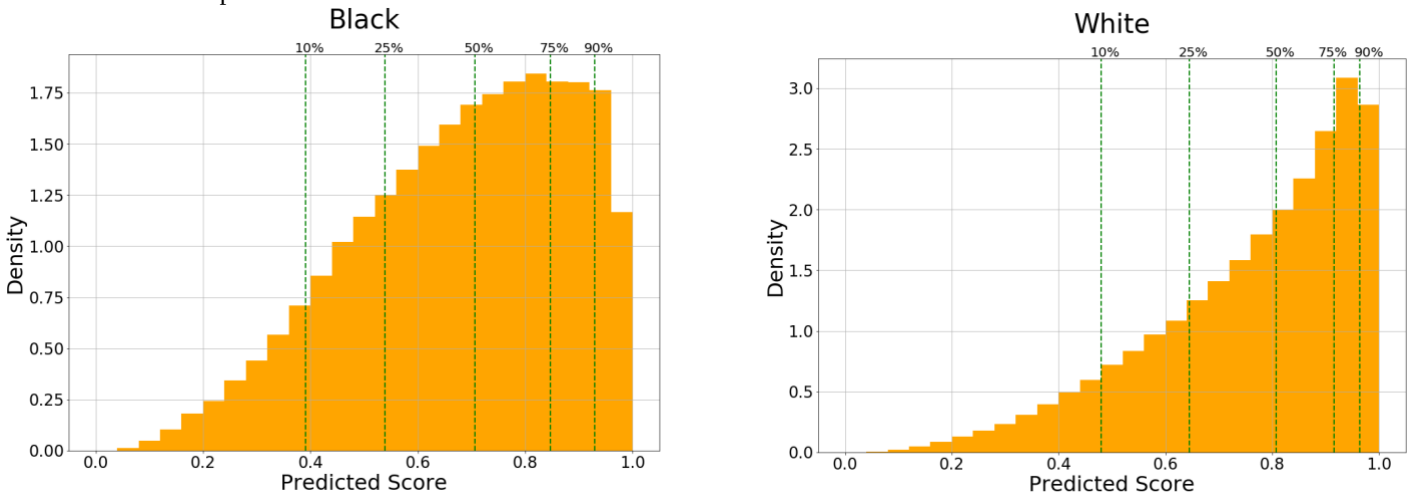
Appendix Figure A1. Calibration Bias from the Logistic Regression Model.



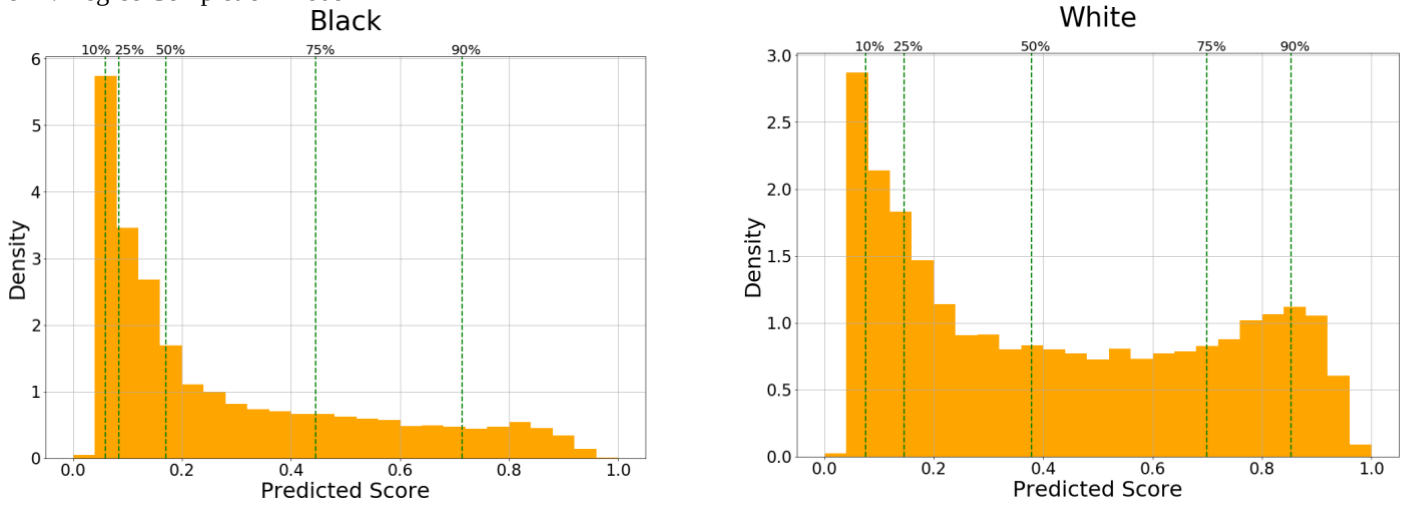
Notes: these plots show the amount of calibration bias (see notes from Figure 2) for logistic regression models using the same samples and set of predictors as the base Random Forest model.

Figure A2. Distribution of Predicted Scores in Base Model, by Race.

Panel A: Course completion model

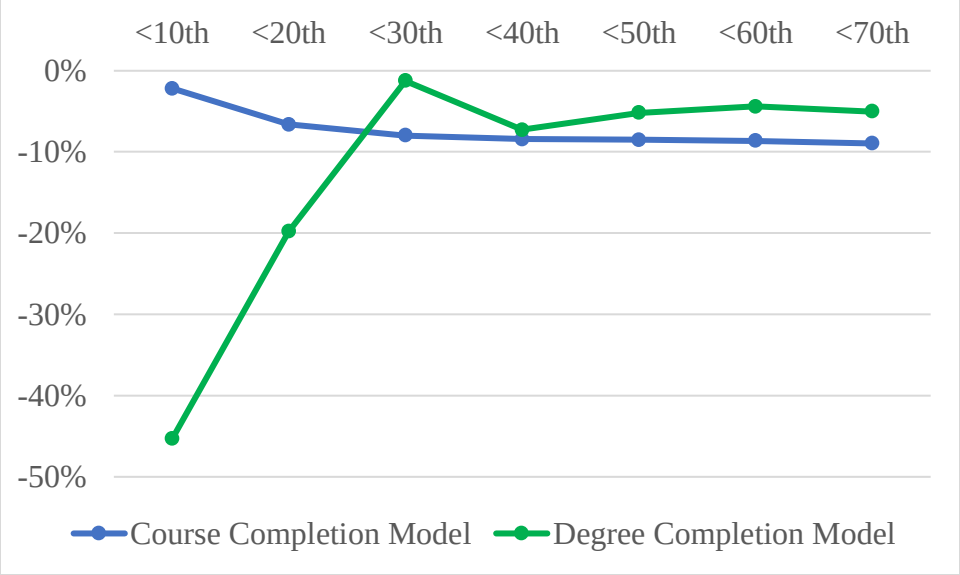


Panel B: Degree Completion Model



Notes: each plot shows the distribution of the raw predicted scores of observations from the relevant validation sample, separately for White and Black students. These predicted scores are generated using the Random Forest model built using the full training sample.

Appendix Figure A3. The Percentage Change in Black Students Labeled At-Risk after Simulating Calibration Bias Removal from the Random Forest Models.

