How Effective are Colleges in Educating a Diverse Student Body? Evidence from West Point*

Dario Cestau

IE Business School

Dennis Epple

Carnegie Mellon University and NBER

Richard Romano

University of Florida

Holger Sieg

University of Pennsylvania and NBER

Carl Wojtaszek
U.S. Military Academy at West Point

*We would like to thank an associate editor of the journal, Flavio Cunha, two anonymous referees, Donna Gilleskie, Christopher Jepsen, Kala Krishna, David Lyle, Paul Oyer, Andy Postlewaite, Petra Todd, and Lesley Turner as well as seminar participants at numerous seminars and conferences for comments and discussions. Financial support for this project was provided by the NSF grant SES-1658746. All opinions expressed in this manuscript are those of the authors and do not represent the opinions of the United States Military Academy (USMA), United States Cadet Command, the United States Army, or the Department of Defense. The Army's Office of Economic and Manpower Analysis provided administrative data for this project to Carl Wojtaszek as part of a restricted use agreement that specifies that data can only be stored, accessed, and analyzed within USMA's information system. Any parties interested in accessing this data must make a direct application to USMA.

Abstract:

We develop an assessment strategy that compares outcomes of majority and minority college students with comparable entering qualifications using matching algorithms. We implement this strategy using data for West Point. We find that minority students at West Point have similar graduation rates as their white counterparts. Moreover, there are no differences in early career outcomes between majority and minority graduates. There is, however, an achievement gap between black and white students but not between Hispanic and white students. Finally, a one-year program provided by the West Point preparatory school substantially improves college readiness for minority students.

1 Introduction

The growing importance of higher education for economic success has brought increasing pressure on colleges and universities to assess their effectiveness in educating a diverse student body. As recently articulated by the U.S. Supreme Court "*A university's goals cannot be elusory or amorphous – they must be sufficiently measurable to permit judicial scrutiny of the policies adopted to reach them.*" Because of differences in pre-college opportunities, minority students often enter selective institutions with lower entry skills than many of their white counterparts. An assessment strategy is, therefore, needed to compare outcomes of majority and minority students with comparable entering qualifications. The central objective of this paper is to develop a strategy for assessment of whether achievement, attainment and career outcomes of minority students are equivalent to those of majority students with comparable measured entry capabilities. We implement such an assessment strategy using data from West Point.[1]

All institutions of higher education confront a common set of challenges: assessing the capabilities of applicants and selecting those best suited to the mission of the institution, fostering diversity, inculcating knowledge, and placing graduates in productive careers. Colleges are under increasing pressure to evaluate their effectiveness in educating a diverse student body. This focus is being stimulated by several forces that have shaped the market for higher education in recent decades. One is stiffening public resistance to the rapid rise in the cost of education. A second is increasing concern by governments at all levels about whether colleges are making effective use of funds from public programs that are designed to help advance the fortunes of disadvantaged members of the population. A third is rulings by the Supreme Court circumscribing what is permissible with respect to diversity policies. Colleges maintain that there are significant benefits to all students from interacting with a diverse group of peers. This position was affirmed by the U.S. Supreme Court in its 2003 decision (Grutter v. Bollinger) stating that "the Equal Protection Clause does not prohibit the law

[1]For simplicity in the discussion that follows, we use "colleges" to mean both colleges and universities.

school's narrowly tailored use of race in admissions decisions to further a compelling interest in obtaining the educational benefits that flow from a diverse student body"; and "the nation's future depends upon leaders trained through wide exposure to the ideas and mores of students as diverse as this Nation of many peoples."In a subsequent case, the Court later took the firm stand regarding the implementation of diversity policies quoted in our opening paragraph.[2]

Our main objective of this study is to lay out an assessment strategy to permit comparisons of college effectiveness for students differing by race or ethnicity. The key challenge in conducting a reliable assessment of a college's effectiveness in educating a diverse student body arises due to differences in college readiness among enrolling students. This is particularly problematic in the most selective colleges and universities in the U.S. where minority students are typically from more disadvantaged backgrounds than many non-minority students. It is important to account for these initial differences in college readiness when assessing the impact of college education on achievement, attainment, and career outcomes. To overcome this challenge, we primarily rely on matching estimators in the empirical analysis. While matching estimators are often used in program evaluation, they are much more rarely used in assessing differences in achievement and attainment by race, gender or ethnicity.[3] A key advantage of the matching estimators we employ is that they do not require specifying the functional form of the outcome equation and are, therefore, less susceptible to misspecification bias along that dimension of the analysis (Rubin, 1973, 1974).

We implement the assessment strategy using data from the U.S. Military Academy at West Point, which provides a good research setting for our objective for the following four reasons. First, matching requires that the quality of the observed covariates is high. West Point collects detailed data on college readiness as part of the admission process. Hence, the scope

[2]For a discussion of affirmative action in higher education see Epple, Romano, and Sieg (2008).

[3]Matching by race has been used in economic research, discussed in our literature review below, and in medical research, for example in comparing black-white breast cancer survival rates (Silber, Rosenbaum, Clark, Giantonio, Ross, Teng, Wang, Niknam, Ludwig, Wang, Even-Shoshan, and Fox, 2013) and black-white colon cancer survival rates (Silber, et.al., 2014).

of observed measurements of college readiness is comprehensive. Second, matching requires that there is a sufficiently large overlap in the distribution of observed covariates of the two types being matched, otherwise, the region-of-common-support assumption is violated.[4] The match quality has to be sufficiently high for the analysis to be meaningful. It is important to achieve full covariate balance. Match quality can be assessed based on standard difference-in-means tests between matched pairs. Assessing covariate balance for continuous variables entails, in addition, comparisons of the distributions of the matching variables between the two groups being matched. For this, Q-Q plots are particularly useful.[5] Using these criteria, we obtain exceedingly good matches in our sample using the matching method of Abadie and Imbens (2006) implemented in the "genetic" algorithm in the R package named MatchIt.[6] Third, we observe the achievement in core courses that all cadets must take. Hence, we can avoid comparing achievements across majors. Finally, we can follow the military career of the cadets and study retention and early promotion. We do not need to compare career outcomes across different employers which also facilitates the analysis. Hence, we can provide a comprehensive assessment of the effectiveness of West Point in educating a diverse student body.

