EdWorkingPaper No. 23-717

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

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. Our results show that algorithmic bias in both models could result in at-risk Black students receiving fewer success resources than White students at comparatively lower-risk of failure. We also find the magnitude of algorithmic bias to vary within the distribution of predicted success. With the degree completion model, the amount of bias is nearly four times higher when we define at-risk using the bottom decile than when we focus on students in the bottom half of predicted scores. Between the two models, the magnitude and pattern of bias and the efficacy of basic bias mitigation strategies differ meaningfully, emphasizing 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 less useful at predicting Black student success compared with White student success, particularly for new students; this suggests that additional data collection efforts have the potential to mitigate bias.

VERSION: February 2023

Suggested citation: Bird, Kelli A., Benjamin L. Castleman, and Yifeng Song. (2023). Are Algorithms Biased in Education? Exploring Racial Bias in Predicting Community College Student Success. (EdWorkingPaper: 23-717). Retrieved from Annenberg Institute at Brown University: https://doi.org/10.26300/yd7z-6e20

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. Our results show that algorithmic bias in both models could result in at-risk Black students receiving fewer success resources than White students at comparatively lower-risk of failure. We also find the magnitude of algorithmic bias to vary within the distribution of predicted success. With the degree completion model, the amount of bias is nearly four times higher when we define at-risk using the bottom *decile* than when we focus on students in the bottom *half* of predicted scores. Between the two models, the magnitude and pattern of bias and the efficacy of basic bias mitigation strategies differ meaningfully, emphasizing 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 less useful at predicting Black student success compared with White 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.

Introduction

Predictive analytics are increasingly pervasive in higher education. By one estimate, nearly 40 percent of higher education institutions now utilize some form of predictive analytics to identify students at risk of dropping out before earning their degree (Barhsay & Aslanian, 2019; Ekowo & Palmer, 2016, Swaak, 2022). The use of predictive analytics in early-alert systems, which flag potentially struggling students within a course, 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 being mis-classified as at-risk or not (Lee & 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 will succeed, and more likely to predict White and Asian students will succeed when they in fact will stop out or struggle (Anderson, Boodhwani, & Baker, 2019, Jiang & Pardos, 2021, Jeong et al, 2021, Yu, Lee, & Kizilcec, 2021, Yu et al, 2020). This algorithmic bias has potential

[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.

important implications for educational policy and practice, since it could result in inefficient targeting of resources to students who are in fact likely to succeed absent additional intervention or who are not as likely to struggle as other students.

Prior research provides important demonstration of the presence of bias in algorithms commonly used in education. 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. As such, we are not aware of prior papers that have discussed algorithmic bias with policy makers, practitioners, and researchers as the primary audiences.

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 in the context of Virginia's state-wide community college system. The first algorithm predicts course completion and the second predicts degree completion; both models use data from the same student population and source. Colleges are most likely to use these types of algorithms to identify students at-risk of not completing a course or of dropping out (henceforth "at-risk"), and to target additional resources (e.g. advising, tutoring) to these students. Therefore, we intentionally focus our analysis and discussion on the types of algorithmic bias that would lead to fewer resources being allocated to at-risk students from historically-disadvantaged groups, relative to at-risk students in the majority. In particular, we focus our investigation of algorithmic bias on the racial margin of Black and White students, given growing efforts by both government and institutional leaders to promote greater racial equity in postsecondary education (e.g. U.S. Department of Education, 2022).

We focus specifically on algorithmic bias along two dimensions: calibration and accuracy. Calibration bias occurs when Black and White students have different actual success rates, conditional on predicted scores.[2] To assess accuracy bias, we compare c-statistics (also known as AUC) of the prediction models for Black versus White students. We then explore multiple hypotheses regarding why these forms of algorithmic bias would be present in our models. First, we test the hypothesis that models simply need to be unaware or blinded of race to be unbiased by comparing our base models (which do not include race) with models that do account for race. Next, we test the hypothesis that data underrepresentation causes algorithmic bias by building a separate model using a sample that over-represents Black students. Finally, we explore whether differential sorting of students by race into programs of study or courses can account for the algorithmic bias by estimating program-specific and course-specific 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. Whereas prior research has focused on overall algorithmic bias, we extend this work by demonstrating that bias is more pronounced at certain points in the distribution of predicted success. In both the course completion model and degree completion model, we find that among students lower in the distribution of predicted success, Black students have lower actual success rates than White students, holding constant predicted scores. In practice, this result means that some White students at lower risk would be targeted for additional resources while some Black students at relatively higher risk would not. For the course completion model, the magnitude of the bias is consistent across the distribution of predicted success. However, the degree completion model exhibits substantially higher bias among the most at-risk populations: the

[2] We quantify the amount of calibration bias following the procedure in Obermeyer et al. (2019), where we simulate how the targeting of students would differ between our actual model and an unbiased simulation.

amount of bias is nearly four times higher when we define at-risk using the bottom *decile* than when we focus on students in the bottom *half* of predicted scores. Separately, we find that overall model accuracy is lower for Black students in both models, indicating a higher rate of mis-classifying student risk-level.

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 highlights the highly contextual nature of efforts to mitigate 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.[3] 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 students (i.e. those with no prior VCCS enrollment history) is 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 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.

[3] The selective labels problem occurs when the outcome measure used in a predictive model is confounded by human decision making, which can make the outcome less comparable across subgroups of interest (e.g. Lakkaraju et al. 2017).

Our paper makes several important contributions. First, our primary focus is on generating insights that are relevant for policy makers, practitioners, and researchers less engaged with the data science community. Our approach is keeping 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. Specifically, when considering racial differences in false positive rates, prior studies conclude that algorithms are *overall* more "pessimistic" about Black student performance and relatively overestimate their risk-level (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. However, our focus of quantifying bias on the margin of being considered at-risk reverses this takeaway: we find that conditional on predicted score, our algorithms relatively underestimate Black students' risk-level.[4] 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

[4] This alternative interpretation is not due to contextual differences: 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 overestimating Black student risk-level.

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.

Finally, whereas prior work has focused mainly on investigating the technical aspects of algorithmic bias in education (e.g. developing novel approaches to bias mitigation), 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 removing bias from an existing algorithm, and to inform future data collection and modeling efforts.

Data and Methods

The data for this study come from VCCS system-wide administrative records from over 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 credentials 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 logistic regression for both prediction models. 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 logistic regression to more complex modeling strategies, such as random forest or recursive neural networks.[5] Given our goal of translating data science insights for a broader set of researchers and policy makers, we also privilege the logistic regression model because of wider familiarity and ease of use. We convert the predicted score from the logistic regression models to a binary prediction of course or degree completion. Specifically, we use a threshold equal to the training sample outcome mean, such that if a student's predicted score is above the average course or degree completion rate, then we classify that as a "positive" prediction -- i.e. that the model predicts the student will successfully complete the course or degree. We set the predicted success threshold based on the overall success rate in the training samples (75.4 percent course completion; 34.0 percent degree completion). 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.[6] Based on this definition, 75.6 percent of student x course observations achieved the outcome of course completion.

[5] We also built Random Forest (RF) versions of the base course and degree completion models. Broadly speaking, the RF results are similar to logistic. For the course completion model, the amount of calibration bias is slightly lower in RF compared to logistic for the riskiest thresholds (e.g. <10th percentile); conversely, the amount of calibration bias is higher in RF compared to logistic for the riskiest thresholds. For the higher thresholds (e.g. <50th percentile), the amount of calibration bias is similar between RF and logistic. In terms of overall model performance, the RF performs slightly better than the logistic (roughly 1% increase in c-statistic); however, the gaps in c-statistic between White and Black students from the RF (3.3% for course completion and 1.3% for degree completion) is quite similar to the gaps in the logistic. Appendix Figure A1 and Appendix Table A1 show these results.

[6] While a grade of D earns the student credit for the course and is 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.

The course completion sample consists of student-by-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.[7] 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 287 predictors from the VCCS administrative data. We describe them at high-level here, and include a full list in Appendix Table A2. The predictors include student demographics (e.g. age); 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, and therefore they have no academic history, we are only able to construct predictors for demographics, student enrollment characteristics of the target term, characteristics of the target course, and instructor characteristics. We handle missing values for predictors related to students' academic history by including a binary predictor indicating whether the observation is for a student's first term at VCCS, and for all first-term observations, we set the values of all predictors that cannot be constructed to zero.

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

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, 34.0 percent of students in our full sample completed a degree. Our sample consists of students who enrolled at a VCCS college as a degree-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.[8] 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).

We construct 294 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 demographics (e.g. age at initial enrollment); 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

[8] For a full discussion of the choice of outcome and this truncation method, see Bird et al (2021).

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).[9] We handle missing values by including indicators for the two potential reasons of missingness (if the term is outside the observation window after truncation, or if the student is not actively enrolled in VCCS during the term), and set the values of all relevant predictors to zero.

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).[10] 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 investigation on algorithmic biases that would result in less resource allocation to at-risk students from historically-disadvantaged groups.

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 outcomes for Black students

[9] 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.

[10] 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.

versus White students? This is referred to as “calibration bias”. To identify calibration bias, we compare actual success rates across subgroups, conditional on the algorithm-generated predicted score. 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 at-risk 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 more at-risk than White students. This calibration bias would result in some at-risk Black students not receiving the additional resources, even though they have the same or higher actual risk level as 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 “at-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 selected Black students is statistically lower than that of the selected White students, add the Black student whose predicted score is lowest among the Black students who are not selected to the selected group, and simultaneously drop the White student whose predicted score is highest among the White students who are selected; (3) repeat step two until the actual success rates of the two selected groups become equal. We then compare the numbers and ratios of Black to White students who would be targeted for intervention both before and after the calibration bias is eliminated. 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.