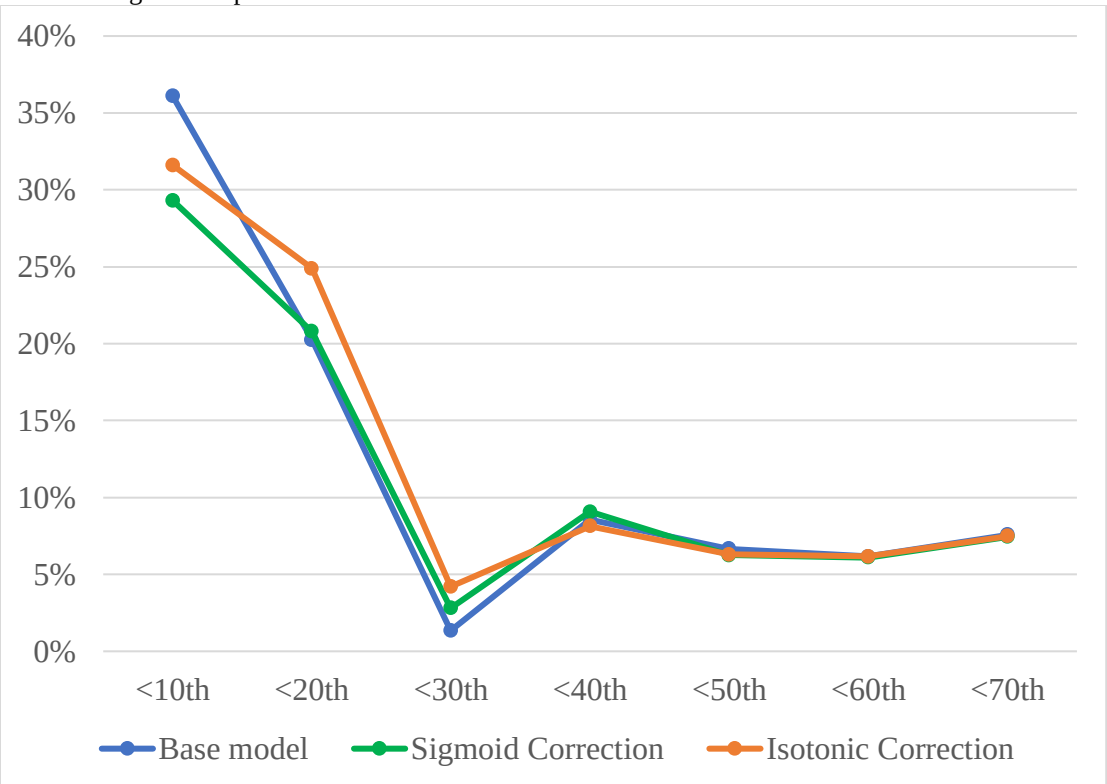


Notes: this plot shows, for a given risk threshold (e.g. < 30th percentile of predicted risk scores), the percentage decrease in the number of White students who would be labeled as at-risk after we simulate the calibration bias removal. Specifically, we calculate the amount of calibration by following these steps: (1) Select all White and Black students whose predicted scores are below the risk threshold; (2) if the actual success rate of the labeled at-risk Black students is lower than that of the labeled at-risk White students, change the label of the Black student whose predicted score is lowest among the Black students not labeled at-risk , and simultaneously drop the White student whose predicted score is highest among the White students labeled at-risk; (3) repeat step two until the actual success rates of the two selected groups become equal.

Appendix Figure A4. Calibration Bias after Sigmoid and Isotonic Corrections.

Panel A: Course Completion Model

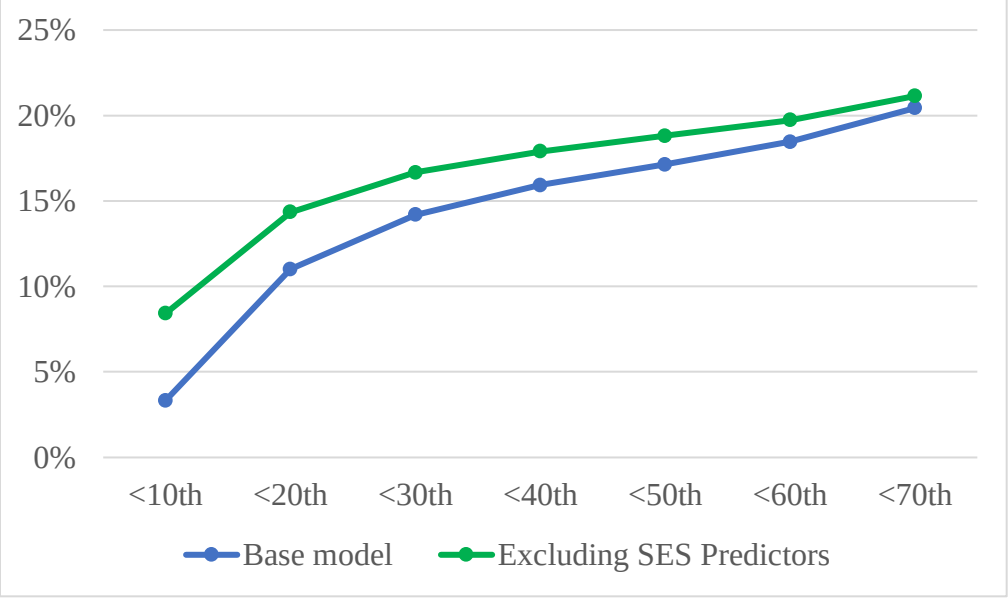
Panel B: Degree Completion Model



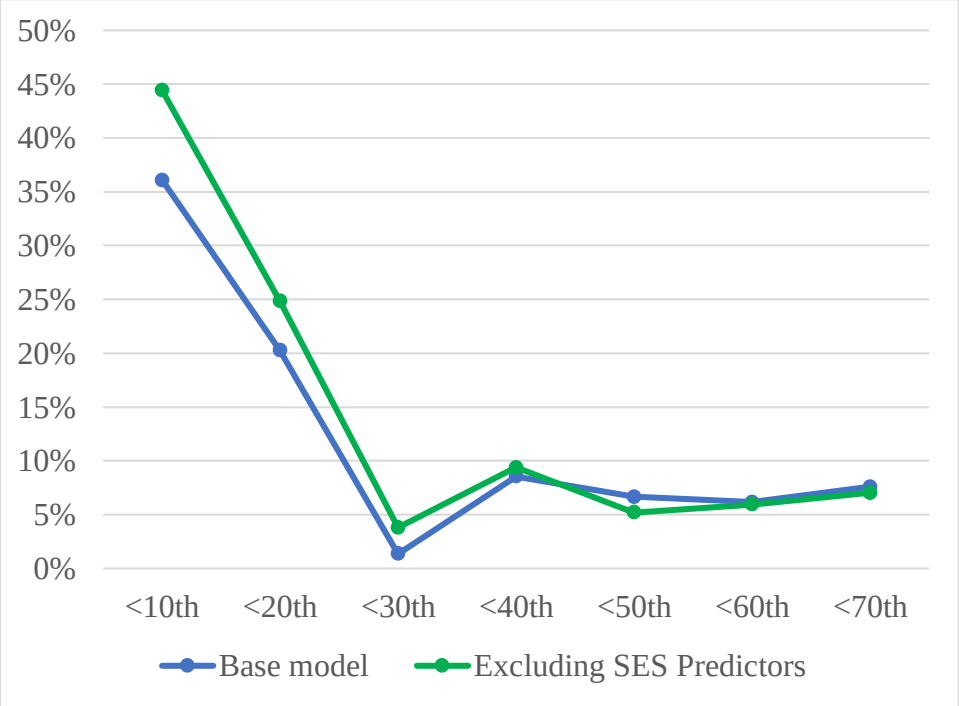
Notes: these plots compare the amount of calibration bias (measured by the increase in Black students targeted as at-risk in the simulated model, see Figure 2) for the base RF model and the versions that incorporate a post-processing calibration correction procedure (Sigmoid Correction, and separately, Isotonic Correction).

Appendix Figure A5. Change in Calibration Bias after Removing All Socio-Economic Status (SES) Predictors.