Matching methods by themselves are not methods of estimation. The approach involves an estimation step following the matching procedure. Hence, in the second stage, we use regressions to test for differences in means for the outcomes of interest. As we noted above, matching methods incorporate the assumption that unobserved covariates can be safely ignored. This is, of course, a strong requirement that may not be met fully in many settings. As part of our analysis, we employ recently developed methods to estimate the effect of unobservables on achievement in our matched samples.[7] This robustness analysis provides an

[4]See, for example, Diamond and Sekhon (2013) and Imbens (2014).

[5]A Q-Q or quantile-quantile plot is a probability plot, which is a graphical method for comparing two probability distributions by plotting their quantiles against each other.

[6]This algorithm uses a nearest neighbor match computed based on a generalized Mahalanobis distance measure with a scaling factor for each covariate. The scaling factors are chosen to maximize a criterion related to covariate balance. See Ho, Imai, King, and Stuart (2007, 2011) for a detailed discussion.

[7]We employ the Altonji, Elder and Taber (2005) approach as extended by Oster (2019) to investigate the potential importance of omitted variable bias.

omitted-variable-bias correction of the estimates.

Our empirical analysis reveals four important findings. First, we study attainment at West Point. We find small, insignificant differences in graduation rates between black and white students and between Hispanic and white students. Second, we focus on achievement among the subsample of cadets who graduate from West Point. Our achievement analysis finds that there are significant black-white achievement gaps for students in our matched samples. This finding holds for broad measures of academic achievement including the graduating GPA and the GPA in core courses. We can rule out a number of potential explanations for this gap. We find that measures of parental education do not affect these findings, nor do measures of cadets' home location. In contrast, we find no significant gaps in achievement between white and Hispanic cadets. Third, we study career outcomes, including retention in the Army and early promotion. We find small and insignificant black-white and Hispanic-white differences. Matched majority and minority cadets are equally likely to have comparable performance on all career outcome measures. Fourth, we study efforts taken by West Point to reduce the gap in college readiness. West Point is unique since it is affiliated with its preparatory school, which provides ten months of preparatory education for less qualified applicants. We find that this preparatory school significantly enhances students' college readiness. In particular, there are substantial and significant achievement gains for both black and white students. This is an important finding suggesting that there may be a potential for developing preparatory programs for other selective colleges.

Finally, we have also conducted a parallel analysis of gender, with the details in Appendix A of this paper. Summarizing, we do not find any differences in academic achievement by gender. However, we find that females have significantly lower graduation and retention rates. We can only speculate on what might explain these three differences, but it is of interest for the Academy to investigate this given the costs of losing students and officers.

The rest of the paper is organized as follows. Section 2 provides a brief literature review. Section 3 introduces our data set. Section 4 presents our analysis of attainment, retention,

and early promotion. Section 5 provides our achievement analysis. Section 6 focuses on college readiness and studies the effectiveness of the preparatory school that is affiliated with West Point. Section 7 concludes and discusses future research. The appendix contains a detailed analysis of gender gaps at West Point.

2 Literature Review

Our paper adds to research analyzing the black-white achievement, attainment, and earnings gaps in the United States. Smith and Welch (1989) published their seminal work on the evolution of black-white inequality during the 20th century. Since that paper, it has been well documented that there have been persistent differences between high school completion rates of white and black students in the United States. Evans, Garthwaite and Moore (2016) report that the gap in high school graduation rates fell by 37% between 1965 and 1986, decreasing from 15.3 to 9.6 percentage points. Then, this progress stopped. Black-white high school graduation rates further diverged until 1997, when the gap was 14.4 percentage points. This gap began to narrow again in roughly the year 2000 as US graduation rates increased, particularly for black and Hispanic students (Murnane, 2013).

A similar pattern arises for achievement measured by standardized test scores. Neal (2006) used data from the National Assessment of Educational Progress. He showed that reading and math scores for black students in urban areas fell during the 1980s relative to scores for other youth. Further, although aggregate black-white gaps in achievement continued to shrink for much of the 1980s, there is considerable evidence that overall black-white skill convergence had already stopped by the time Smith and Welch (1989) published their findings. In 2012, black-white gaps in NAEP math and reading scores of 13-year-olds were virtually the same as in 1990. Assessment of whether this gap has changed awaits results of the NAEP 2019-20. The achievement gap arises prior to high school. Fryer and Levitt (2004) study the early emergence of the black-white achievement gap, focussing on the first two years of school. They show a substantial initial gap in cognitive skills entering kindergarten that can be fully

explained by non-race controls. However, by the end of second grade, the gap increases significantly, their best explanation being school quality differences. Hanushek and Rivkin (2009) show that the black-white achievement gap continues to widen in grades 3 through 8 and that most of this occurs at the upper end of the distribution. They provide evidence that school characteristics, specifically inexperienced teachers and a high proportion of black students, can explain some of this divergence. There are also persistent differences in labor market outcomes by race. Card and Krueger (1992) document differences in earnings between black and white workers. Neal and Rick (2016) show that, relative to white men, labor market outcomes among black men are no better now and possibly worse than they were in 1970. Neal and Johnson (1996) provide evidence using AFQT scores that about 3/4 of the black-white wage gap of those in their late 20's can be explained by achievement differences in the mid-teens. Black, Haviland, Sanders, and Taylor (2006) employ a matching estimator to estimate racial wage gaps of college-educated individuals. They find that all of the wage gap of college educated Hispanics and blacks not from the south is explained by premarket factors, but most of the gap remains for blacks from the south.

There are a number of hypotheses to explain the earlier black-white convergence in educational outcomes including improved parental education (Cook and Evans, 2000), reduced segregation (Jaynes and Williams, 1989), increased school spending (Boozer, Krueger and Wolkon, 1992), changes in within-school factors for integrated schools (Cook and Evans, 2000), and parenting practices (Thompson, 2018). Less attention has been given to understanding the long lull in the convergence and research has struggled to determine why it occurred. Evans, Garthwaite and Moore (2017) examine the emergence of crack markets as an explanation for the stalled progress in black high school completion rates. Neal and Rick (2014) argue that the rise in the incarceration rate for black men largely explains why there has been no progress in labor market outcomes during the past decades. Murnane (2013) provides a summary of this body of research as well as a discussion of factors that may have resulted in the increase in graduation rates from 2000 to 2010 and the narrowing

of the black-white gap during that period.