The second form of algorithmic bias we consider is whether algorithms are equally accurate at identifying at-risk students for Black students as they are for White students. To assess this "accuracy bias", we compare several metrics across subgroups. To measure differences in overall accuracy, we report subgroup-specific c-statistics.[11] Lower levels of model accuracy imply that more students would be mis-classified 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. 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. In some cases, the correlations between race and non-race predictors are strong enough that the latter are considered proxies for the former, such as zip code (Pope & Sydnor, 2011). While we do not include zip code or other measures of socio-economic status in our lists of predictors, we do observe many important differences in predictors between the White and Black students in our sample. Appendix Table A4 shows that among the top 10 predictors of each model, most have statistically significant differences in the average predictor value between the White and Black students (9 out of 10 predictors for course competition model; 7 out of 10

[11] C-statistic 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). .

predictors for degree completion model). For example, Black students have 15-20 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 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 estimate a logistic regression for the White subgroup, and a separate logistic regression for the Black subgroup). We do so because, when we estimate the prediction models on the full sample of students, there is a single coefficient estimated for each predictor. Because a larger share of the sample are White students, then the coefficients should be more reflective of the determinants of success of White students, which may differ from those of Black students. When we estimate race-specific versions of the models, we allow the coefficient estimates to differ across racial lines, so that the determinants of success can be customized to the subgroup.[12]

Results

Demonstrating Algorithmic Bias

We begin by showing results from the base versions of the course completion and degree completion models. Figure 1 shows the distribution of predicted scores from the validation

[12] We explore whether the determinants of success differ between Black and White students by comparing the top 10 predictors from the race-specific models. Appendix Table A5 shows that two race-specific course completion models share 8 of the top 10 predictors; similarly, the two race-specific degree completion models share 9 of the top 10 predictors. However, the ordering of the predictors differs across race-specific models, as well as the magnitude of the coefficients of the top 10 predictors. For example, the student's age is the second most important predictor of course completion in the White-specific model, but the tenth most important predictor in the Black-specific model. For another example, the coefficient estimate on college-level credits earned is over twice as large in the White-specific model. These patterns indicate that there are subtle differences in the best way to predict success between the subgroups.

sample, separately for each subgroup. For both the course completion (Panel A) and degree completion (Panel B) outcomes, we observe higher predicted scores for White students compared to Black students -- this pattern is not surprising given that the actual success rates are also quite different across these groups. Specifically, the actual course completion rates for White and Black students are 79.8 percent and 69.4 percent, respectively; the actual degree completion rates for White and Black students are 39.2 percent and 22.4 percent, respectively.

Figure 2 presents plots that illustrate calibration bias; these plots compare the average success rates within a specific predicted score percentile range. Specifically, 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, 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. In Panel A, we observe clear evidence of calibration bias: within a given predicted score bin, White students have significantly higher actual success rates compared to Black students. This pattern occurs at nearly all points in the distribution of predicted scores. To provide a concrete example, suppose that students with predicted scores at or below the 30th percentile were identified as at-risk and targeted for intervention. In this case, the actual success rate among selected White students would be 53.8 percent, while the actual success rate among selected Black students would be 47.2 percent; this difference is statistically significant with a p-value of less than 0.01. This imbalance implies, that while the intervention would target a population of mostly at-risk students and many at-risk Black and White students would receive the intervention, there are some at-risk Black students who would not be targeted, even though they have the same (or higher) risk level as some White students who would be targeted. In other words, at the margin of identifying a student as at-risk or not, the course completion model is overestimating the success of Black students and underestimating the success of White students.

In Panel B of Figure 2, we observe a different pattern for the degree completion model. While the predicted scores appear to be better calibrated between the White and Black students for the lower and middle portions of the predicted score distribution, the low levels of the degree completion rates in the left half of the plot mask statistically significant differences between the White and Black success rates conditional on predicted score. Again supposing that students with predicted scores at or below the 30th percentile were targeted for intervention, the actual success rate among selected White students would be 4.1 percent, while the actual success rate among selected Black students would be 3.4 percent; this difference is statistically significant with a p-value of 0.029. 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. If the model was used in this way, then among the students with the highest predicted scores, Black students would have lower actual success rates than White students (i.e. at this margin, the model would again be underestimating Black student success and overestimating White student success). In this instance, the calibration bias would actually be beneficial to Black students, since they would be more likely to receive the positive treatment despite having lower success rates.

We further show how the amount of calibration bias differs across the predicted score distribution in Figure 3, with results from the course completion model in Panel A and the degree completion model in Panel B. This figure shows the percentage change in the number of targeted students after we simulate the removal of the calibration bias, for a variety of at-risk thresholds (i.e. the percentile of predicted score at or below which the students are considered at-risk and targeted for additional resources). In Panel A, we see that if all students in the bottom decile were targeted as at-risk, then the "unbiased simulation" would target 23.3 percent more Black students than the actual model; the unbiased simulation would target 14 percent fewer White students than the actual model. This translates to the share of all Black students who are targeted as at-risk being 14.1 percent in the actual model and 17.3 percent in the unbiased simulation. In other words, calibration bias would cause too few Black students to be targeted. Using this metric comparing the actual model and unbiased simulation, we see very similar levels of calibration bias at all points in the distribution of predicted scores for the course completion model, with the increase in at-risk Black students ranging from 23.3 percent to 25.4 percent.[13]

In Panel B, however, we observe a significantly different pattern. The amount of calibration bias is highest for the most stringent at-risk definitions (using the 10th, 20th, or 25th percentiles as at-risk definitions), but the amount of bias declines substantially such that at the 50th percentile, the simulated algorithm would target only 4.4 percent more Black students than the actual model. The amount of bias increases slightly to 6.7 percent at the 70th percentile, though this is still less than one-third of the amount of bias we observe at the 20th percentile.

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 1.5 percent lower, respectively). These racial differences in c-statistics are statistically significant at the $p < 0.01$ level.

[13] We do not display distributional points above the 70th percentile for two reasons. First, we do not view these as relevant points when considering the targeting at-risk students, because students above the 70th percentile would be very unlikely to be considered at-risk. Second, the calibration bias correction can no longer be implemented once the Black success rate within a percentile is equal to the actual Black success rate; for the course completion model, this occurs by the 75th percentile. For the degree completion model, this occurs by the 90th percentile.

Other data science papers investigating algorithmic bias in education contexts also consider other overall accuracy metrics, such as true and false positive rates and precision. In our context, the most relevant would be the true negative rate (among actually at-risk students, what share is correctly identified as at-risk?) and non-success precision (what share of students who the algorithm predicts to be at-risk are actually at risk?). Consistent with prior studies, we find higher true negative rates and non-success precision values for Black students than White students (Anderson, Boodhwani, & Baker, 2019, Jiang & Pardos, 2021, Jeong et al, 2021, Yu, Lee, & Kizilcec, 2021, Yu et al, 2020).[14] Based on these metrics 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 and non-success precision measure *overall* accuracy and efficiency of identifying at-risk students, however, the calibration bias measures the *marginal* accuracy and efficiency of identifying at-risk students. Based on our interpretation of which is most important to consider in the context of identifying 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.[15]

Making Models More Attuned to Race Impacts Algorithmic Bias

[14] Specifically, the true negative rate is 15.1 percent higher for Black students in the course completion model (relative to 52.4 percent for White students), and 8.7 percent higher for Black students in the degree completion model (relative to 84.1 percent for White students). The non-success precision is 12.7 percent higher for Black students in the course completion model (relative to 49.3 percent for White students), and 3.8 percent higher for Black students in the degree completion model (relative to 86.6 percent for White students).

[15] 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).

Next, we compare algorithmic bias from the base model (Figure 3, Table 1) to models including racial predictors and race-specific models. Figure 4 plots the amount of calibration bias (measured by the increase in Black students targeted as at-risk in the simulated model) across these three specifications. For the course completion model (Plot A), we see that making the models more attuned to race significantly reduces the amount of calibration bias compared to the base model; this is especially true for at-risk definitions focused on the "riskiest" students (e.g. bottom decile). This finding indicates that, for the course completion model, there is some useful information contained in race to allow for better calibration. 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. 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 4 for the course completion model does not come at the expense of overall model accuracy: Table 2, 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. The racial accuracy gaps are also very consistent across model specifications, approximately three percent for the course completion model and 1.5 percent for the degree completion model. Again, all racial differences in c-statistics are statistically significant at the $p < 0.01$ level.

However, including race information has a significantly different impact on calibration bias for the degree completion model. Figure 4, Plot B shows that the model including race predictors has slightly more calibration bias than the base model, and that the race-specific models have significantly higher calibration bias than the base model (with the exception of at-risk thresholds above the 50th percentile). This finding is noteworthy for two reasons. First, it 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 opposite effects on the level of calibration bias demonstrates just how idiosyncratic algorithmic bias can be across models. 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. We offer two related explanations as to why we observe these divergent patterns regarding making models more attuned to race. First, in the case of course completion, there are clearly unobservable factors correlated with both race and the outcome; in the model including race predictors, the coefficient estimate for the Black indicator is -0.3 ($p < 0.01$). Therefore, including race significantly improves the model's ability to provide better *marginal* estimation across racial groups, i.e. by changing the rank ordering of Black and White students within a similar predicted score range (even though the overall model performance measured by the c-statistic remains the same). However, in the case of the degree completion model, this is not the case: the coefficient estimate for the Black indicator is much smaller (-0.03) and not statistically significant ($p = 0.35$).

Second, estimating race-specific models can result in significant changes in the distribution of predicted scores with respect to race. In a race-specific model, exactly 10 percent of the Black sample will be in the bottom decile of predicted scores. However, in the full model, this won’t necessarily be true -- particularly in the degree completion model, significantly more than 10 percent of black students are in the bottom decile of predicted scores (see Figure 1). This redistribution of students in the race-specific models has important implications for calibration bias at particular points in the distribution of predicted scores. This is less relevant to the course completion model, because the racial differences are not as stark in the lower portions of predicted score distributions. Therefore, our results imply that there is minimal benefit to including race in the degree completion model, and the differences in predicted score distribution that results from estimating race-specific degree completion models leads to significantly more calibration bias.

Taken as a whole, our results make it clear that algorithmic bias should be considered on a case-by-base basis.

Exploring Why Algorithmic Bias Exists

One reason why algorithmic bias may arise is referred to as the “selective labels problem,” where the outcome that we use in the model is confounded by human decision-making and therefore does not translate to the same true outcome across different subgroups. Obermeyer et al. (2019) provide one clear example of the selective labels problem: the algorithm they analyze uses health care expenditures as a proxy for actual health, but they show that expenditure behavior differs in meaningful ways across races. 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. 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.[16] In this sense, the selective labels problem in our context translates to a differential sorting problem.