Panel A: Course Completion Model



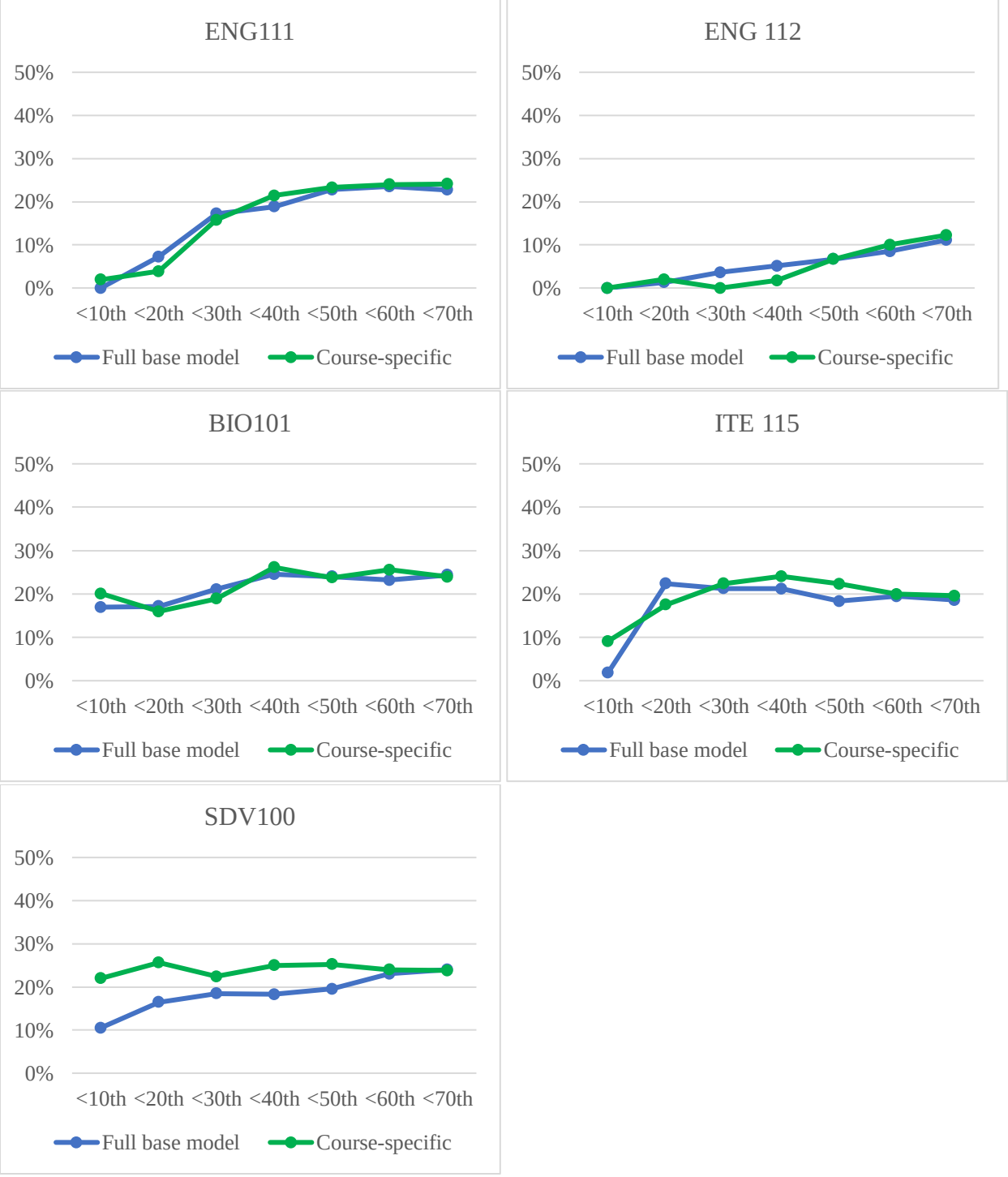
Panel B: Degree Completion Model



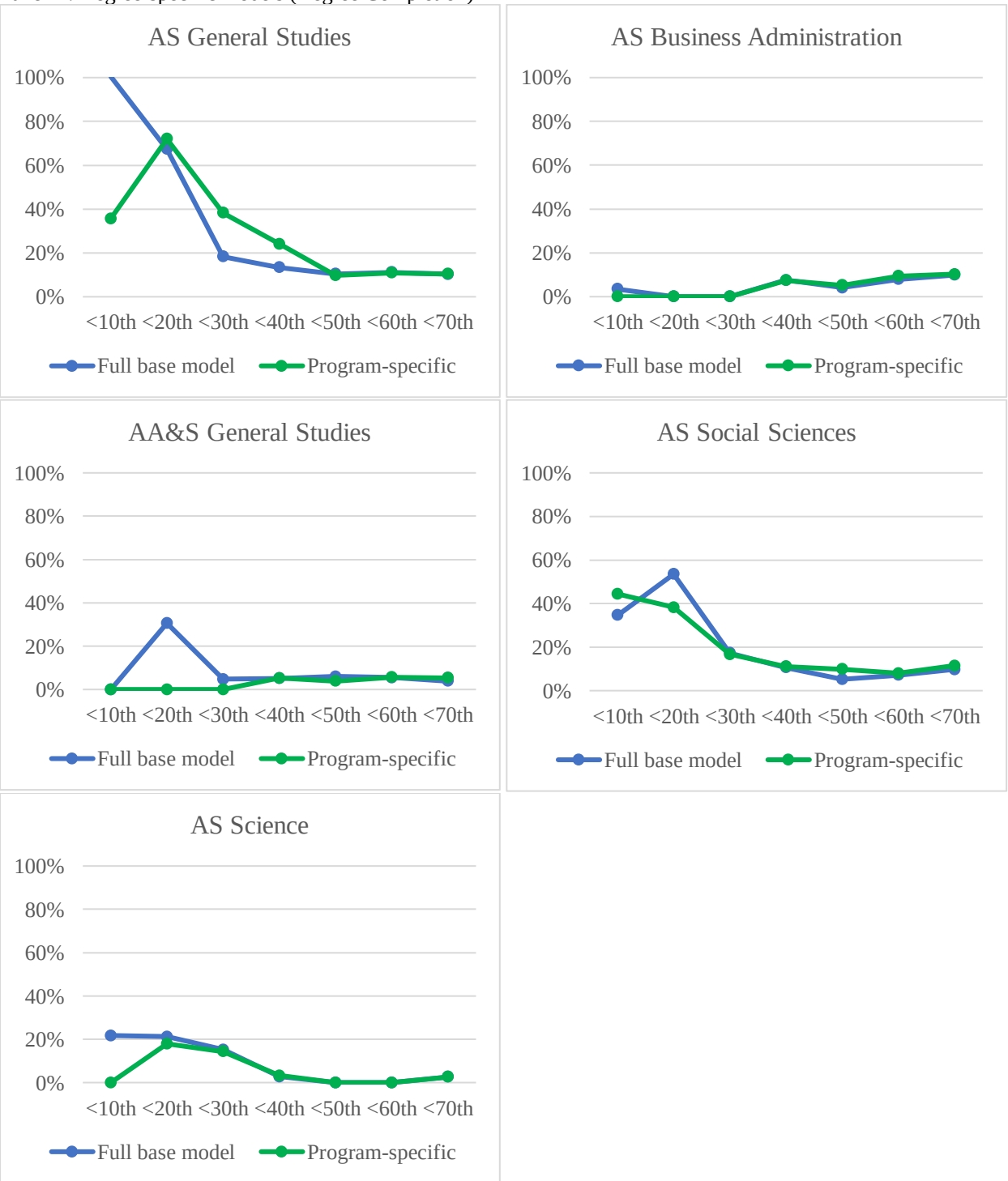
Notes: these plots compare the amount of calibration bias (measured by the increase in Black students targeted as at-risk in the simulated model, see Figure 2) for the base RF model and the RF models that do not include any socio-economic predictors (see Appendix Tables A3 and A4)

Appendix Figure A6. Calibration Bias in Course-Specific, Degree-Specific, and Program Level-Specific Models

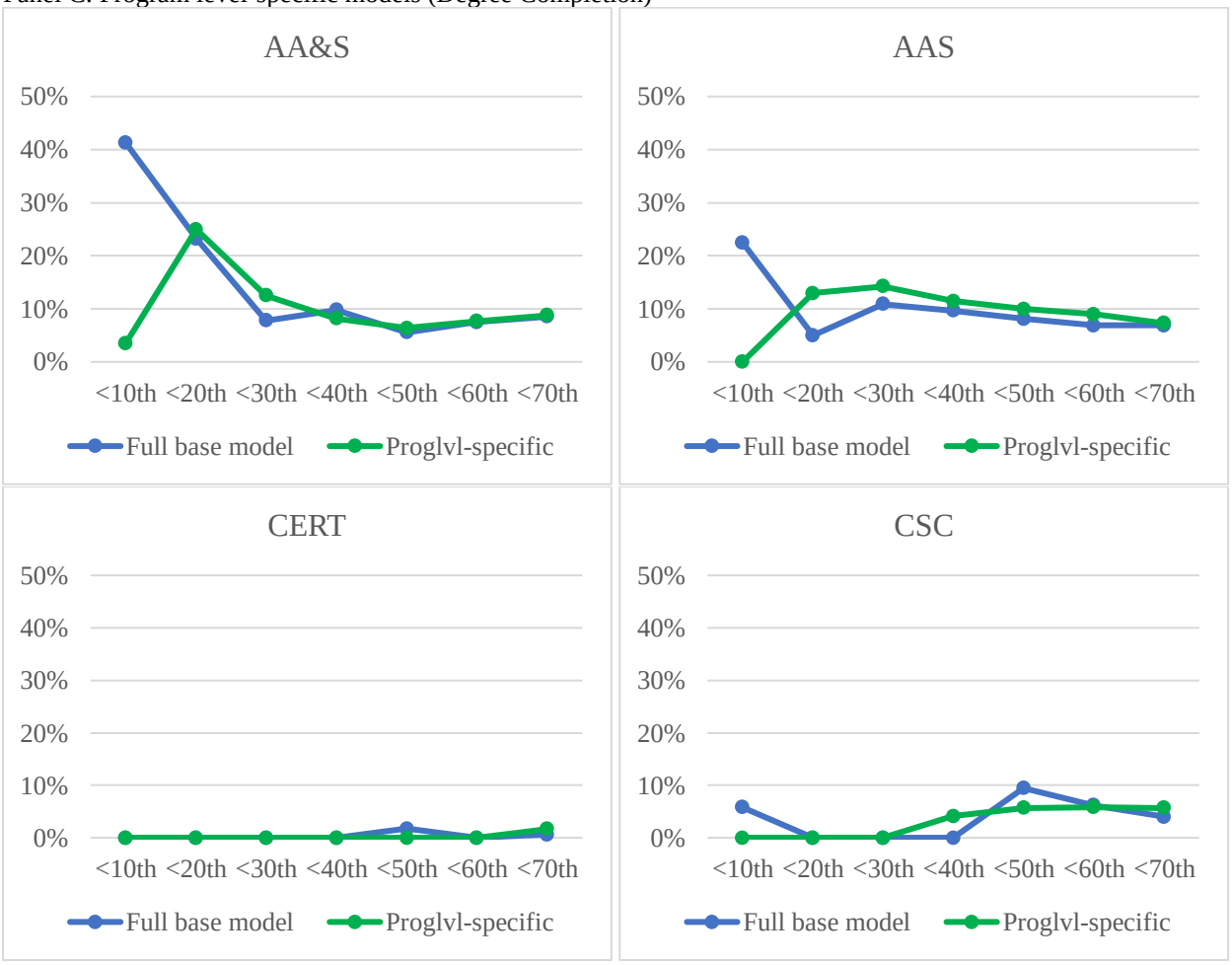
Panel A: Course-specific models (Course Completion)