West Point is most similar to highly selective colleges and universities with strong STEM programs. Some research in higher education has focused on minority participation and graduation in STEM majors. This research has shown that graduation of minority and women students that choose STEM majors is low and, respectively, significantly below that of non-minorities and men. While the proportion of minority students that begin as STEM majors in four-year colleges has a been somewhat higher than whites: 18.6 percent of blacks and 22.7 percent of Hispanics compared to 18.5 percent of whites in 1995-96 (Anderson and Kim, 2006), the respective percentages that persisted and graduated in a STEM major were 41.8, 48.6, and 69.3. These persistence values are high relative to those found in other studies, perhaps because of the inclusion of non-selective colleges. Griffith (2010) calculates persistence-to-graduation rates in a survey of 28 selective colleges and universities of minorities and females that began a STEM major in 1999 equal to, respectively, 35.8 percent and 36.5 percent. The respective values for non-minorities and males were 46.2 percent and 43.1 percent. Griffith provides evidence that students in schools with higher undergraduate to graduate student ratios are more likely to remain in major, consistent with West Point's undergraduate focus, but graduation rates are much higher at West Point across all sub-groups. Arcidiacono, Aucejo, and Hotz (2016) estimate a discrete choice model of school, major, and persistence-to-graduation using late 1990s data from California's university system, during a period when affirmative action in admissions was practiced at the top universities in the system (e.g., Berkeley). Throughout the UC system, persistence to graduation of minority STEM majors was 24.6 percent (within 5 years). Their estimates predict this could have been modestly increased by minorities attending lower-ranked UC schools for those in the bottom two quartiles of prior achievement.[8] They predict that minorities in the upper two quartiles of prior achievement would not have gained by attending a lower-ranked school. The persistence to graduation in STEM majors of the top quartile minority and non-minority students (on

[8]See Arcidiacono and Lovenheim (2016) for a lucid review of the literature on "mismatch," the hypothesis that less prepared minorities attend too rigorous colleges, e.g., as a result of affirmative action in admissions.

the same scale) in the two highest-ranked schools were not drastically different, respectively 52.1 percent and 58.1 percent, but these values dropped to 28.9 percent and 45.1 percent among the third quartile students (Table 4, p. 538). Again, we find much higher persistence at West Point and virtually no difference in graduation rates between matched minorities and non-minorities, while being able to use much more detailed data on prior achievement.

We discussed in the introduction the variables required for evaluating college effectiveness in educating different demographic groups. In addition, the environment being studied must serve a sufficient number of minority students to permit making meaningful comparisons with majority students. For smaller institutions, data for multiple cohorts will be required to obtain adequate sample sizes. Our analysis for West Point utilizes data for 11 cohorts. The approach can also be applied, for example, to large graduate professional programs and seems particularly well suited for large MBA programs.

Our work complements Arcidiacono, Aucejo, and Hotz (2016), who model college choices head-on and use college application sets to control for non-observables among students, following the approach of Dale and Krueger (2002,2014). We add to this body of research by studying attainment and achievement of students by race in a single institution, West Point, with a large database, a diverse body of students, commonality of types of courses across academic measures, extensive measures of entering qualifications of students, and measures of achievement, attainment, and post-college outcomes. Another educational realm where matching has been employed is to assess the performance on achievement of charter schools relative to traditional public schools. See CREDO (2009, 2013), and Sass, Zimmer, Gill, and Booker (2016). One difference is that student matches are between schools in these charter-traditional public school comparisons while our analysis is within a college.

Finally, our paper is related to research that has studied educational practices and outcomes at the USMA. Lyle (2007, 2009) estimates the impact of peer effects and role model effects on human capital accumulation, exploiting random assignments of cadets to social groups at the USMA. Lyle and Smith (2014) estimate the effect of high-performing mentors

on the promotion of junior officers.

3 Data

We implement our assessment strategy for West Point which is similar to other undergraduate colleges in many ways. It is a four-year coeducational undergraduate institution offering 36 academic majors. Students take 40 courses of which 32 are on subjects typical of other undergraduate colleges. The remaining 8 focus on the development of military knowledge and skills. Implications drawn from West Point are likely to apply most directly to technically oriented undergraduate colleges. Of the 36 academic majors at West Point, 23 are in STEM areas, and all graduates of West Point receive a Bachelor of Science degree. In USNews 2021 rankings, West Point was ranked number 11 among National Liberal Arts Colleges and number 2 among Top Public Colleges.

Admission to West Point is largely determined by the Whole Candidate Score (WCS) which is a comprehensive measure of entering capabilities. The WCS is a weighted composite score that incorporates high school academic performance, high school rank, SAT scores, leadership potential, and physical fitness. In particular, 60 percent of the WCS is based on the college entrance examination rank (CEER). The CEER score in turn factors in SAT or ACT scores, as well as the high school rank convert score (HSRCS), which accounts for high school rank adjusted for differences in high school quality. The remaining 40 percent of the WCS is computed based on the three leadership scores and one physical fitness score, each accounting for 10 percent of the WCS. The four measures are the following: (1) the faculty appraisal score (FAS); (2) the athletic activities score (AAS); (3) the extracurricular activities score (EAS); and (4) the candidate fitness assessment (PAE). The community leader score (CLS) score is the sum of the first three of the preceding. We observe all these skill measures. In the analysis below, we restrict attention to the academic entry score (CEER), the leadership entry score (CLS), and the physical fitness entry score (PAE).

In addition, we observe several student characteristics including cohort, whether the

student had prior-service in the enlisted ranks, whether the student attended US Military Academy Preparatory School (USMAPS), detailed information about the father's and mother's education level, and the home state of each student. Note that there is no separate application for USMAPS. All students who attend USMAPS will have been evaluated for admission to West Point based on the metrics discussed in the preceding paragraph. Admission officers may choose to offer USMAPS to potential West Point cadets who lack the grades or skills necessary for immediate admission to West Point. Recruited athletes and those with prior Army service typically attend USMAPS.

We have several outcome measures including college attainment and early career outcomes in the Army. We observe whether the student: a) graduated from USMA, b) obtained a commission in the U.S. Army as an officer, c) was retained beyond 5 and 8 years of service, and d) was promoted "below the zone" to major. Graduates have a five-year obligation and can reenlist for (initially) three years with mutual consent. Early promotion to major is termed "below the zone" promotion. We study each outcome separately below.