To explore the potential bias caused by differential sorting of students, we estimate course-specific and degree-specific models for the five most popular courses and degree programs offered by VCCS. If there was significantly less algorithmic bias in models where we are focusing on subgroups of students who self-selected into the same course or program,then that would point to the selective labels problem. Figure 5 compares the amount of calibration bias in the full models versus course- or program-specific models.[17] In Figure 5, Panel A, when we look within-course, we observe relatively similar levels of calibration bias between the full base model and the course-specific models. In most cases, the amount of calibration bias is slightly higher for the course-specific models.[18] Panel B of Figure 5 compares the calibration

[16] 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 A6 for these results.

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

[18] It is also interesting to note that there is significantly less calibration bias predicting success in ENG112 compared to the other four courses, regardless of whether we use full or the course-specific model. While our inference is limited from considering only five courses, because ENG112 is the only course that students typically cannot take in their first term at VCCS (because ENG111 is a prerequisite), this pattern suggests that calibration bias is lower when we predict success for returning students compared with new students; we find additional support for this hypothesis in below.

bias for the full model and the program-specific models. While the levels of calibration bias are more variable across programs compared to courses, we do not observe any meaningful pattern of bias reduction using the program-specific models instead of the full model.[19] The lack of bias reduction in Figure 5 suggests that the calibration bias is not due to differential sorting of students across courses or programs.[20] [21]

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 when estimating the coefficients of the model (i.e. which characteristics are most important in predicting success). 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 Figure 6. For the course completion model (Plot A), we find very similar levels of calibration bias between the base model and downsampled model. For the degree completion model (Plot B), we find that the downsampled model exhibits more calibration bias than the base model for "riskiest" populations (e.g. at-risk defined using the bottom decile), but that there is

[19] The one exception is the AS in Science program where the program-specific model exhibits very little to no calibration bias across the distribution of possible thresholds.

[20] 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.

[21] We find a similar pattern of results when we estimate college-specific models for the five largest VCCS colleges, as shown in Appendix Figure A2: 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) that 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.

little difference in the calibration bias at higher thresholds. These results suggest that data underrepresentation is not the source of calibration bias in our models.[22]

While the typical form of data underrepresentation described above does not appear to be driving the algorithmic bias in our models, 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. 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. For students in their first-term, the 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 7.[23] 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 and orange bars to the gray and yellow bars). We also observe some reduction in bias when we estimate subgroup-specific models (i.e. comparing the blue bar to orange bar, and separately the gray bar to yellow bar), although these reductions are mostly modest. This pattern of results supports the hypothesis that Black students having shorter enrollment histories is a contributing factor to the calibration bias of our models, and suggests that the additional predictors available for returning students, such as cumulative

[22] 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.
[23] Again, the comparison “full” models in Figure 7 are the models trained on the full training set but applied to the subgroup-specific validation set.

GPA and credits completed from prior terms, partially mitigates the calibration bias. Taken together, these results suggest: that shorter-enrollment histories of Black students accounts for some but not all of the calibration bias.

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 5, 6, 7, and A2 are in Appendix Tables A7 through A10. 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 three additional goodness of fit metrics from the training sample: c-statistic, Efron's R-squared (Efron, 1978), and McFadden's Adjusted R-squared (McFadden, 1974). Under certain assumptions, the c-statistic from the training set represents the limit of the prediction accuracy the model can attain.[24] We compute these statistics for the race-specific models, so that the determinants of success are allowed to differ across groups. Table 3 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.

[24] This is because the logistic regression model fitting procedure is an optimization routine which identifies the set of coefficient values that maximize the prediction accuracy of observations in the training set. The relevant assumptions are no overfitting and perfect generalization (the probability distribution of the unseen data is identical to the observed training set).

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 outcomes than White students, which would lead to some at-risk Black students being less likely to receive additional resources than White students who are comparatively at lower risk of dropping out; 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 underestimating Black student performance at the threshold of being classified 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. 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, 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 Black students more at-risk of dropping out receiving less support that White students at comparatively lower-risk. 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

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

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-analyt ics-higher-education/#:~:text=In%20a%20new%20paper%2C%20The,well%20in%2C%20and%20provide%20digital

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.

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/

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

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.

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

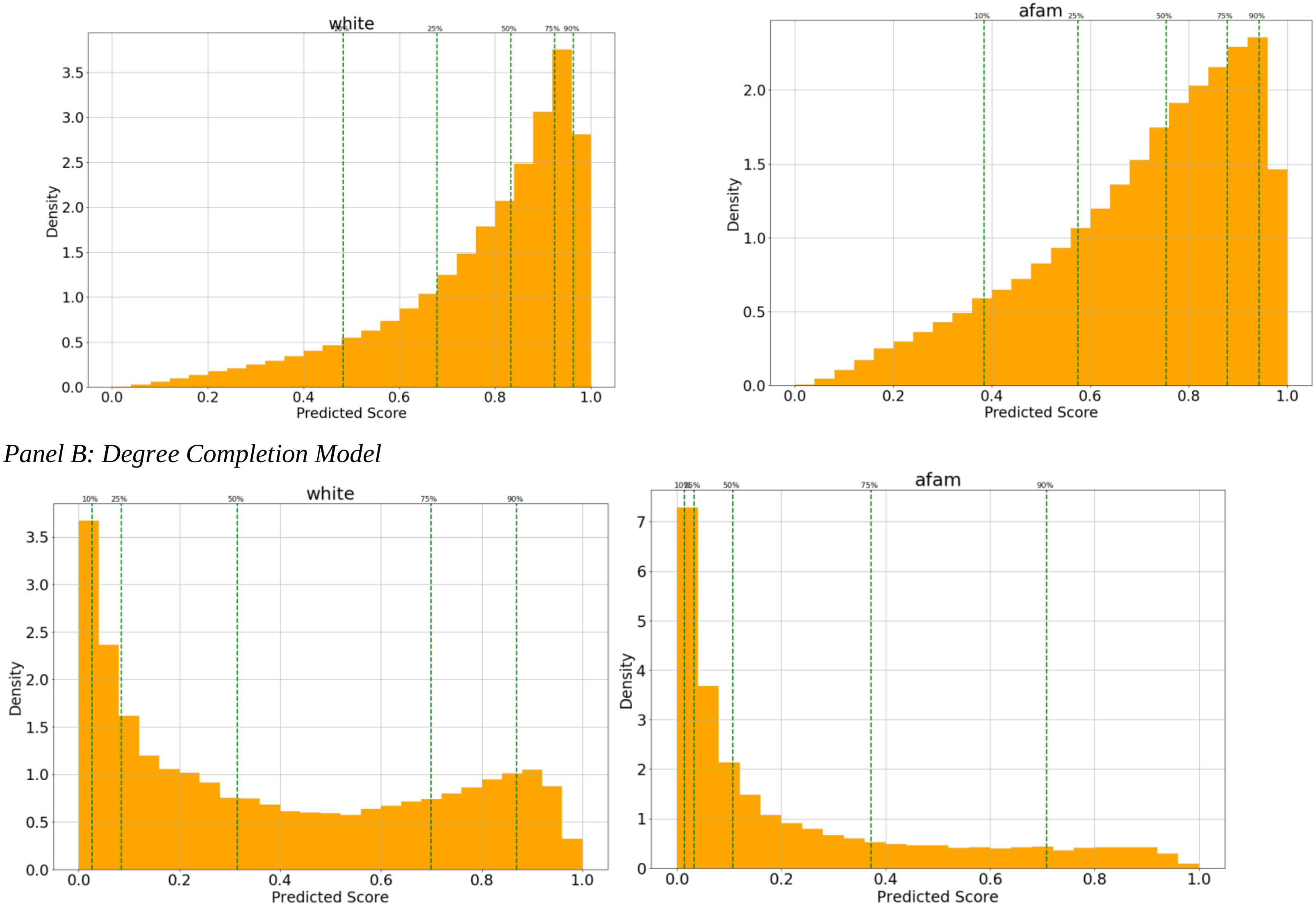
U.S. Department of Education (2022). Department of Education Releases Equity Action Plan as Part of Biden-Harris Administration's Efforts to Advance Racial Equity and Support Underserved Communities. Press release, April 14, 2022. https://www.ed.gov/news/press-releases/department-education-releases-equity-action-plan-part-biden-harris-administrations-efforts-adv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