Panel B: Degree-specific models (Degree Completion)

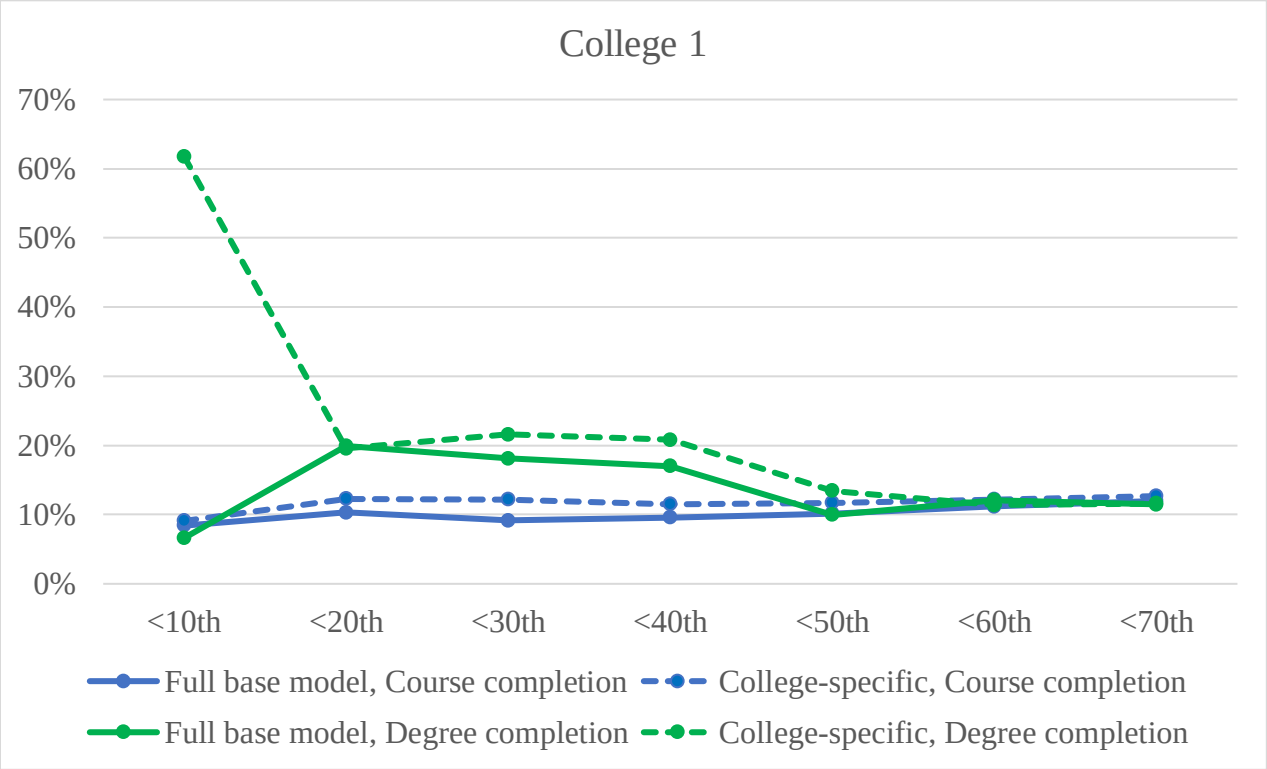


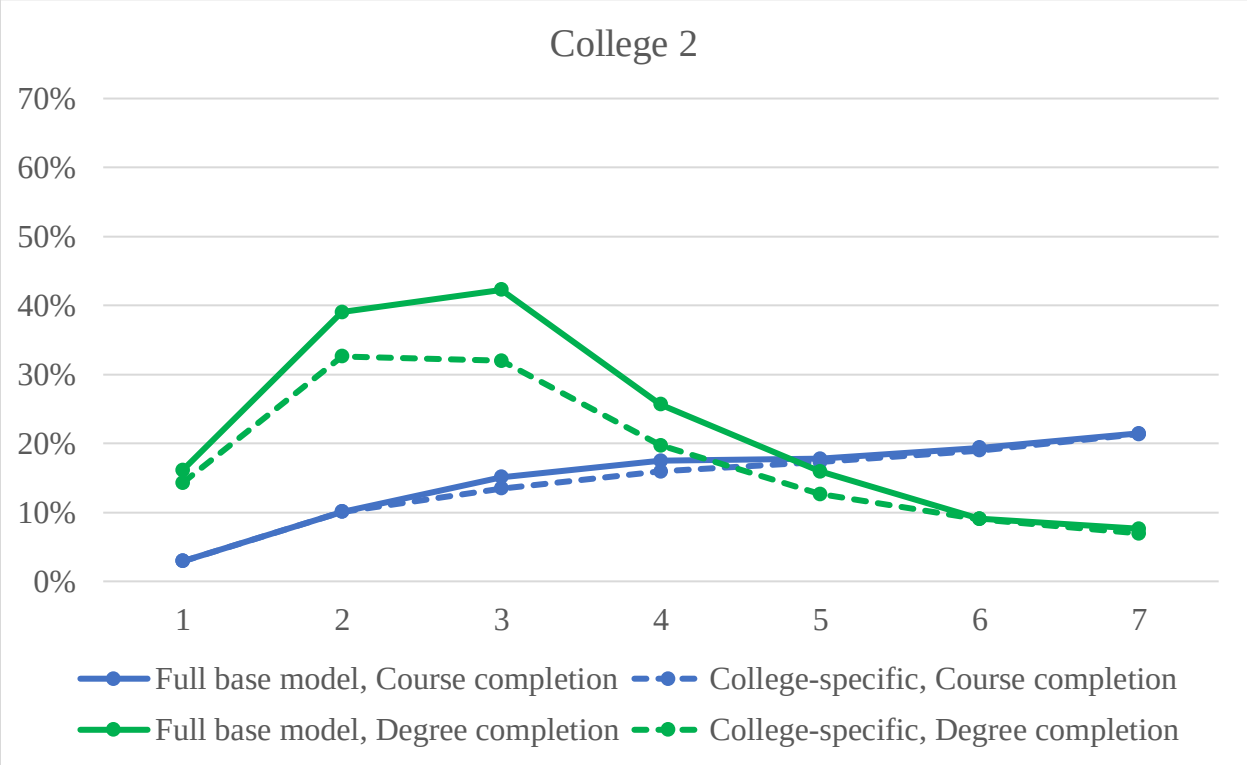
Panel C: Program level-specific models (Degree Completion)