We also observe several achievement measures for those cadets who graduate from West Point. The most important measure at graduation is rank on the Order of Merit List (OML) which is a comprehensive measure formed as a weighted average of measures of academic accomplishments, leadership potential, and physical capabilities, supplemented by a judgment of relative merit by a board of Army officers. The OML ranks graduating students from best, a rank of one, to worst. The OML is prestigious. Until 2005 it also established the order in which candidates chose among the 16 military branches, and hence determined which candidates obtained the limited positions available in the most highly sought-after branches. We also observe the cumulative GPAs for the three main skill domains, academic, military leadership, and physical skills, as well as in each core course.

Our sample consists of 11 cohorts of cadets that enrolled at West Point between 1998 and 2008. The sample size of all enrolled cadets is 12,992. The final sample we use for our analysis has a total of 11,503 cadets. This sample has 9,892 white cadets, 840 black cadets,

771 Hispanic cadets, 1,450 white female cadets, 191 black female cadets, and 124 Hispanic female cadets. We have complete records for these 11,503 cadets from their time of entry to up to 16 years following graduation. Cadets not included in our analysis are from racial groups too small in numbers to permit accurate comparisons to matched majority students or cadets with missing data for one or more variables.

Summary statistics are provided in Table 1 for black, Hispanic and white cadets. The "retain" variables are for reenlistment beyond the minimum 60 months and for subsequent reenlistment 36 months later. Other variables have obvious definitions.

This table shows that there are substantial differences in entering academic, leadership, and physical scores by race and ethnicity. The academic score differences are of particular importance since the academic score comprises 60% of the Whole Candidate Score. Black and Hispanic cadets are also much more likely to attend the preparatory school. Finally, black cadets are more likely to be female than Hispanic or white cadets. Given these large differences in college readiness and other demographic characteristics, it is essential to account for these differences when assessing the effectiveness of the college.

4 Attainment, Retention, and Early Promotion

Given that minority students more typically come from disadvantaged educational backgrounds, they are likely to be farther from reaching their potential than majority students when starting college. Hence, not surprisingly, minority students enter selective colleges and universities with, on average, lower academic and leadership skills. The central objective of this paper is to develop a strategy for assessment of whether achievement, attainment and career outcomes of minority students are equivalent to those of majority students with comparable measured entry capabilities. We demonstrate the application of this methodology with black-white and Hispanic-white comparisons using data for West Point.

We first report our findings comparing black and white cadets. As noted above, we have data for 840 black cadets and 9,892 white cadets. To assess the effectiveness of West Point

training by race and ethnicity of cadets who have comparable skills upon entry we employ matching. Matching is a particularly promising approach in the West Point setting because there is a large pool of white cadets for matching, and there is an overlap of the score distributions. This "common support" is portrayed in Figure 1 for black and white cadets. Inspection of these plots reveals that, for each score, the histogram for black candidates falls within the histogram for white cadets.[9]

To clarify the matching procedure, let $x_1, x_2, .., x_K$ be the covariates to be matched. The objective is to maximize the covariate balance between black cadets and the white cadets chosen as matches. The match need not be one-to-one. For example, suppose a black cadet i has covariate values $x_{1i}, x_{2i}, .., x_{Ki}$ and two white cadets have the same values of all covariates. Then both of the white cadets would be chosen as matches. The reverse can also be true, with two black cadets matched to a single white cadet. These examples are illustrative; the closest available matches may not be exact. These examples also illustrate that the number of matched white cadets need not equal the number of black cadets. If a cadet from one race is matched to more than one cadet from the other race, the algorithm provides weights that permit the use of weighted least squares when estimating differences in outcome for the matched samples.[10]

We match cadets based on entry scores and prior-service measures available to the admissions office of West Point at the time admissions decisions are made. The variables we use for matching are the academic, leadership, and physical scores, gender, an indicator of prior service, and indicator for attendance at USMAPS, and an indicator for both prior service and attendance at USMAPS. We then investigate whether there are differences in achievement, graduation, and career outcomes for the matched cadets.

To assess the quality of the matching algorithm, we begin by comparing the means of

[9] A close inspection of the upper left panel of Figure 1 reveals that there is an outlier at the lower end of the CEER distribution. We have investigated robustness and find that the results reported below are not sensitive to whether this outlier is included.

[10] For details regarding the computation of the weights, see Ho, Imai, King and Stuart (2011). In our case, WLS and OLS yield almost identical results, as shown in Table 4.

the covariates that we use in the matching algorithm for black cadets to the means for the matched white cadets. This comparison is done in Table 2 using standard difference-in-means tests. It reveals that the means in both subsamples match up exceedingly well for all of the variables used in the analysis. Assessing the match quality is not only the most natural starting point from a research perspective, but it is exceedingly important from the perspective of the academy to determine whether any systematic differences by race remain once one matches on the variables that impact admissions. As shown in the note to the table, the number of matched white cadets is somewhat lower than the number of black cadets. This arises when a given white student is the closest match to more than one black student. The same is true for the match of Hispanic and white cadets. As explained above, we also report weighted least squares estimates that take account of the presence of a match of a cadet of one race/ethnicity to more than one cadet of another race/ethnicity.

We next compare the distributions of the three continuous entry score variables for the matched sample. This is done in Figure 2 which provides quantile-quantile plots of the three continuous variables used in matching. For example, the upper-left graph for black cadets plots the quantiles of the academic score for black cadets (vertical axis) and the matched white cadets (horizontal axis). A perfect match would have all observations lying on a 45-degree line. The graphs for academic, leadership and physical scores show that the distribution of each of these variables for black cadets is very close to the distribution of the corresponding variable for the matched sample of white cadets.[11]

Having established that we have a high-quality black-white match, we turn to the analysis of outcomes, i.e., the second stage of the analysis. Table 3 reports our findings concerning four binary outcome variables: graduation, retention in the Army after 5 years from graduation, retention after 8 years, and early promotion to the rank of Major. These are important outcome measures for West Point. The top panel reports the results of four regressions for our matched sample of black and white cadets. In each of these regressions, the dependent

[11]The three graphs on the right side of Figure 2 reveal high-quality matches for Hispanic and matched white students.

variable is an outcome variable, and the independent variable is an indicator equal to 1 if the cadet is black and 0 if white. Hence, for each regression, the intercept is the mean of the dependent variable for white cadets, and the coefficient of black is the difference in the means of the dependent variable between black and white cadets. We also report heteroskedasticity-robust standard errors.