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

Figure 1: Distribution of predicted scores in base model, by race

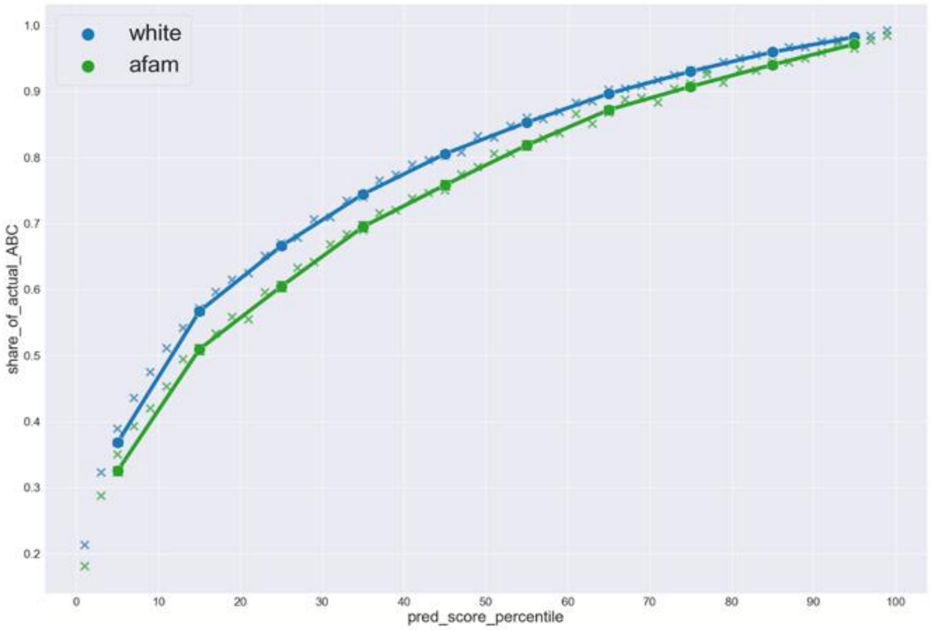
Panel A: Course completion model



Panel B: Degree Completion Model

Figure 2: Calibration of base models, by race

Panel A: Course completion model



Panel B: Degree Completion Model

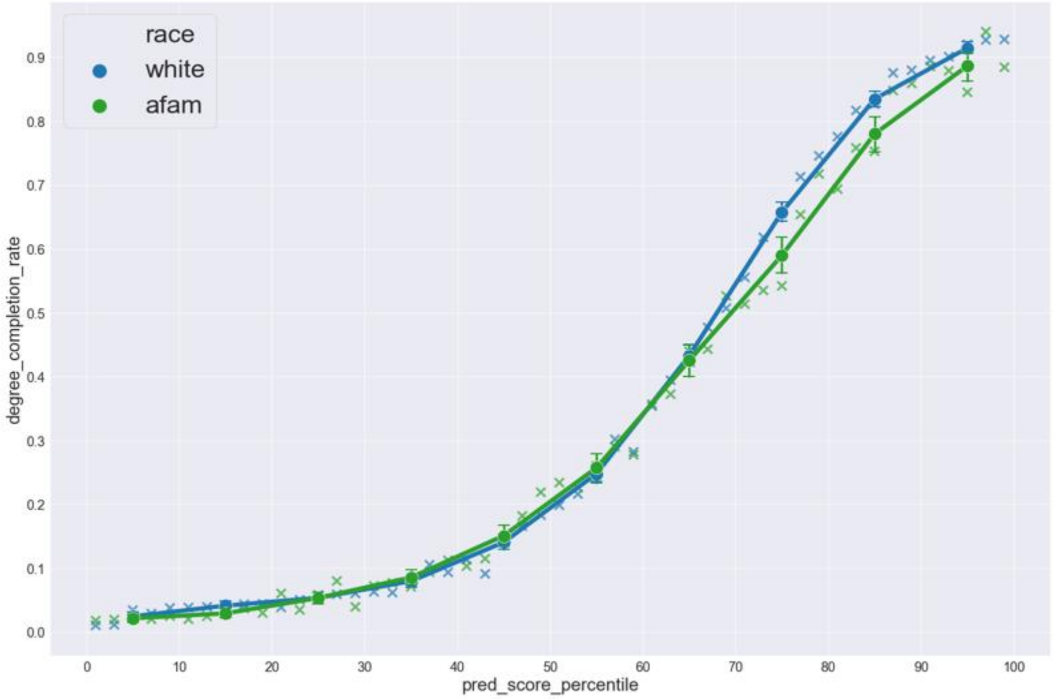
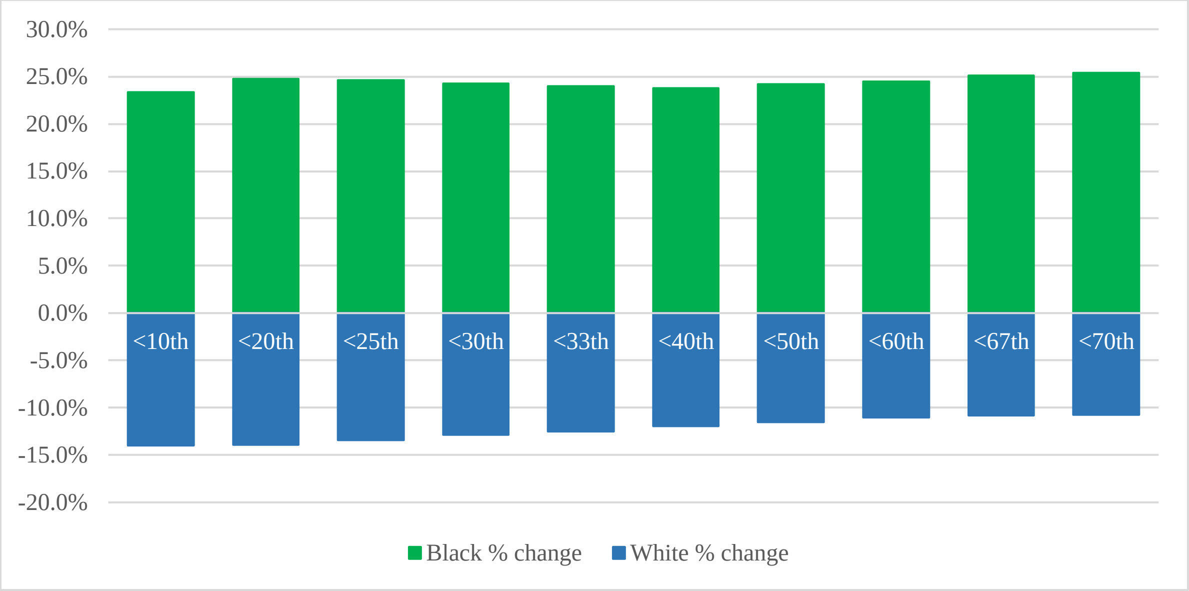


Figure 3: Quantifying calibration bias (percentage change in the number of students targeted as at-risk between true model and unbiased simulation)

Panel A: Course completion model



Panel B: Degree Completion Model

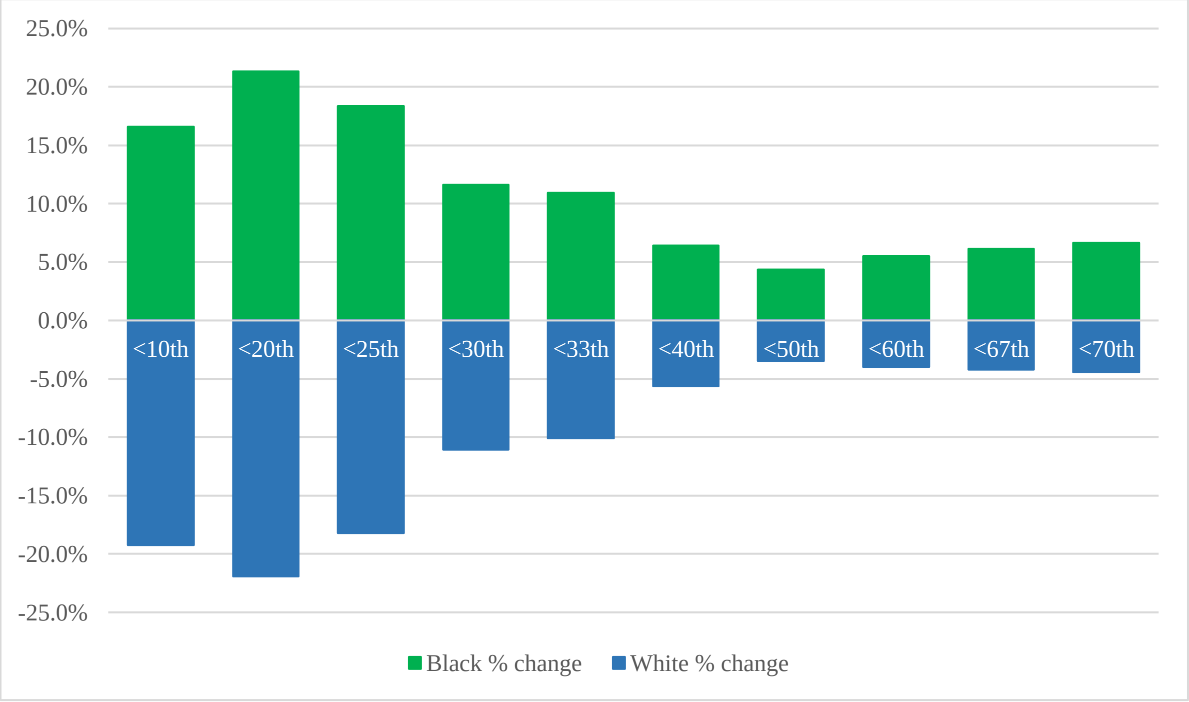
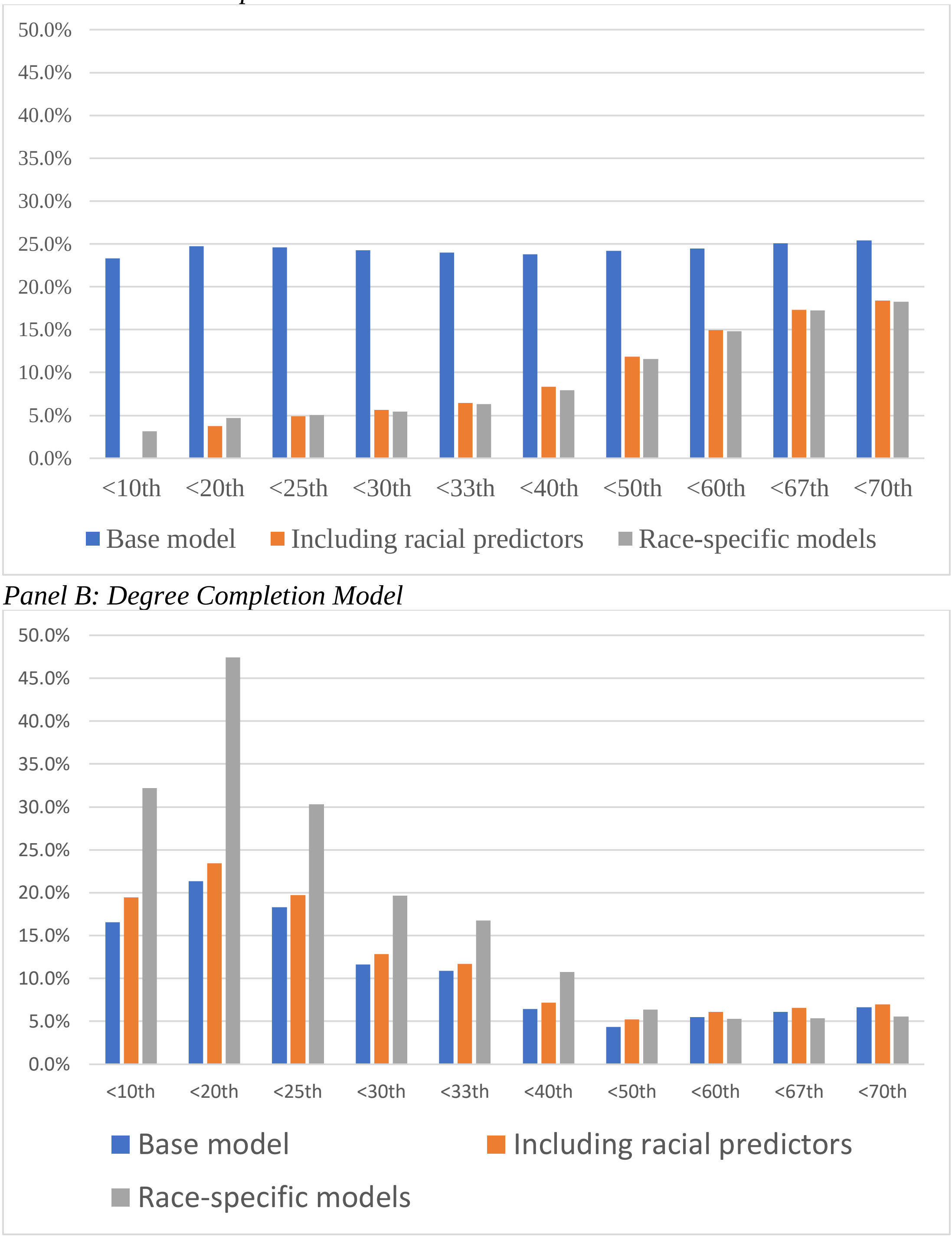