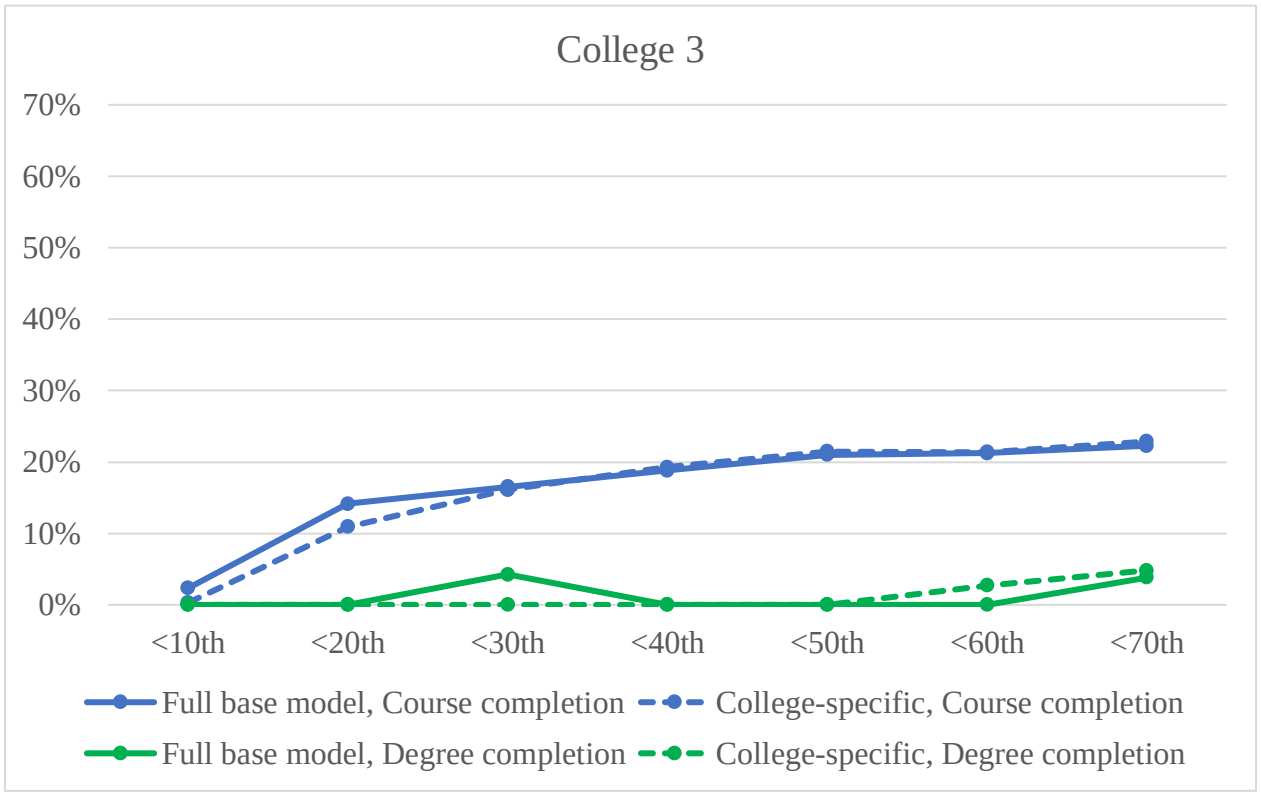


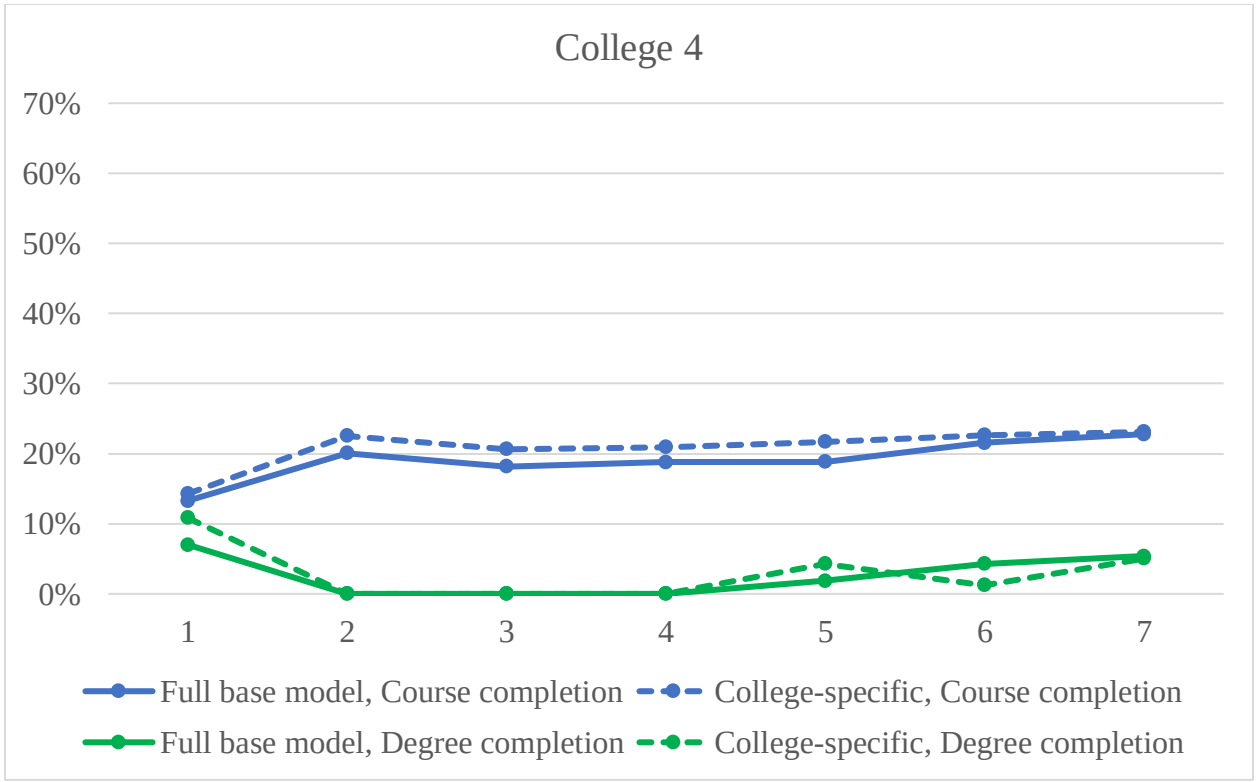
Notes: This figure plots the amount of calibration bias (measured by the increase in Black students targeted as at-risk in the simulated model, i.e. Figure 2) for the base RF model to the RF models built using a training sample restricted to observations from the particular course, degree, or program-level. For both the base models and the unit-specific models, the amount of calibration bias is calculated using the validation sample restricted to observations from the particular course, degree, or program-level.

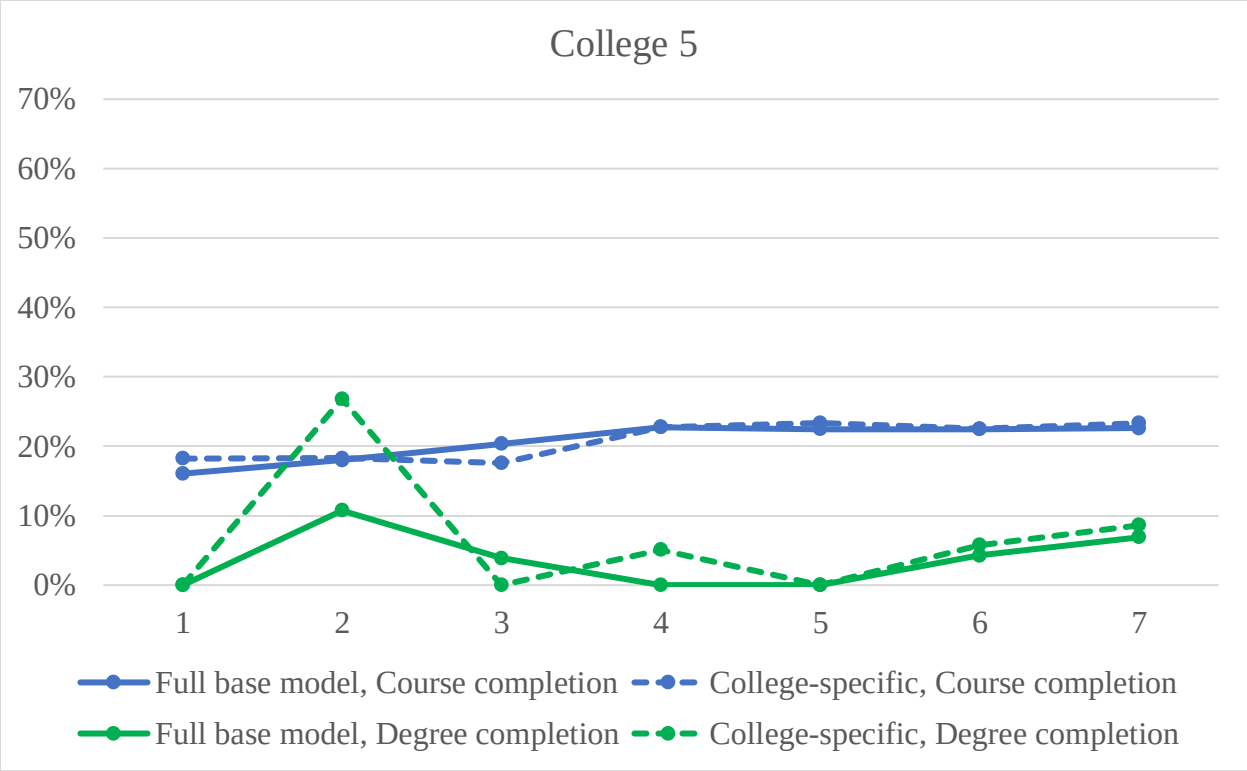
Appendix Figure A7. Calibration Bias in College-Specific Models.