From the regression in the Column entitled "Graduation", we see that the estimated graduation rate for white cadets is 77.7% while the estimated graduation rate for black cadets is 75.6%. The estimated -2.1 percentage point difference in graduation rates between black cadets and the matched white cadets has a p-value of .327. Hence, there is not a significant difference in graduation rates. Students do not incur any obligation to military service unless they attend West Point beyond their second year. Put differently, students can obtain two years of tuition-free education at West Point if they leave at the end of the second year. Considered in this light, the West Point graduation rates are impressive.

From the second and third columns in Table 3, we see that the estimated differences in retention rates between black cadets and the matched white sample are all quantitatively small and statistically insignificant. Thus, five-year and eight-year retention rates are comparable for black and matched white cadets; just under 60% are retained beyond 5 years and roughly one-third are retained beyond eight years. Rates of early promotion to major are also comparable at approximately 2.5% as shown in the last column of Table 3.

We also conducted robustness checks summarized in Table 4. First, we added the variables that we use in matching during the second stage of the regression analysis. Including these variables may improve the efficiency of the second-stage estimator. Second, we add parental educational background variables to the regression. Third, we matched on all of the previously enumerated variables and on parental education variables. In particular, we included eight parental education variables in matching. Those variables denote, for each parent, whether the parent is a high school dropout, a high school graduate, has some college, or has a bachelor's degree. All results are negligibly changed by the inclusion of these extensive

measures of parental education as matching variables. Finally, we estimate the models using weighted least squares with matched variables as controls. Overall, we find the coefficients of black continue to be negligible in magnitude and statistically insignificant. Hence, this reinforces our finding of small and insignificant differences in these outcomes between black and white students.

Matching implicitly invokes the assumption that the covariates used in the match impound the relevant information for the evaluation of outcomes. Put differently, variables not included in the match are ignorable. As explained above, we analyzed whether parental education variables would enhance the prediction of outcomes for admitted students. This is a strong test because the education of parents undoubtedly plays an important role in influencing the education of their children. Moreover, as noted above, the measures of parental education obtained by West Point are extensive. Regression results for binary outcome variables with the match variables and parent education variables included reveal that the estimated black-white differences continue to be small and insignificant. As shown in Table 4, the inclusion of these parental education variables has a negligible effect on any of the estimated black-white differences. In the interest of space, we do not report the coefficients of those variables, but the p-values of the parental variables are all very high. Joint tests of the significance of parental variables also yield very high p-values. These findings continue to hold if parental variables are aggregated into a smaller group of categories, e.g., one category for college graduates or higher. These findings provide valuable evidence for the robustness of our matching analysis.

We next turn to the results for Hispanic and white cadets. Since the analysis proceeds along the same lines as above, we just summarize the main findings. From the lower panel of Table 2, we see that the means of the variables for Hispanic and white cadets are virtually identical in the matched sample that we created. The QQ plots for the three continuous variables shown in the right panel of Figure 2 also indicate that the quality of the match is very good. We, therefore, conclude that the matching algorithm works well in this application. From the regressions in Table 3, we see that the differences in binary outcomes for Hispanic

and white cadets are all quantitatively small and statistically insignificant once we control for differences in the key characteristics that are used during the admission process at West Point. Hence, we find that graduation, retention, and early promotion rates are very similar for comparable Hispanic and white cadets.

In summarizing, we have shown that there is not a significant difference in graduation rates between black cadets and their matched white counterparts. In addition, the career outcomes of black cadets and their white counterparts are very similar. The gaps in 5-year and 8-year retention rates and rates of early promotion to major are quantitatively small and statistically insignificant. Similarly, there are no systematic attainment or career gaps between comparable Hispanic and white cadets.

5 Achievement

Thus far, we have compared outcomes for the matched sub-samples of admitted cadets. Since we can only measure achievement for those cadets who graduate, we now restrict our attention to matched subsamples of graduating cadets. Comparison of students matched well at the point of admission is very informative about a college's success in educating a diverse student body. Table 14 in Appendix B shows that our matched sample of entering black and white cadets proves to be well balanced among those who persist to graduation. We, therefore, continue our analysis by focusing on differences in achievement among graduating cadets drawn from the same set of students matched at enrollment. The empirical findings are summarized in Table 5. Our first measure of achievement is the position on the Order of Merit List which is a comprehensive ranking of all graduating cadets. Table 5 shows that the estimated difference in graduating OML rank between black and white cadets is 84.6 and statistically significant. Recall that lower OML is better. Hence this result tells us that black cadets who graduated had less favorable rankings than the matched white cadets. This difference is quantitatively large.

Next, we focus on academic, physical, and leadership measures. These skills are measured

by cumulative grade point averages in the relevant courses at the time of graduation. Table 5 shows that black cadets have significantly lower graduating academic scores and academic scores in common core areas. The point estimates are -0.132 for all academic courses and -0.145 for core courses, which are approximately a third of a standard deviation of these scores in the matched sample. Table 5 also documents that black students have significantly lower leadership and physical scores than their matched white counterparts.

To investigate the robustness of these results, we undertook a variety of sensitivity checks. In the first panel of Table 6 we include all variables used in matching in the regression. In the second panel, we also add parental education to the regression model. In the third panel, we use weighted least squares instead of least squares to estimate the regression model. In the fourth panel we rematch the sample, using only data for graduates. As with previous matches, the match quality for all of these variables is exceeding good. A comparison of these results reported in Table 6 with the baseline results in Table 5 reveals that the results are little affected by the changes.

To gain more insight into the achievement gaps, it is useful to compare the performance of black and white cadets in individual courses. We converted scores to standard deviations to facilitate interpretation. The coefficient denoted "black" in Table 7 is the black-white difference measured in standard deviations. For example, we find that the average score of black students in the International Relations course was .143 standard deviations below the average score of matched white students. In this table, we distinguish between three types of courses. The first set of courses is mandatory; students cannot test out of them. The second set of courses can be tested out, but it happens rarely. Finally, the last courses are those for which students frequently test out. Table 7 summarizes our results. While the rate of testing out of the last group of courses is high in the West Point student body, relatively few students in our matched black-white sample test out of those courses. Hence, we can meaningfully compare black-white performance in all of the courses shown in Table 7.