Figure 4: Differences in calibration bias between base model and models that incorporate race information

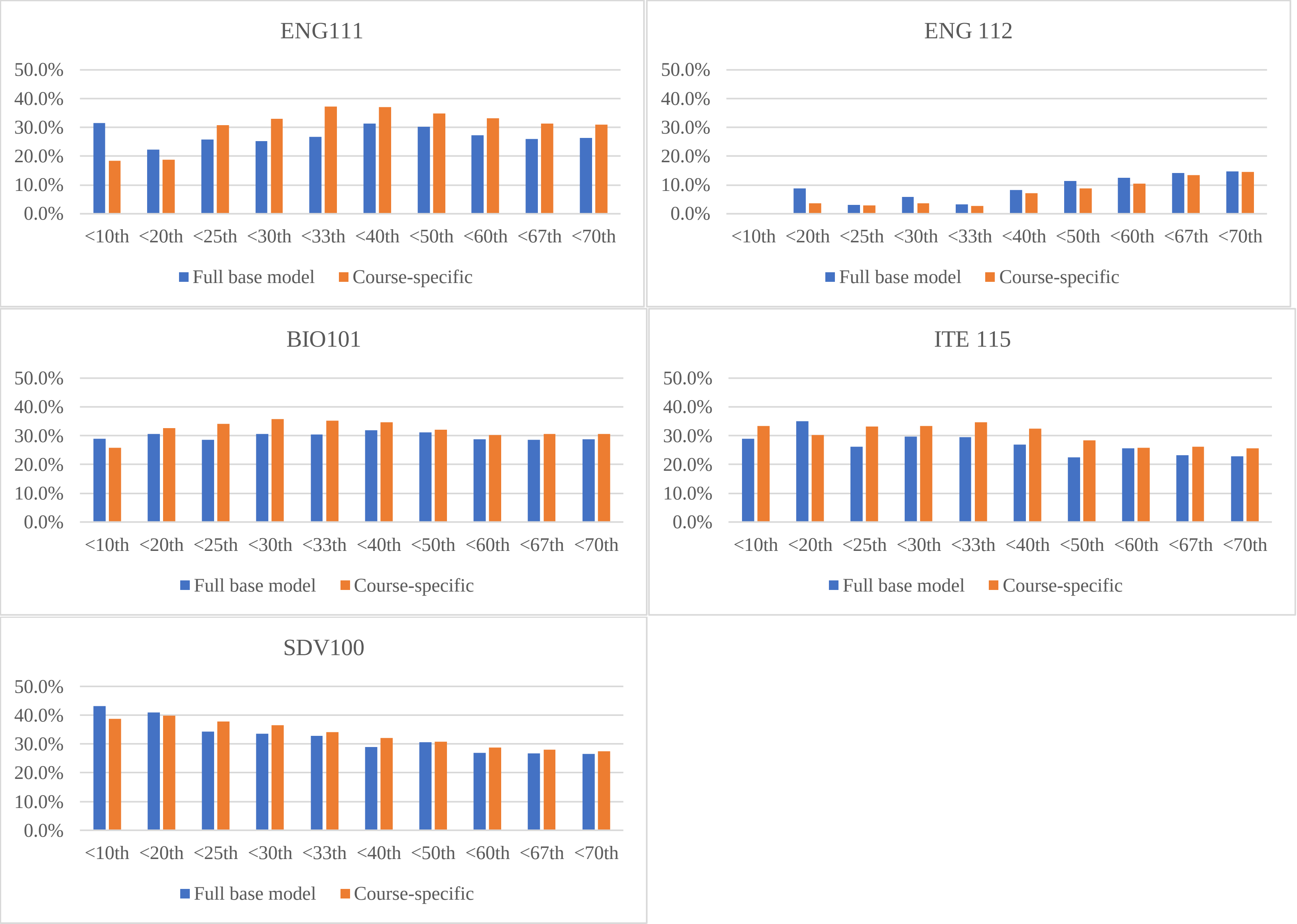


Panel A: Course completion model

Panel B: Degree Completion Model

Figure 5: Comparing calibration bias of course-specific and program-specific models

Panel A: Course completion model



Panel B: Degree Completion Model

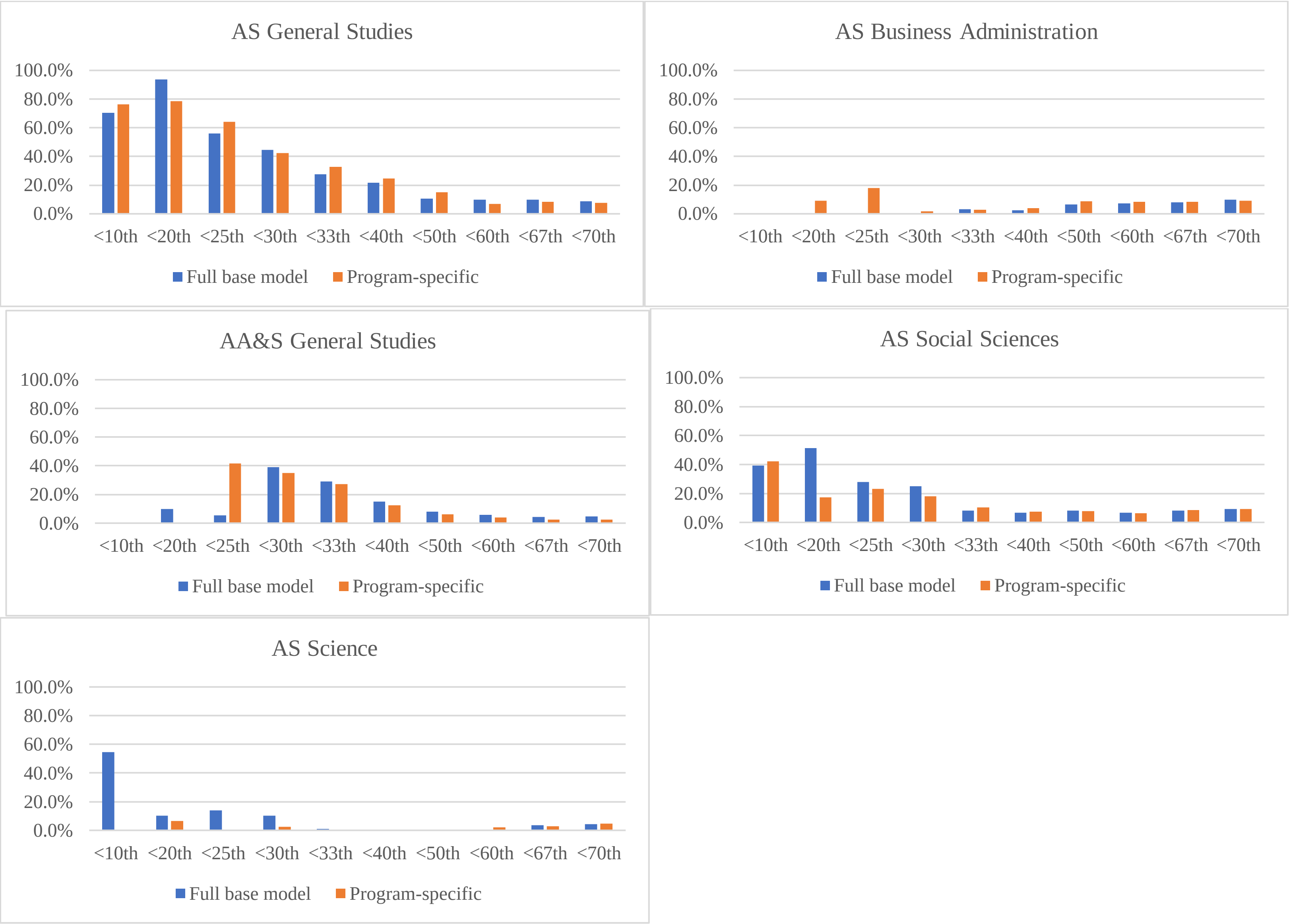
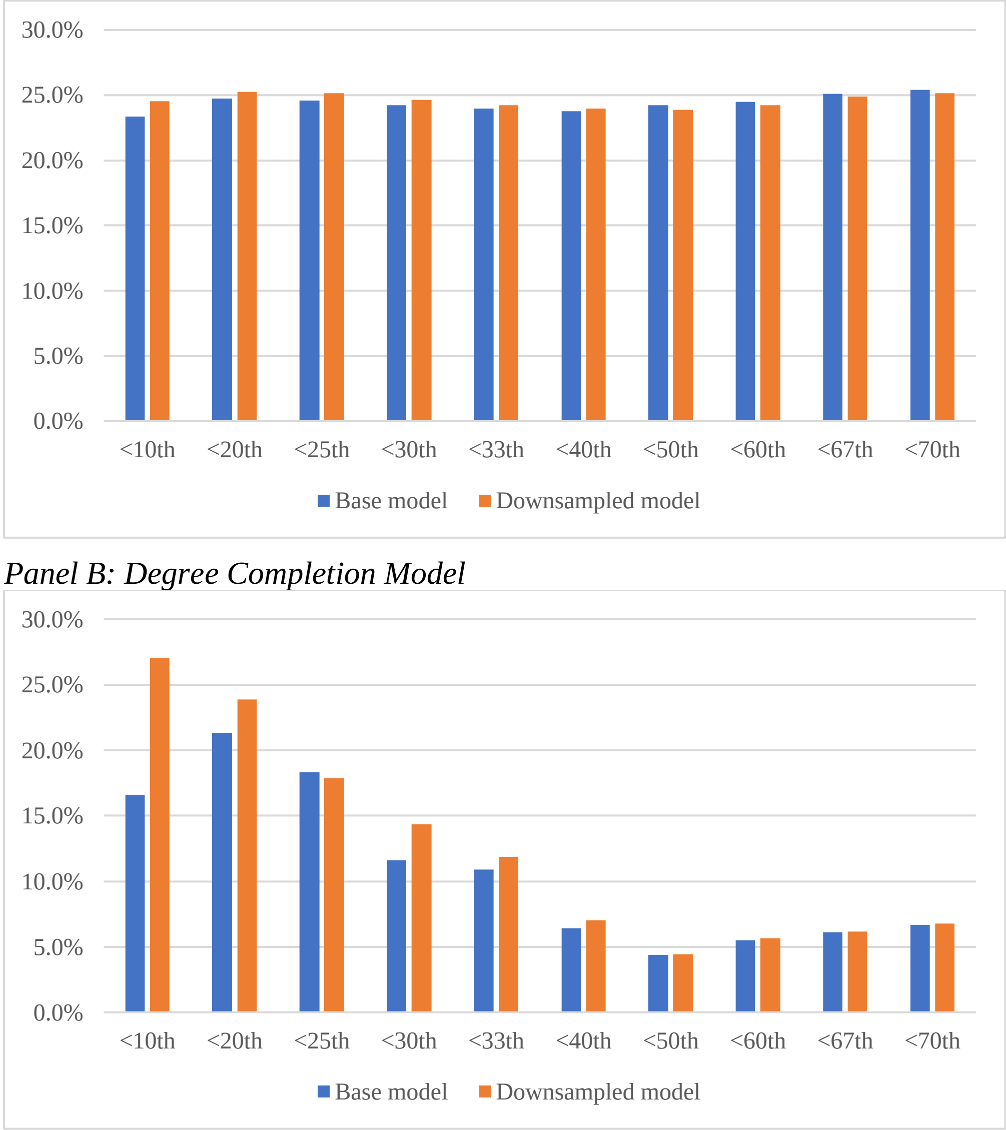