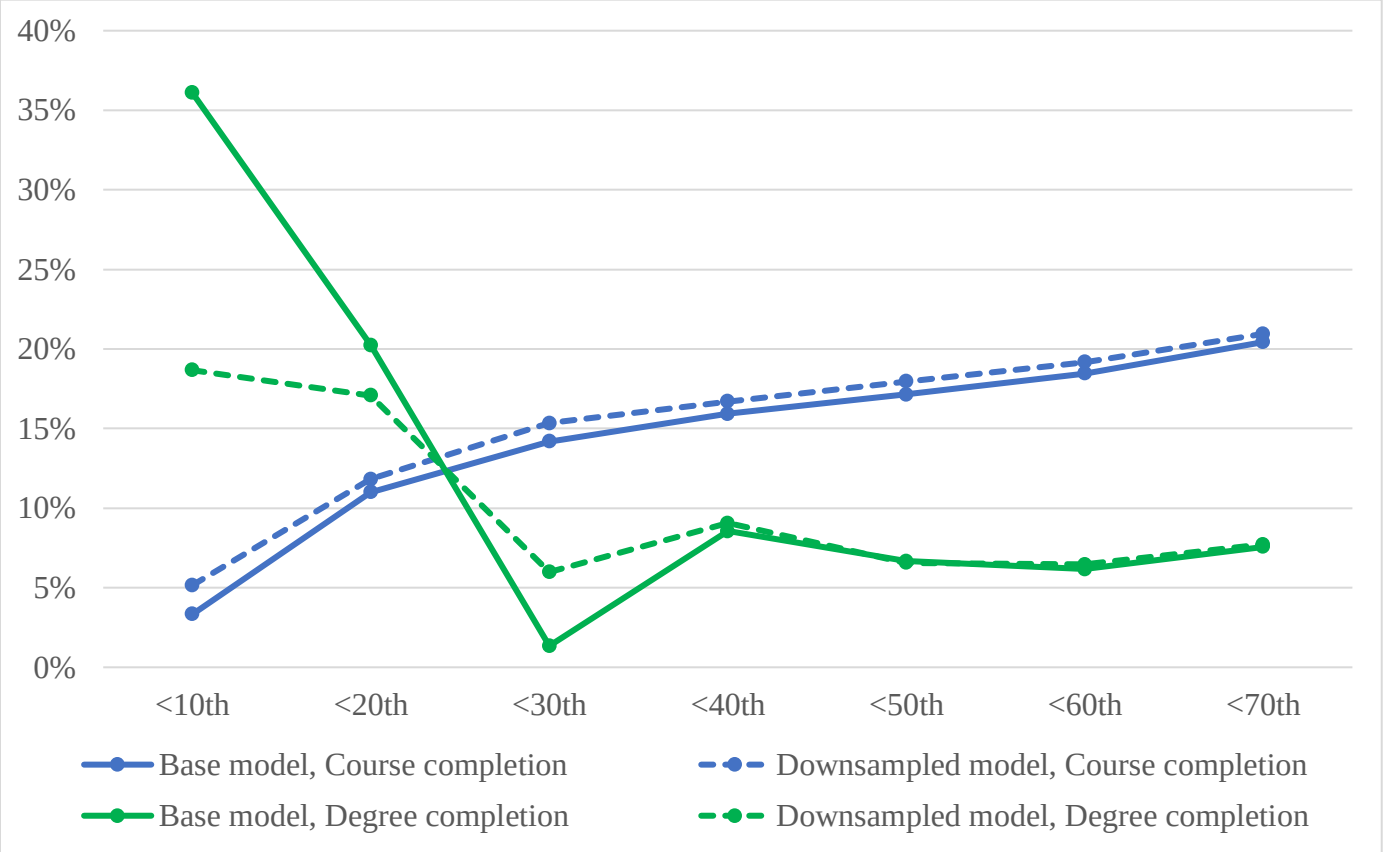

College 3
70%
60%
50%
40%
30%
20%
10%
0%
<10th
<20th
<30th
<40th
<50th
<60th
<70th
Full base model, Course completion
College-specific, Course completion
Full base model, Degree completion
College-specific, Degree completion


College 4
70%
60%
50%
40%
30%
20%
10%
0%
1
2
3
4
5
6
7
Full base model, Course completion
College-specific, Course completion
Full base model, Degree completion
College-specific, Degree completion



Notes: This figure plots the amount of calibration bias (measured by the increase in Black students targeted as at-risk in the simulated model, see Figure 2) for the base RF model to the RF models built using a training sample restricted to observations from a particular college. We build these college-specific models for the five largest VCCS colleges, based on enrollment counts. For both the base models and the unit-specific models, the amount of calibration bias is calculated using the validation sample restricted to observations from the particular college.

Appendix Figure A8. Calibration Bias in Down-Sampled Models.



Notes: This figure plots the amount of calibration bias (measured by the increase in Black students targeted as at-risk in the simulated model, i.e. Figure 2) for the base RF model to the down-sampled RF models, in which we down-sample the non-Black subgroup to be the same size as the Black subgroup.

Appendix Table A1. C-statistic of base logit models.

Panel A: course completion model

White	Black	% diff
0.8015	0.7787	-2.84%
(0.0007)	(0.0012)	

Panel B: degree completion model

White	Black	% diff
0.8938	0.8805	-1.49%
(0.0020)	(0.0037)	

Notes: standard errors in parentheses; all differences between White and Black are significant at the $p < 0.01$ level

Appendix Table A2. Optimized values of base Random Forest model parameters.

	Course Completion (1)	Degree Completion (2)
Maximum Tree Depth	34	16
Number of Trees	260	120
Number of random predictors used in node splitting	15	12

Notes: optimal model parameters chosen using five-fold cross-validation.

Appendix Table A3. Full list of predictors, course completion model.

Predictor description	Category	Available for 1st term
Average historical grade in the target course	Course characteristics	X
Average historical grade in the concurrent courses	Course characteristics	X
23 college indicators	Course characteristics	X
Course meeting time is in the evening	Course characteristics	X
Target course is 200-level	Course characteristics	X
Target course section is online	Course characteristics	X
Average grade in target course's prerequisites	Course characteristics	X
Enrollment in target course section	Course characteristics	X
Target course is in a Summer term	Course characteristics	X
Student is taking concurrent courses with historic grades available	Student's academic, course-specific	X
Student took the target course's prerequisites (if applicable)	Student's academic, course-specific	X
Student has previously taken the target course	Student's academic, course-specific	
Student's average prior grade in the target course (if repeating the course)	Student's academic, course-specific	
Has taken prior Arts courses (target course = X subject)	Student's academic, course-specific	
Average grade in prior Arts courses (target course = X subject)	Student's academic, course-specific	
Has taken prior Business/Finance courses (target course = X subject)	Student's academic, course-specific	
Average grade in prior Business/Finance courses (target course = X subject)	Student's academic, course-specific	
Has taken prior Engineering courses (target course = X subject)	Student's academic, course-specific	
Average grade in prior Engineering courses (target course = X subject)	Student's academic, course-specific	
Has taken prior Foreign Languages courses (target course = X subject)	Student's academic, course-specific	
Average grade in prior Foreign Languages courses (target course = X subject)	Student's academic, course-specific	
Has taken prior Humanities courses (target course = X subject)	Student's academic, course-specific	
Average grade in prior Humanities courses (target course = X subject)	Student's academic, course-specific	
Has taken prior Medical Sciences courses (target course = X subject)	Student's academic, course-specific	
Average grade in prior Medical Sciences courses (target course = X subject)	Student's academic, course-specific	

Has taken prior Mathematics courses (target course = X subject)	Student's academic, course-specific	
Average grade in prior Mathematics courses (target course = X subject)	Student's academic, course-specific	
Has taken prior Applied Technologies courses (target course = X subject)	Student's academic, course-specific	
Average grade in prior Applied Technologies courses (target course = X subject)	Student's academic, course-specific	
Has taken prior Natural Sciences courses (target course = X subject)	Student's academic, course-specific	
Average grade in prior Natural Sciences courses (target course = X subject)	Student's academic, course-specific	
Has taken prior Social Sciences courses (target course = X subject)	Student's academic, course-specific	
Average grade in prior Social Sciences courses (target course = X subject)	Student's academic, course-specific	
Age at time of target course enrollment	Student Demographic and SES	X
Gender	Student Demographic and SES	X
Parental education (3 categories: first-gen, not-first gen, missing)	Student Demographic and SES	X
Distance from student home to college (miles, based on ZCTA of target term)	Student Demographic and SES	X
Median household income of student home ZCTA during target term	Student Demographic and SES	X
Percent below poverty line of student home ZCTA during target term	Student Demographic and SES	X
Pell status during the target term (received, did not receive, did not apply)	Student Demographic and SES	X
Pell status throughout prior VCCS enrollment history	Student Demographic and SES	
Instructor works full-time at VCCS	Instructor characteristics	X
Instructor has taught the target course in the past	Instructor characteristics	X
Average grade assigned by the instructor in the target course	Instructor characteristics	X
Instructor has been teaching at VCCS for 6+ years	Instructor characteristics	X
15 field of study indicators (2 digit CIPs)	Student's academic characteristics, general	X
Enrolled in a transfer-oriented associate degree program	Student's academic characteristics, general	X
Enrolled in an occupation-oriented associate degree program	Student's academic characteristics, general	X
Enrolled in a certificate program	Student's academic characteristics, general	X
Enrolled in any development courses in the target term	Student's academic characteristics, general	X
# credits attempted in the target term	Student's academic characteristics, general	X
% attempted credits during target term that are evening	Student's academic characteristics, general	X
% attempted credits during target term that are the 200-level	Student's academic characteristics, general	X
% attempted credits during target term that are online	Student's academic characteristics, general	X