The results in Table 7 have the potential to help West Point focus efforts to close the

academic black-white performance gap by identifying courses in which black-white differences are greatest. Among courses with a .15 sd gap or more, the gap in Environment and Geography is by far the largest (.41 sd), followed by English Literature (.24 sd), Physics II (.23 sd), Military III (.21sd), Economics (.2 sd), Leadership (.18 sd), Military I (.18 sd), American Politics (.16 sd), and Military II (.15sd). In addition, the results highlight some courses in which there are no significant differences. Most noteworthy among the latter, there are not significant differences in scores between black students and their matched white counterparts in probability and statistics and the two mathematics courses. Performance in those courses suggests that the foundation is laid for closing the achievement gap in physics between matched black and white students. The ordering by the magnitude of gaps highlights substantial and significant gaps in the Military Science courses taken in each of the first three years and the Leadership course taken in the third year. It seems likely that there is a large overlap in the skills needed for success in the military science and leadership courses, suggesting an effort to emphasize the development of those skills for black cadets. Some encouragement in this regard is provided by the absence of a significant difference in the fourth-year Military Science course. This type of analysis of performance by race in core courses might prove fruitful for other colleges and universities.

As a final robustness check, we also use the techniques suggested by Oster (2019) to test how sensitive the estimates are to omitted variables. This technique extends the methods developed by Altonji, Elder, and Taber (2005) who formalize the idea that "the selection on the unobservables is the same as the selection on the observables (p. 169)."[12] In their framing of the approach, they assume that a regression including both observables and unobservables would have an R^2 equal to one. Oster (2019) extends their analysis to permit the corresponding R^2 to be less than one. She notes, for example, that a regression with both observed and unobserved explanatory variables would have an R^2 less than one if there is a measurement

[12]The computations were done using the R package Robomat. We calculate standard errors using a leave-out-one-at-a-time approach, known as jackknifing. Using the jackknife results, we calculate confidence intervals using the approach set forth in Sawyer (2005).

error in the dependent variable. The framework of Altonji, Elder, and Taber (2005) is designed for applications in which the focal indicator variable in the regression is itself a choice variable. In their application, the choice variable denotes attendance at a Catholic or public school. In our application, the focal indicator variable denotes race, which is, of course, not a choice variable. Our interest is in the extent to which differential academic performance of black and white students can be accounted for by unobserved characteristics that reflect the different pre-college experiences of black and white students. The approach as extended by Oster (2019) is well-suited to this purpose.

Recall, that our baseline regression uses graduating academic GPA as the dependent variable. We include an indicator for black, prior academic, physical skills, and leadership scores, an indicator for USMAPS, an indicator for prior service, and an indicator for both USMAPS and prior service. In the least squares regression reported in the upper panel of Table 6, the coefficient of black is -.103 with a standard error of .019. The $R^2 = .35$ (not reported in Table 6). To apply Oster's procedure, it is necessary to specify R^2_{max}, which is the R^2 that would be obtained if the unobserved variables were included in the regression. Of course, this value is unknown. Hence, a natural strategy is to apply the method for a range of possible values of R^2_{max}. Figure 3 shows the estimates of the coefficient of black adjusted for bias from omitted variables and the 95% confidence intervals as a function of R^2_{max}.

The estimate of the coefficient or black decreases as R^2_{max} increases, which suggests that the uncorrected estimator is biased away from zero. A higher value of R^2_{max} implies the omitted variables have more explanatory power. That the adjusted coefficient declines, then means these variables, if included, would help to explain the lower achievement scores of blacks. As the graph shows, the estimated coefficient of black is negative throughout the range of values of R^2_{max}. The null hypothesis that the coefficient of black is zero is rejected for values of R^2_{max} up to $R^2_{max} = .85$. While the R^2_{max} is unknown, we might expect the value to be on the order of twice that for the observed measures, e.g., a range of .60 to .80.[13] If

[13]One way to bound the value of R^2_{max} is to estimate a lower bound of the variance of the measurement error in the outcome variable. Given that measurement error in achievement is likely to non-negligible, it

so, the graph shows that the estimate of black would be in the range from -.08 to -.06, and significantly different from zero at the 5% level.

Figure 4 shows a similar analysis for graduating GPA in core courses. The observed explanatory variables for the analysis of graduating GPA in core courses are the same as in the analysis of graduating GPA. The least squares regression coefficient is -.110 and the standard error is .02 as reported in Table 6. The findings for core GPA are very similar to those in the previous graph for graduating GPA. The coefficient corrected for omitted variables is negative throughout, and the null hypothesis of zero is rejected at the 5% level for values R^2_{max} up to $R^2_{max} = .85$. If R^2_{max} falls in the range from .60 to .80, the graph below shows that the estimate of black would be in the range from -.085 to -.065, and significantly different from zero at the 5% level. We thus conclude that the estimated black-white difference is reduced on the order of 20 to 25 percent by accounting for unobservables.

We thus conclude that black cadets have significantly lower achievements measured by cumulative GPA scores at graduation, and significantly less favorable positions on the order of merit list than comparable white cadets. These differences can only partially be explained by unobserved characteristics.

Next, we turn to the results for Hispanic and white cadets.[14] From the regressions in Table 5, we see that the differences in OML rank and academic, leadership and physical scores are quantitatively small, with none being close to significant except for the leadership measure. We thus conclude that there are no systematic achievement gaps between comparable Hispanic and white cadets in the matched sample.

plausible that .8 is a realistic upper bound for R^2_{max}.

[14]Again we conduct some additional balance tests on the subsample of the graduates. The results are reported in Table 14 of Appendix B of the paper. We find that the differences between white and Hispanic cadets in entering academic and leadership scores are quantitatively small. We find that only the difference in physical fitness scores is significant. Thus, there is little indication of differential selective attrition between matched Hispanic and white cadets.