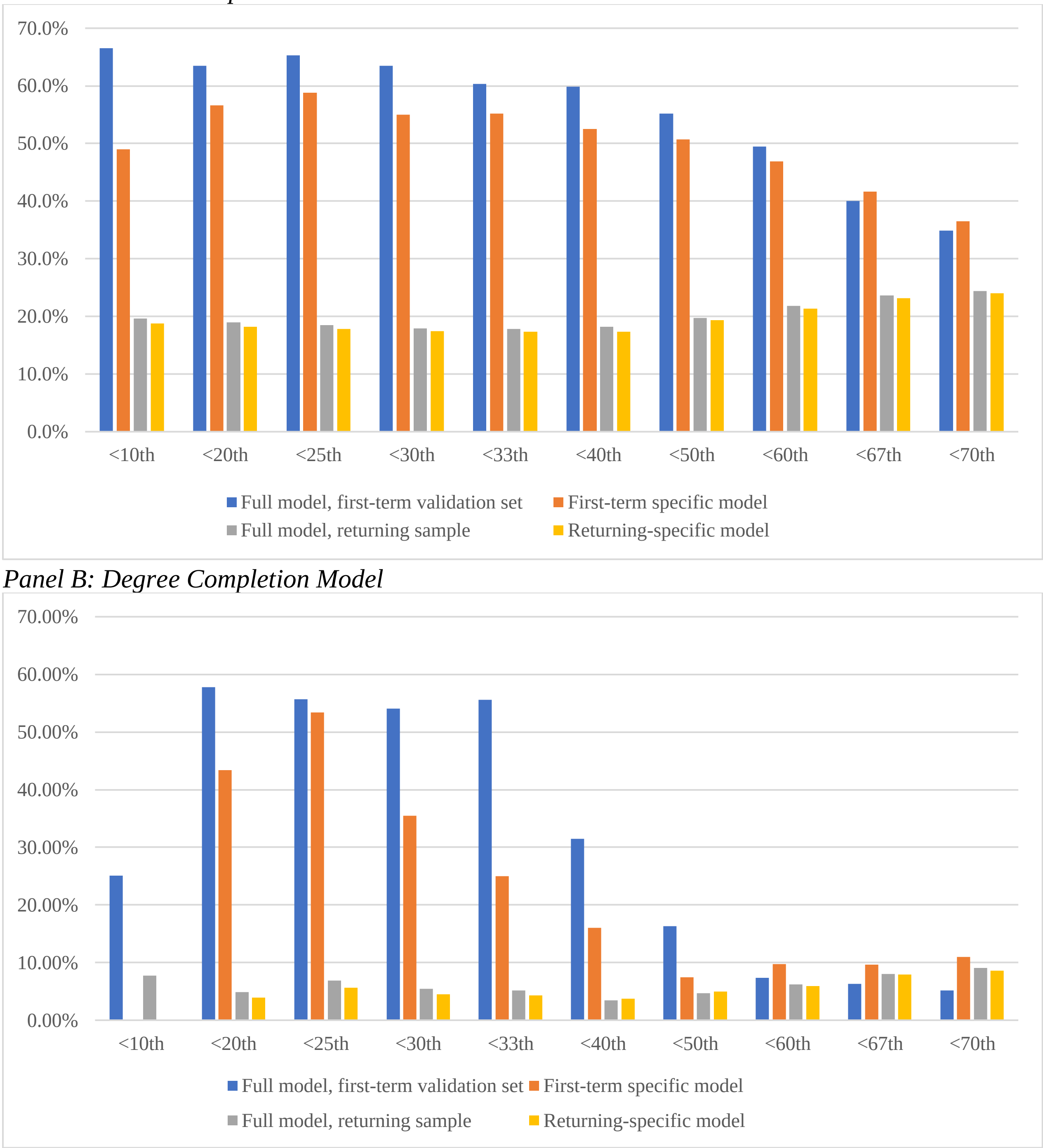
Figure 6: Calibration bias in downsampled model



Panel A: Course completion model

Panel B: Degree Completion Model

Figure 7: Calibration bias in first-term specific and returning specific models



Panel A: Course completion model

Panel B: Degree Completion Model

Table 1: Accuracy of base models

Panel A: course completion model			
	White	Black	% diff
C-statistic	0.8007	0.777	-2.96%
	(0.0007)	(0.0012)	
True Negative Rate	0.5236	0.6026	15.09%
	(0.0018)	(0.0024)	
Precision non-success	0.4933	0.5559	12.69%
	(0.0017)	(0.0023)	
Panel B: degree completion model			
	White	Black	% diff
C-statistic	0.8933	0.8802	-1.47%
	(0.0020)	(0.0037)	
True Negative Rate	0.8407	0.9136	8.67%
	(0.0026)	(0.0025)	
Precision non-success	0.8658	0.8986	3.79%
	(0.0025)	(0.0026)	

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

Table 2: 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
0.8007	0.777	-2.96%	0.8011	0.7768	-3.03%	0.8019	0.7782	-2.96%
(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
0.8933	0.8802	-1.47%	0.8933	0.8802	-1.47%	0.8936	0.8799	-1.53%
(0.0020)	(0.0037)		(0.0020)	(0.0037)		(0.0020)	(0.0037)	