Total credits accumulated prior to target term	Student's academic characteristics, general
Cumulative GPA	Student's academic characteristics, general
Credits attempted in last term (prior to target term)	Student's academic characteristics, general
Slope of credits attempted in prior terms	Student's academic characteristics, general
Ever dually enrolled	Student's academic characteristics, general
Slope of term-level GPA in prior terms	Student's academic characteristics, general
Missing indicator for term GPA of the last term	Student's academic characteristics, general
Missing indicator for term GPA of the second-to-last term	Student's academic characteristics, general
# terms enrolled at VCCS prior to target term	Student's academic characteristics, general
% prior attempted credits completed	Student's academic characteristics, general
% prior attempted credits that were developmental courses	Student's academic characteristics, general
% prior attempted credits "Incomplete"	Student's academic characteristics, general
# stop-out terms between initial enrollment and target term	Student's academic characteristics, general
% prior attempted credits "Withdrawn"	Student's academic characteristics, general
Stddev of term-level credit completion rate	Student's academic characteristics, general
Term GPA of the last term prior to the target term	Student's academic characteristics, general
Term GPA of second-to-last term prior to the target term	Student's academic characteristics, general

Appendix Table A4. Full list of predictors, degree completion model.

Predictor description	Category	Available for 1st term
Age at initial enrollment	Student Demographic and SES	X
Gender	Student Demographic and SES	X
Parental education (first-gen, not-first gen, missing)	Student Demographic and SES	X
Distance from student home to college (miles, based on ZCTA of current term)	Student Demographic and SES	X
Median household income of student home ZCTA during current term	Student Demographic and SES	X
Percent below poverty line of student home ZCTA during current term	Student Demographic and SES	X
Pell status during the current term (received, did not receive, did not apply)	Student Demographic and SES	X
Pell status during the prior term	Student Demographic and SES	
Pell status throughout prior VCCS enrollment history	Student Demographic and SES	
Percentage of terms enrolled at VCCS through the last term	Non-term specific VCCS academics	
Cumulative GPA	Non-term specific VCCS academics	
Share of total credits earned (credits passed / credits attempted)	Non-term specific VCCS academics	
Average number of credits attempted during each enrolled term at VCCS	Non-term specific VCCS academics	
Standard deviation of term proportion of credits earned	Non-term specific VCCS academics	
Share of total credits withdrawn	Non-term specific VCCS academics	
Share of developmental credits among total credits attempted	Non-term specific VCCS academics	
Share of 200-level credits among total credits attempted	Non-term specific VCCS academics	
Trend of term enrollment intensity (term credits attempted)	Non-term specific VCCS academics	
Trend of term GPA	Non-term specific VCCS academics	
Ever repeated a course	Non-term specific VCCS academics	
Ever dually enrolled at VCCS	Academics prior to initial VCCS enrollment	X
College-level credit hours accumulated	Academics prior to initial VCCS enrollment	X
Cumulative GPA prior to initial enrollment	Academics prior to initial VCCS enrollment	X
Share of total credits earned	Academics prior to initial VCCS enrollment	X
Enrolled in any non-VCCS institutions in the past 3 years	Academics prior to initial VCCS enrollment	X

Number of terms enrolled at non-VCCS institutions	Academics prior to initial VCCS enrollment	X
Seamless enrollee indicator (if student enrolled in the same year as HS graduation)	Academics prior to initial VCCS enrollment	X
Ever enrolled in non-VCCS colleges since initial enrollment	Non-term specific non-VCCS academics	
Total number of enrolled terms at non-VCCS	Non-term specific non-VCCS academics	
Total number of non-VCCS colleges attended	Non-term specific non-VCCS academics	
Non-VCCS institution type ever attended (sector x level x in-state)	Non-term specific non-VCCS academics	
Admission rates of institutions attended (averaged & weighted if multiple)	Non-term specific non-VCCS academics	
Graduation rates of institutions attended (averaged & weighted if multiple)	Non-term specific non-VCCS academics	
25th and 75th percentiles of SAT scores, by subject (averaged & weighted if multiple)	Non-term specific non-VCCS academics	
Average grants received by all enrolled terms at VCCS	Non-term specific financial aid	X
Average subsidized loans received by all enrolled terms at VCCS	Non-term specific financial aid	X
Average unsubsidized loans received by all enrolled terms at VCCS	Non-term specific financial aid	X
Average other aids received by all enrolled terms at VCCS	Non-term specific financial aid	X
Indicator for whether the student was actively enrolled in VCCS or not	Term-specific VCCS academics	X
Credits attempted	Term-specific VCCS academics	
Share of credits earned	Term-specific VCCS academics	
Term GPA	Term-specific VCCS academics	
Proportion of credits withdrawn	Term-specific VCCS academics	
Proportion of development credits among credits attempted	Term-specific VCCS academics	X
Proportion of 200-level credits among credits attempted	Term-specific VCCS academics	X
Repeating a previously attempted course in the current term or not	Term-specific VCCS academics	
Degree-seeking of not	Term-specific VCCS academics	X
Attended any non-VCCS institution	Term-specific non-VCCS academics	X
Total enrollment intensity in non-VCCS institutions	Term-specific non-VCCS academics	X
Amount of grants received	Term-specific financial aid	X
Amount of subsidized loans received	Term-specific financial aid	X
Amount of unsubsidized loans received	Term-specific financial aid	X
Amount of other aid received	Term-specific financial aid	X

Appendix Table A5. Subgroup difference in top 10 predictors.

Panel A: Course completion

	White mean (1)	Black mean (2)	Black - White % diff (3)
# credits attempted in the target term	10.62	9.99	-5.9%
Cumulative GPA	2.98	2.61	-12.4%
Average historical grade in the target course	2.85	2.81	-1.7%
Term GPA in last term (prior to target term)	2.96	2.53	-14.4%
Average grade assigned by the instructor in the target course	2.84	2.80	-1.6%
Average historical grade in the concurrent courses	2.84	2.81	-1.2%
Median households income corresponding to the zcta	$80,034	$74,931	-6.4%
% below poverty corresponding to the zcta	0.11	0.12	16.6%
% prior attempted credits completed	0.90	0.84	-7.1%

Panel B: Degree completion

	White mean (1)	Black mean (2)	Black - White diff (3)
Cumulative GPA	2.60	2.05	-21.3%
% prior attempted credits completed	0.80	0.66	-16.9%
Standard deviation of term-level % attempted credits completed	0.16	0.21	26.7%
Proportion of credits withdrawn, first Spring term	0.82	0.69	-15.7%
% prior credits withdrawn	0.11	0.15	42.0%
Term GPA, first Fall term	2.63	2.20	-16.2%
Term GPA, first Spring term	2.60	2.12	-18.6%
Standard deviation of % term credits withdrawn	0.12	0.15	17.8%
Proportion of credits completed, first Fall term	0.82	0.70	-14.3%
% prior developmental credits	0.11	0.22	101.9%

Notes: This table compares the mean values of the top 10 predictors of the course and degree completion models, as determined by the feature importance score. All mean differences between the Black and White samples are statistically significant with a p-value < 0.01

Appendix Table A6. Summary statistics, by race and training/validation split.