6 College Readiness

From a broader policy perspective, we would like to know what colleges can do to close the racial and ethnic achievement gaps. A unique feature of West Point is that it is affiliated with its preparatory school, the US Military Academy Preparatory School, known as USMAPS. An offer of admission to USMAPS may be provided to a West Point applicant who initially lacks the grades or skills necessary to succeed at West Point. This school provides an opportunity for would-be cadets to improve their skills and increase their college readiness. Students who are recruited as athletes and students with prior Army service in the enlisted ranks also typically attend USMAPS.[15] The preparatory school serves a substantially higher proportion of minority students than West Point. Here we focus on the black-white comparison.

Next, we analyze the gains in college readiness scores focusing on academic, leadership, and physical scores. One nice feature of this analysis is that entrants to USMAPS take the entry examinations to West Point and they take those examinations again after completing USMAPS. Hence, we have measures at the beginning and the end of the preparatory school for all metrics used in admission by West Point. We can differentiate the scores and compute the gains for each student. Table 8 summarizes the empirical results for our analysis of the gains.

Our analysis reveals that cadets who attended USMAPS significantly improved their academic and leadership skills during that year. It is useful to normalize these estimates and put them on a common scale. Dividing the estimates in Table 8 by the standard deviations of all entering West Point students, we find that black students gain .47 standard deviations in academic score, .13 standard deviation in fitness, .35 standard deviation in SAT, and .44 standard deviation in leadership. Matched white students gain .62 standard deviation in the academic score, .47 standard deviation in the leadership score, .45 in SAT, but lose

[15]USMAPS enrolls approximately 240 students per year. Our sample has 1,650 students who subsequently were accepted at West Point. Thus, more than 60% of USMAPS graduates ultimately gain admission to West Point. Nevertheless, there is a potential selection problem associated with this analysis since we do not have access to the full sample of students enrolled in the USMAPS.

.2 standard deviation in the fitness score. The academic and leadership score gains are impressively large, especially given that they are achieved in a ten-month program. While the academic and leadership gains are somewhat smaller for black students than for white students, the gains for both groups are very large. Black students also have modest gains in physical fitness whereas comparable white students have a significant decline in fitness. We thus conclude that the one-year remedial program provided by the West Point preparatory school substantially improves college readiness for all students including minority students.

A fundamental challenge for West Point and other selective colleges and universities is increasing the number of minority graduates. This in turn requires attracting more minority applicants and taking measures to compensate for the difference in preparation between minority and majority students. Our analysis of the preparatory school demonstrates the effectiveness of the additional year of education with a curriculum designed to enhance the capabilities required for admission to West Point.

There is a large literature in labor economics and the economics of education that has analyzed various interventions and programs that have tried to improve higher education, especially for students with disadvantaged backgrounds.[16] The lack of academic preparation often limits students' ability to benefit from the opportunities that higher education offers to students from advantaged backgrounds. Oreopoulos (2021) argues that most interventions discussed in Holzer and Baum (2018) are too expensive to justify implementation. Nevertheless, he recommends making mandatory some existing services, such as application assistance and advice, proactive tutoring and advising, and greater career transition support has the most immediate potential. Our analysis of USMAPS suggests another promising intervention. Selective colleges and universities can potentially benefit from the experiences of West Point since they face similar challenges in attracting low-income and minority students who are often not sufficiently well-prepared for the academic rigors of advanced undergraduate education. It is worth noting that the preparatory program is not remedial education pro-

[16]Holzer and Baum (2018) provide a recent survey of this literature and formulate a plan for action for policymakers.

vided to students who are admitted to West Point. Students graduating from USMAPS gain admission to West Point only if they perform well enough on the West Point entry examinations. This aligns the incentives. Students at USMAPS have an opportunity to ramp up their skills, and West Point admits students only if they have higher admissions scores than those not admitted. This should be borne in mind as other institutions consider the design of preparatory programs to help close the college readiness gap. Selective colleges might join together in a collaborative effort to create and operate such a preparatory program. The success of the Consortium for Graduate Study in Management (cgsm.org) demonstrates the impact that can be made by cooperative programs of selective institutions.

7 Conclusions

There are large differences in college readiness between minority and majority students enrolling in most selective colleges and universities, including West Point. These differences reflect the reality that minority students typically come from more disadvantaged backgrounds than majority students. To assess the relative effectiveness of a college in educating minority and majority students, it is necessary to control for differences in the college readiness of entering students. We present an approach for making meaningful comparisons of outcomes across demographic groups, employing modern matching estimators and using data from West Point.

Our empirical analysis reveals several important findings. First, we find that minority students at West Point have similar graduation rates. We find no evidence that would suggest the existence of an attainment gap at West Point. Second, there is an achievement gap between black and matched white students while in college. We have ruled out several possible explanations such as differences in parental education, home location of the student, and athletic status. Moreover, we can pinpoint the courses that seem to be driving the achievement gap. This in turn suggests a domain in which more resources might be invested to reduce the achievement gap. It is important to note that we do not find an achievement

gap when we compare Hispanic and matched white students, which is encouraging. Third, we find that minority students at West Point have similar early career outcomes as their matched white counterparts. This finding is important since it shows that the achievement gap may not have a large impact on the early stages of the cadets' careers. Fourth, the one-year program provided by the West Point preparatory school substantially improves college readiness for minority students. Hence, there is a proven intervention that can be used to reduce the college readiness gap between minority and majority students. Finally, we have conducted a gender-gap analysis. We find that there are no significant differences in achievement by gender. However, females are less likely to graduate from West Point and have lower retention rates than males.

In the context of West Point, we see our work as providing findings that may help the U.S. Army address the large imbalance in racial composition of the officer corps relative to the enlisted ranks. An exceptionally distinguished group of military officers wrote cogently of the importance of diversity of the officer corps for achieving the objectives of the army: "Based on decades of experience, amici have concluded that a highly qualified, racially diverse officer corps educated and trained to command our nation's racially diverse enlisted ranks is essential to the military's ability to fulfill its principal mission to provide national security."[17]

The findings of this paper provide ample scope for future research. West Point is, undoubtedly, a special institution, and there are some clear differences between military academies and selective liberal arts colleges. Given the range of skills required for a military officer to be effective, the challenge of measuring the effectiveness of the education may be more daunting for service academies than for civilian academic institutions. However, the main focus of our analysis has been on studying achievement and attainment gaps in college. When it comes to the academic education of cadets, West Point ranks among the top liberal arts colleges. Hence, we think that there is external validity to this part of our analysis. In contrast, the analysis of early career paths is special to military academies. Nevertheless, we think that

[17]Military Amicus Brief cited in Supreme Court's Decision in the University of Michigan Case, Grutter v. Bollinger.

this analysis is insightful, and a comprehensive assessment of any college should seek to assess career outcomes.