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

Table 3: Goodness of fit metrics

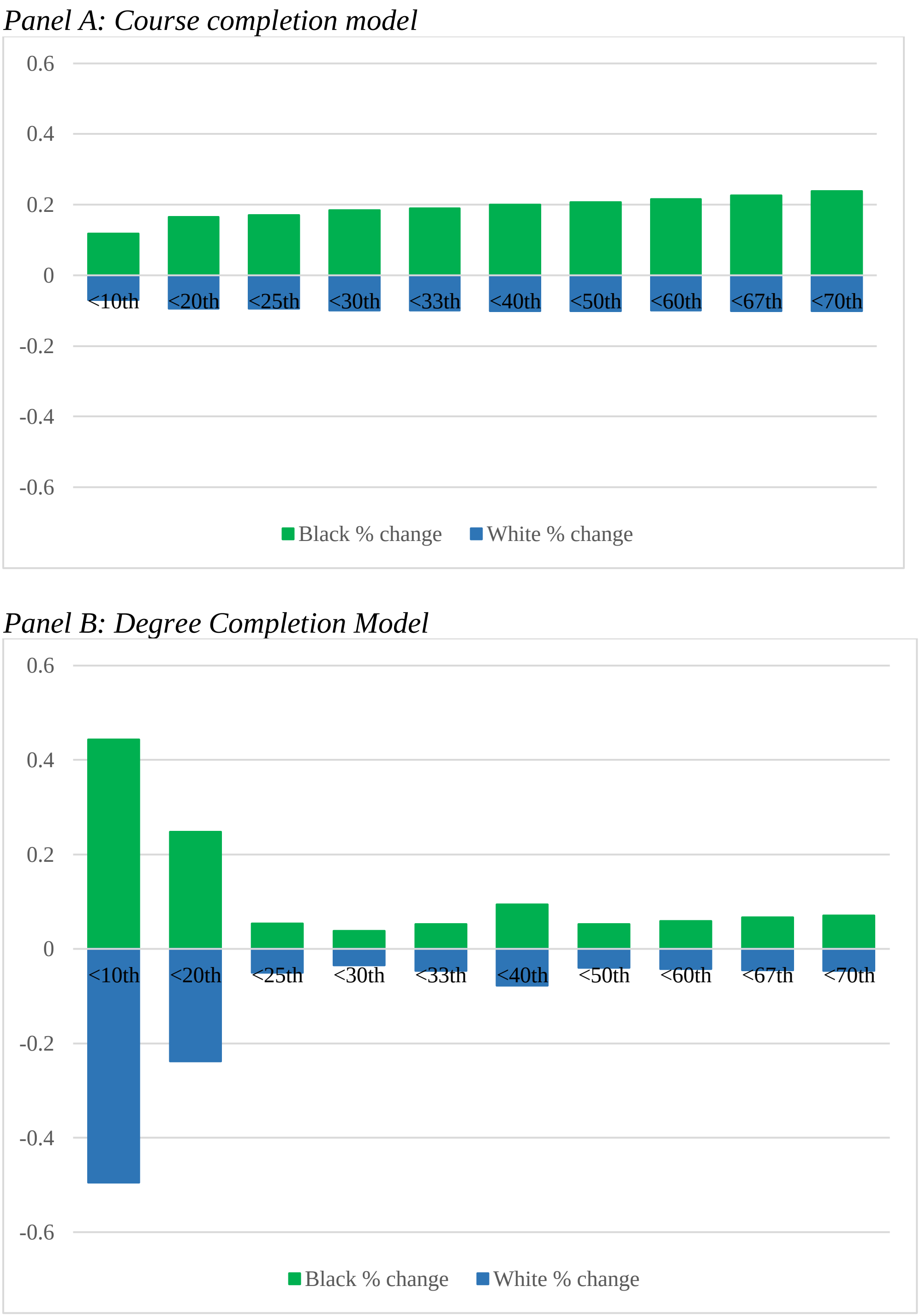
Panel A: Course completion model

	White	Black
C-statistic (training set)	0.7976	0.78
Efron's R-squared	0.2092	0.2161
McFadden's Adjusted R-squared	0.4375	0.2308

Panel B: Degree completion model

	White	Black
C-statistic (training set)	0.8846	0.8599
Efron's R-squared	0.4458	0.3623
McFadden's Adjusted R-squared	0.6037	0.4235

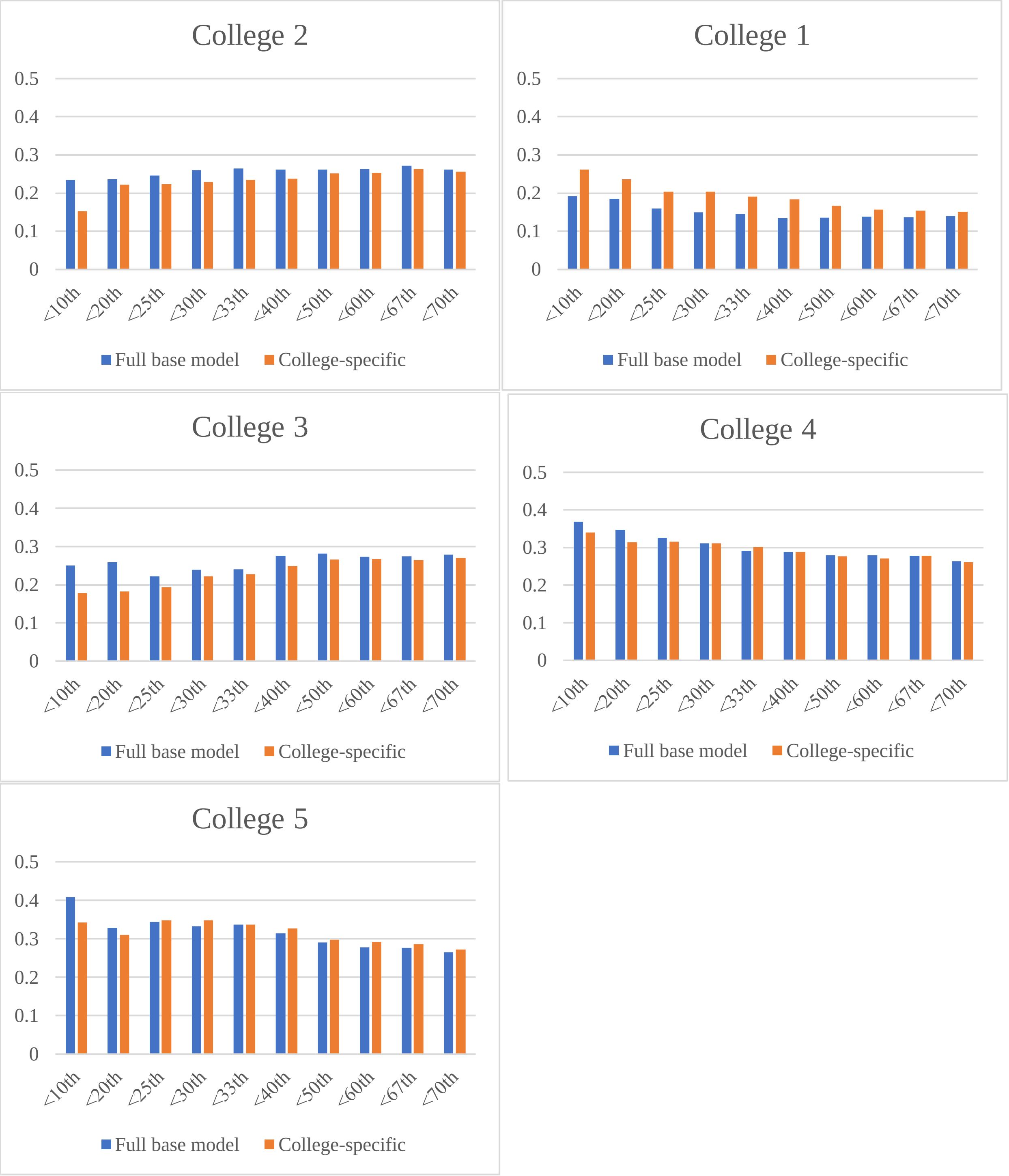
Appendix Figure A1: Calibration bias in base Random Forest models


Panel A: Course completion model

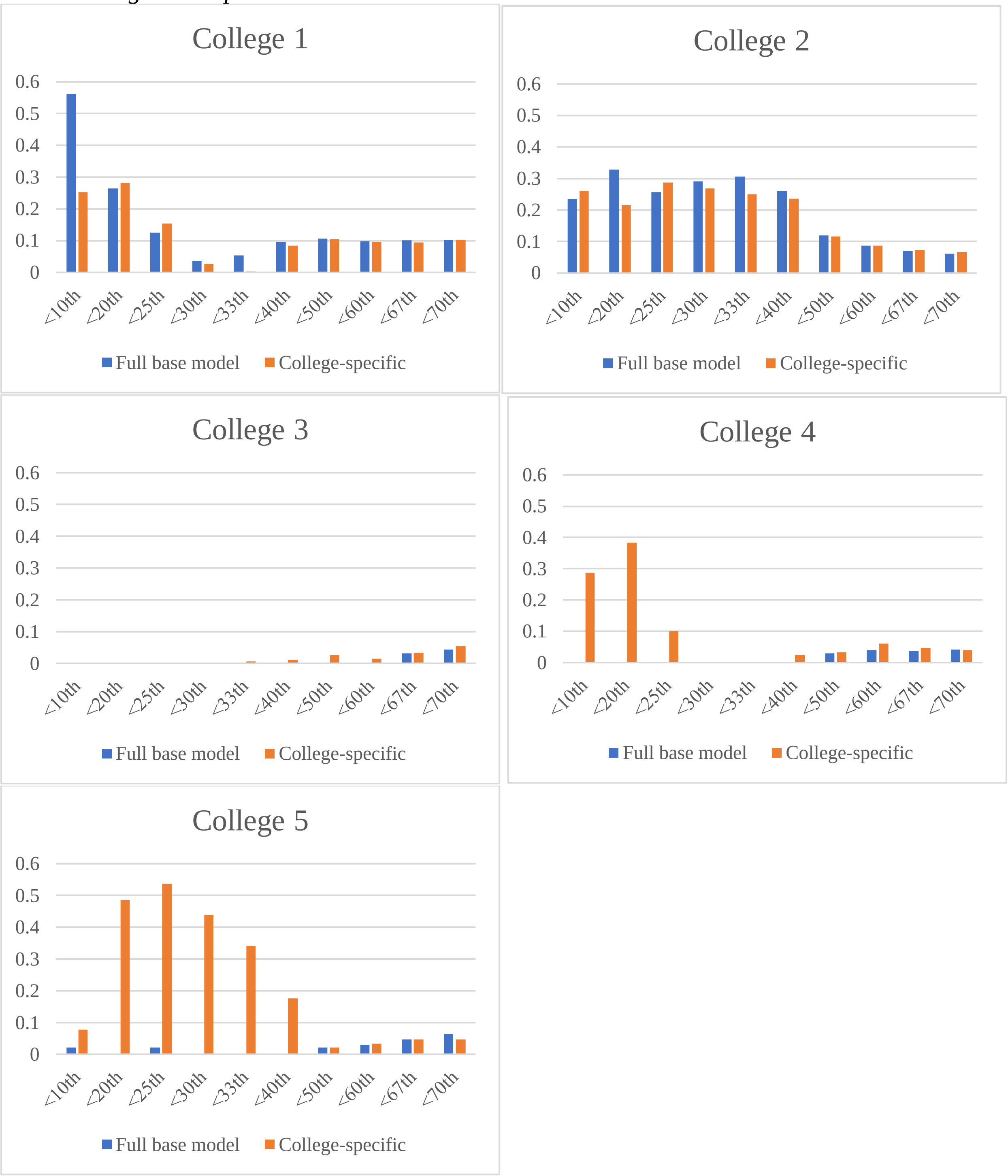
Panel B: Degree Completion Model

Appendix Figure A1: Calibration bias in college-specific models

Panel A: Course Completion Model



Panel B: Degree completion model



Appendix Table A1: C-statistic of base RF models

Panel A: course completion model

White	Black	% diff
0.8143	0.7871	-3.34%
(0.0007)	(0.0012)	

Panel B: degree completion model

White	Black	% diff
0.902	0.8906	-1.26%
(0.0019)	(0.0036)	