Panel A: Course completion model

	Black		White	
	Training	Validation	Training	Validation
	(1)	(2)	(3)	(4)
Course completion rate	67.4%	69.4%	78.5%	79.8%
Age	27.3	25.9	24.5	23.8
Male	39.5%	39.0%	45.6%	44.9%
Not first-generation	44.6%	38.0%	62.2%	51.3%
First-generation	24.5%	18.8%	21.3%	16.1%
Completed FAFSA, no Pell	17.6%	18.1%	19.6%	21.2%
Pell recipient	54.8%	51.5%	32.7%	31.7%
ZCTA median household income	$66,128	$74,950	$72,920	$80,124
ZCTA percent below poverty	13.7%	12.3%	11.1%	10.5%
ZCTA distance to college (miles)	23.1	23.4	31.2	31.8
Cumulative GPA	2.6	2.6	2.9	3.0
Cumulative credits earned	29.4	29.8	32.7	33.1
Percent of prior credits withdrawn	7.4%	7.0%	6.1%	5.7%
% of prior credits from developmental courses	15.2%	12.4%	7.1%	5.5%
Ever dual enrollment students	12.1%	15.7%	27.2%	31.8%
Number of prior VCCS terms	4.1	4.1	4.0	4.0
Credits enrolled	9.9	10.0	10.5	10.6
AA&S	56.8%	62.3%	56.6%	58.9%
AAS	28.9%	24.7%	24.2%	23.7%
CERT	10.2%	8.6%	10.2%	9.1%
Online course	29.6%	34.2%	30.1%	34.9%
Evening course	13.6%	11.6%	11.8%	9.3%
200 level course	30.7%	32.4%	32.0%	34.1%
Developmental course	18.2%	15.0%	9.9%	8.1%
Course section enrollment	22.1	21.3	22.7	22.0
N observations	904,152	139,076	2,392,786	394,447

Panel B: Degree completion model

Degree completion rate	24.1%	22.4%	37.7%	39.2%

Age (entry)	25.9	26.3	24.4	23.8
Male	41.0%	44.7%	45.9%	46.9%
Not first-generation	29.9%	38.7%	40.6%	58.2%
First-generation	20.5%	27.5%	18.9%	24.1%
Completed FAFSA, no Pell (prior term)	49.6%	56.8%	27.3%	33.2%
Pell recipient (prior term)	13.5%	13.6%	14.4%	16.5%
ZCTA median household income	$60,321	$60,810	$68,415	$70,221
ZCTA percent below poverty	14.1%	14.7%	11.1%	11.0%
ZCTA distance to college (miles)	26.2	34.8	30.6	30.3
Cumulative GPA	2.3	2.2	2.7	2.7
Cumulative credits earned	1.0	0.9	2.3	2.3
Number of terms at VCCS	4.5	4.3	4.7	4.8
Percent of prior credits completed	72.9%	72.0%	84.0%	84.0%
GPA trend	-0.20	-0.19	-0.13	-0.14
Credits attempted 1st Fall	9.2	9.3	10.0	10.1
Credits attempted 1st Spring	9.0	9.0	9.5	9.6
Any NSC enrollment	5.5%	5.3%	5.4%	5.6%
AA&S	49.4%	50.5%	54.2%	57.6%
AAS	25.3%	27.2%	22.6%	21.9%
CERT	6.1%	5.0%	6.5%	5.3%
CSC	11.9%	10.6%	9.7%	8.9%
N observations	82,784	16,796	179,919	32,288

Notes: all measurements are made at the point of the beginning of the target term, unless otherwise specified.

Appendix Table A7. Top 10 predictors of race-specific models.

Panel A: Course completion model

White		Black	
Predictor	FI Score	Predictor	FI Score
# credits attempted in the target term	0.073	# credits attempted in the target term	0.061
Cumulative GPA	0.049	Average historical grade in the target course	0.048
Term GPA in last term (prior to target term)	0.048	Cumulative GPA	0.048
Average grade assigned by the instructor in the target course	0.045	Average grade assigned by the instructor in the target course	0.047
Average historical grade in the target course	0.044	Term GPA in last term (prior to target term)	0.040
	0.037	Average historical grade in the concurrent courses	0.036
Average historical grade in the concurrent courses Median households income corresponding to the zcta	0.030	% prior attempted credits completed	0.030
% below poverty corresponding to the zcta	0.030	Median households income corresponding to the zcta	0.030
% prior attempted credits completed	0.028	% below poverty corresponding to the zcta	0.029
Distance to college	0.028	Enrollment in target course section	0.028

Panel B: Degree completion model

White		Black	
Predictor	FI Score	Predictor	FI Score
Cumulative GPA	0.064	Cumulative GPA	0.062
% prior attempted credits completed	0.061	% prior attempted credits completed	0.058
Standard deviation of % term attempted credits completed	0.034	Standard deviation of % term attempted credits completed	0.034

% attempted credits completed in the 1st Spring term	0.028	Share of total credits withdrawn	0.028
Term GPA in the 1st Spring term	0.027	Term GPA in the 1st Fall term	0.028
Term GPA in the 1st Fall term	0.026	% attempted credits completed in the 1st Spring term	0.025
Share of total credits withdrawn	0.026	Term GPA in the 1st Spring term	0.022
Standard deviation of % term credits withdrawn	0.026	Proportion of development credits among credits attempted	0.021
Enrollment intensity in the 1st Spring term	0.021	Standard deviation of % term credits withdrawn	0.021
Trend of term GPA	0.021	Term GPA in the 2nd Fall term	0.019

Notes: This table shows the top ten predictors, based on the feature importance (FI) score, of the race-specific RF models.

Appendix Table A8. Relationship between share Black enrollment and success rate.

	Course Completion	Degree completion
	(1)	(2)
Success rate	0.1633***	-0.273***
	(0.019)	(0.063)
Level of data	College x course	College x program of study
R-squared	0.004	0.038
N	5610	475

Notes: results from a regression of the share of Black students enrolled in a particular course or degree program on the success rate among White students in that course or program.

Appendix Table A9. C-statistics of course-specific, degree-specific, and program level-specific models (standard errors in parentheses).

Panel A: Course completion model

ENG 111			SDV 100			ENG 112			ITE 115			BIO 101		
White	Black	%diff	White	Black	%diff	White	Black	%diff	White	Black	%diff	White	Black	%diff
0.757	0.7449	-1.6%	0.778	0.7563	-4.1%	0.8017	0.7692	-1.7%	0.7672	0.7587	-1.1%	0.7938	0.7805	-1.7%
(0.0039)	(0.0058)		(0.0043)	(0.0063)		(0.0041)	(0.0068)		(0.0052)	(0.0076)		(0.0051)	(0.0085)	

Panel B: Degree Completion model -- Program-specific

AS General Studies			AS Business Admin			AA&S General Studies			AS Social Sciences			AS Science		
White	Black	%diff	White	Black	%diff	White	Black	%diff	White	Black	%diff	White	Black	%diff
0.8811	0.8884	0.8%	0.8847	0.8848	0.0%	0.9144	0.914	-0.0%	0.8903	0.907	1.9%	0.8751	0.8629	-1.4%
(0.0068)	(0.0111)		(0.0079)	(0.0121)		(0.0058)	(0.0149)		(0.0074)	(0.0106)		(0.0082)	(0.0141)	

Panel C: Degree Completion model -- Program-level-specific

AA&S			AAS			CERT			CSC		
White	Black	%diff	White	Black	%diff	White	Black	%diff	White	Black	%diff
0.9026	0.8941	-0.9%	0.904	0.8956	-0.9%	0.902	0.8855	-1.8%	0.8469	0.8733	3.1%
(0.0025)	(0.0048)		(0.0042)	(0.0072)		(0.0094)	(0.0173)		(0.0087)	(0.0119)	

Appendix Table A10. C-statistics of down-sampled model (standard errors in parentheses).

Panel A: course completion model

White	Black	% diff
0.8217	0.8003	-2.60%
(0.0007)	(0.0012)	

Panel B: degree completion model

White	Black	% diff
0.8944	0.885	-1.05%
(0.0020)	(0.0037)	

Appendix Table A11. C-statistics of first-term and returning subgroup-specific models (standard errors in parentheses).

Panel A: Course completion model

First-term specific model			Returning-specific model		
White	Black	%diff	White	Black	%diff
0.7547	0.7376	-2.27%	0.8402	0.8161	-2.87%
(0.0021)	(0.0032)		(0.0007)	(0.0012)	

Panel B: Degree Completion model

First-term specific model			Returning-specific model		
White	Black	%diff	White	Black	%diff
0.8945	0.8692	-2.83%	0.8885	0.8832	-0.60%
(0.0056)	(0.0099)		(0.0022)	(0.0040)	

Appendix Table A12. College specific models.

Panel A: Course completion model

College 1			College 2			College 3			College 4			College 5		
White	Black	%diff	White	Black	%diff	White	Black	%diff	White	Black	%diff	White	Black	%diff
0.8156	0.7978	-2.2%	0.8289	0.7975	-3.8%	0.8134	0.8015	-1.5%	0.8189	0.8074	-1.4%	0.8171	0.8101	-0.9%
(0.0015)	(0.0022)		(0.0018)	(0.0024)		(0.0031)	(0.0038)		(0.0032)	(0.0040)		(0.0029)	(0.0043)	

Panel B: Degree Completion model

College 1			College 2			College 3			College 4			College 5		
White	Black	%diff	White	Black	%diff	White	Black	%diff	White	Black	%diff	White	Black	%diff
0.8857	0.8819	-0.4%	0.8543	0.884	3.5%	0.8918	0.895	0.4%	0.8638	0.8275	-4.2%	0.8887	0.8907	0.2%
(0.0043)	(0.0072)		(0.0060)	(0.0070)		(0.0086)	(0.0119)		(0.0096)	(0.0146)		(0.0097)	(0.0166)	