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

Appendix Table A2: 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 characteristics, course-specific	X
Student took the target course's prerequisites (if applicable)	Student's academic characteristics, course-specific	X
Student has previously taken the target course	Student's academic characteristics, course-specific	
Student's average prior grade in the target course (if repeating the course)	Student's academic characteristics, course-specific	
Has taken prior Arts courses (target course = X subject)	Student's academic characteristics, course-specific	
Average grade in prior Arts courses (target course = X subject)	Student's academic characteristics, course-specific	
Has taken prior Business/Finance courses (target course = X subject)	Student's academic characteristics, course-specific	
Average grade in prior Business/Finance courses (target course = X subject)	Student's academic characteristics, course-specific	
Has taken prior Engineering courses (target course = X subject)	Student's academic characteristics, course-specific	
Average grade in prior Engineering courses (target course = X subject)	Student's academic characteristics, course-specific	
Has taken prior Foreign Languages courses (target course = X subject)	Student's academic characteristics, course-specific	
Average grade in prior Foreign Languages courses (target course = X subject)	Student's academic characteristics, course-specific	
Has taken prior Humanities courses (target course = X subject)	Student's academic characteristics, course-specific	
Average grade in prior Humanities courses (target course = X subject)	Student's academic characteristics, course-specific	
Has taken prior Medical Sciences courses (target course = X subject)	Student's academic characteristics, course-specific	
Average grade in prior Medical Sciences courses (target course = X subject)	Student's academic characteristics, course-specific	
Has taken prior Mathematics courses (target course = X subject)	Student's academic characteristics, course-specific	
Average grade in prior Mathematics courses (target course = X subject)	Student's academic characteristics, course-specific	
Has taken prior Applied Technologies courses (target course = X subject)	Student's academic characteristics, course-specific	
Average grade in prior Applied Technologies courses (target course = X subject)	Student's academic characteristics, course-specific	
Has taken prior Natural Sciences courses (target course = X subject)	Student's academic characteristics, course-specific	
Average grade in prior Natural Sciences courses (target course = X subject)	Student's academic characteristics, course-specific	
Has taken prior Social Sciences courses (target course = X subject)	Student's academic characteristics, course-specific	

Average grade in prior Social Sciences courses (target course = X subject)	Student's academic characteristics, course-specific	
Age at time of target course enrollment	Student Demographics	X
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 A3: Full list of predictors, degree completion model

Predictor description	Category	Available for 1st term
Age at initial enrollment	Demographics	X
Gender	Demographics	X
Race/Ethnicity (four binary indicators for White, Black, Hispanic, other)	Demographics	X
Parents' highest education level (categorized)	Demographics	X
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 A4: Subgroup difference in top 10 predictors

Panel A: Course completion

	White mean	Black - White diff	p-value
	(1)	(2)	(3)
# credits attempted in the target term	10.6154	-0.629	**0
Average grade assigned by the instructor in the target course	2.5203	-0.0029	0.3083
Instructor has taught the target course in the past	0.8868	0.0132	**0
Slope of credits attempted in prior terms	0.2837	-0.1921	**0
Credits attempted in last term (prior to target term)	8.1232	-0.5114	**0
Cumulative GPA	2.5381	-0.3782	**0
% prior attempted credits "Withdrawn"	0.0488	0.0094	**0
Average historical grade in the target course	2.8542	-0.0489	**0
% prior attempted credits "Incomplete"	0.005	0.0071	**0
Student age	23.8405	2.1047	**0

Panel B: Degree completion

	White mean	Black - White diff	p-value
	(1)	(2)	(3)
College-level credit hours accumulated	0.0756	-0.0459	**0
Credits attempted, second Summer term	0.7846	-0.0877	**0
Proportion of credits withdrawn, sixth Fall term	0.0016	-0.0003	0.3618
Cumulative GPA	2.6014	-0.5534	**0
Total enrollment intensity in non-VCCS institutions, first Fall term	0.0162	-0.002	0.0673
Credits attempted, first Fall term	8.9384	-1.1436	**0
Share of total credits withdrawn	0.1089	0.0457	**0
Number non-VCCS enrolled terms prior to initial VCCS enrollment	0.5002	0.0967	**0
Credits attempted, second Fall term	4.2929	-1.3087	**0
Proportion of credits withdrawn, sixth Summer term	0.0002	0.0001	0.4108

Appendix Table A5: Top 10 predictors of race-specific models

Panel A: Course completion model

White			Black		
Predictor	Coef	95% CI	Predictor	Coef	95% CI
# credits attempted in the target term	0.18	(0.1789, 0.1811)	# credits attempted in the target term	0.1571	(0.1554, 0.1587)
Student Age	0.0158	(0.0152, 0.0163)	Average grade assigned by the instructor in the target course	0.9058	(0.8767, 0.9349)
Average grade assigned by the instructor in the target course	0.8012	(0.7827, 0.8198)	Instructor has taught the target course in the past	-2.4877	(-2.5705, -2.4049)
Instructor has taught the target course in the past	-2.1543	(-2.2076, -2.101)	Credits attempted in last term (prior to target term)	-0.034	(-0.0362, -0.0317)
Share of credits earned	0.0016	(0.0011, 0.002)	Slope of credits attempted in prior terms	0.0195	(0.0168, 0.0223)
Slope of credits attempted in prior terms	0.0146	(0.0127, 0.0164)	Cumulative GPA	0.1281	(0.1136, 0.1427)
Credits attempted in last term (prior to target term)	-0.0425	(-0.0441, -0.041)	% prior attempted credits "Withdrawn"	-1.3454	(-1.3919, -1.2989)
Term GPA in last term (prior to target term)	0.3533	(0.3462, 0.3604)	Average historical grade in the target course	0.3521	(0.3218, 0.3824)
% prior attempted credits "Withdrawn"	-1.2219	(-1.2542, -1.1895)	Enrollment in target course section	0.0068	(0.0062, 0.0074)
Average historical grade in the target course	0.3557	(0.3363, 0.3751)	Student Age	0.0118	(0.0112, 0.0124)

Panel B: Degree completion model

White			Black		
Predictor	Coef	95% CI	Predictor	Coef	95% CI
College-level credit hours accumulated	0.7165	(0.6433, 0.7898)	Cumulative GPA	0.2314	(0.1913, 0.2715)
Proportion of credits withdrawn, sixth Fall term	-2.0439	(-3.1037, -0.9841)	Credits attempted, second Fall term	0.0505	(0.0405, 0.0606)

Cumulative GPA	0.3635	(0.3326, 0.3943)	Number non-VCCS enrolled terms prior to initial VCCS enrollment	0.1985	(0.1729, 0.2241)
Total enrollment intensity in non-VCCS institutions, first Fall term	0.3591	(0.0815, 0.6368)	Total enrollment intensity in non-VCCS institutions, first Fall term	0.7829	(0.3862, 1.1797)
Credits attempted, first Fall term	0.0776	(0.0733, 0.0819)	Credits attempted, first Fall term	0.0645	(0.0578, 0.0712)
Share of total credits withdrawn	-1.3585	(-1.5141, -1.203)	Share of total credits withdrawn	-0.7456	(-0.9588, -0.5324)
Number non-VCCS enrolled terms prior to initial VCCS enrollment	0.2747	(0.2558, 0.2935)	Trend of term enrollment intensity (term credits attempted)	0.0263	(0.0187, 0.034)
Credits attempted, second Fall term	0.0587	(0.0526, 0.0647)	College-level credit hours accumulated	0.3287	(0.1927, 0.4647)
Credits attempted, first Summer term	0.0395	(0.0333, 0.0456)	Credits attempted, first Summer term	0.0295	(0.0201, 0.0388)
Trend of term enrollment intensity (term credits attempted)	0.0274	(0.0225, 0.0322)	Standard deviation of term proportion of credits earned	-0.7312	(-0.913, -0.5495)

Appendix Table A6: 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

Table 4: Program and course specific models

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
0.7271	0.7235	-0.50%	0.7514	0.729	-2.98%	0.7953	0.7559	-4.95%	0.7545	0.7457	-1.17%	0.7875	0.77	-2.22%
(0.0041)	(0.0061)		(0.0046)	(0.0067)		(0.0042)	(0.0070)		(0.0053)	(0.0078)		(0.0051)	(0.0087)	

Panel B: Degree Completion model

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
0.8803	0.8892	1.01%	0.8843	0.8869	0.29%	0.9224	0.9182	-0.46%	0.8932	0.9047	1.29%	0.8843	0.8671	-1.95%
(0.0068)	(0.0111)		(0.0079)	(0.0120)		(0.0055)	(0.0146)		(0.0073)	(0.0107)		(0.0080)	(0.0139)	

Appendix Table A8: C-statistics of downsampled model

Panel A: course completion model

White	Black	% diff
0.7988	0.7779	-2.62%
(0.0008)	(0.0012)	

Panel B: degree completion model

White	Black	% diff
0.8927	0.8805	-1.37%
(0.0020)	(0.0037)	

Appendix Table A9: Accuracy of first-term and returning subgroup-specific models

Panel A: Course completion model

First-term specific model			Returning-specific model		
White	Black	%diff	White	Black	%diff
0.7149	0.6991	-2.21%	0.8151	0.7916	-2.88%
(0.0023)	(0.0034)		(0.0008)	(0.0013)	

Panel B: Degree Completion model

First-term specific model			Returning-specific model		
White	Black	%diff	White	Black	%diff
0.8522	0.8339	-2.15%	0.8898	0.8814	-0.94%
(0.0065)	(0.0109)		(0.0022)	(0.0041)	

Appendix Table A10: 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
0.7898	0.7762	-1.72%	0.8033	0.7738	-3.67%	0.8007	0.7824	-2.29%	0.8053	0.7934	-1.48%	0.7923	0.7893	-0.38%
(0.0016)	(0.0023)		(0.0020)	(0.0025)		(0.0033)	(0.0040)		(0.0034)	(0.0042)		(0.0031)	(0.0045)	

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
0.8855	0.8771	-0.95%	0.8562	0.8816	2.97%	0.8905	0.8892	-0.15%	0.8652	0.829	-4.18%	0.8974	0.888	-1.05%
(0.0043)	(0.0074)		(0.0060)	(0.0071)		(0.0087)	(0.0122)		(0.0096)	(0.0146)		(0.0094)	(0.0167)	