The methodology developed in this paper is not specific to West Point but can be broadly applied to other colleges and professional schools. More applications are needed to obtain a better understanding of how effective selective colleges are in educating a diverse student body.[18] Future research that focuses on civilian institutions of higher education will also help to assess the external validity of this study.[19]

[18]The external validity of our analysis may be stronger for males than for females who account for less than 15 percent of the enrollment at West Point.

[19]Appendix C discusses in more detail how our method can be applied to civilian universities and professional schools.

A The Gender-Gap Analysis

In this section, we undertake a matching analysis to investigate outcomes by gender. Table 9 provides a comparison for male and female white cadets. Here we focus on white students because of limited observation of female minorities. The table shows that there are only small differences in entry scores by gender. However, there are significant differences in prior service. Moreover, females are less likely to attend the USMAPS than males.

To account for differences in observed characteristics among male and female cadets we match on a vector of characteristics. Table 10 shows that the means of the key variables used in the matching analysis for female and male cadets are quite close. We, therefore, conclude that the quality of the match is high in our sample.

Turning to the regression results, we see in Table 11 that there are significant differences between female and male cadets in graduation rates and retention rates. Female cadets have a 4.7 percentage point lower graduation rate than male cadets, a 9.3 percentage point lower rate of retention after five years, and an 11.5 percentage point lower rate of retention after eight years. There is not a significant difference between female and male cadets in early promotion (promotion below zone) to major.

In Table 12, we investigate whether there is differential attrition by comparing entry scores of female and male graduates. The regressions reveal that the means for female cadets and matched male cadets are nearly identical. The estimated differences for academic, leadership and physical fitness entry scores are all exceedingly small relative to the means for male cadets (the intercepts in the regressions), and none are anywhere near significant. Hence, the mean values of these three variables for those who graduated are virtually the same for the female cadets and the matched sample of male cadets, indicating no differential selective attrition.

Table 13 compares several achievement measures for female graduates and male graduates in the matched sample. The outcomes of interest are the order of merit (OML), as well as academic, physical fitness and leadership cumulative GPA scores at graduation. For all the

measures, the differences between female and male graduates are negligible in magnitude and statistically insignificant.

In summary, we do not find any differences in academic achievement by gender. However, we find that females have significantly lower graduation and retention rates.

B Additional Balance Checks

The results reported in Table 14 show that our matched sample of entering black and white cadets is still well balanced among those who persist to graduation. Hence, we find no evidence that may suggest selective attrition before graduation.

Table 15 presents differences in means tests for the matched black and white cadets at the US Military Academy Preparatory School. It shows that matching works very well. We construct matched subsamples with virtually identical observed characteristics.

C Scope for Broader Application

Here we discuss the scope for and issues of implementation in applying the methodology generally in higher education. Because there are many types of educational institutions and fields of study, we adopt "program" to denote the entity to be assessed. In some applications, a program may be an area of study within an educational institution. In other applications, it may be the entire institution, as with our study of West Point. We discuss data required for an assessment, and we identify types of programs that appear to be promising candidates for the approach.

To conduct an evaluation of a program, it is necessary to have measures of the skills of students entering the program, and measures of outcomes. The latter will include measures of performance in the program, and, ideally, measures of outcomes following graduation.

The sample size is a key consideration. The program being studied must have a large enough number of minority students to provide statistical power for reasonably precise com-

parisons of outcomes between matched majority and minority students. Some large professional programs may meet this criterion with students from a single cohort. For other programs, data for multiple cohorts will be required. It is thus important that programs obtain and maintain data over time, both with respect to students' entering credentials and outcomes.

Evaluation of performance in a program can most readily be conducted for programs that have a commonality of courses (e.g., a core curriculum) and associated metrics that permit comparison across all students (e.g., GPA in core courses, persistence to graduation). These criteria will be met in many professional programs including masters in business and public policy, nursing, law, and education. These criteria will also be met by other undergraduate programs, as we discuss later in this section.

Measures of outcomes following graduation will be available for many professional programs. Business schools gather data on starting salaries for their MBA graduates, and many business schools conduct exit surveys to assess student satisfaction. Such data can be used to undertake comparisons of starting salaries and satisfaction by race and ethnicity for matched minority and majority students. For some professions, certification is required to practice in a state. Law graduates must pass their state's bar exam. Nurses and dentists must pass their state's licensing exam. The certification exams can be employed to compare first-time pass rates of matched minority and majority graduates.[20]

Consider undergraduate institutions more generally. As we noted above, for colleges that have a common core, GPA in the core can be compared between matched minority and majority students. We employed this approach in our West Point application. This approach can be applied at CalTech, MIT, Georgia Tech, and other institutes of technology that require a common core. Many undergraduate institutions have multiple schools (e.g., engineering,

[20]Passing these exams is by no means automatic. In 2018, the bar exam pass rate for first-time test takers was 74.82%. First-time pass rates in the first half of 2020 for RN and PN were 89.25% and 85.32% respectively. A three-part exam is administered for licensing dentists, and all three parts must be passed. Failure rates for first-time test takers in 2017 for the three exam parts were 10.6%, 8.3%, and 6.2%. Very few graduates in dentistry are black, so there are presently very few dental schools that have sufficient data to make comparisons for black and white graduates.

mathematics and physical sciences, humanities and social sciences). Here we use "school" to denote an area of specialization within a college or university. For many colleges and universities, admission is school specific. Consider engineering. A student wishing to study engineering will typically be required to apply for admission to the engineering school. If an engineering school serves a large enough number of minority students, matching of minority and majority students admitted to engineering can be undertaken. GPA in core courses in engineering can be compared between matched minority and majority students. Retention in engineering can also be compared between the matched students to investigate whether there is differential switching out of engineering by minority and majority students.

In colleges and universities that do not have sufficient numbers of minority students to evaluate outcomes within schools, assessment is more challenging. Nonetheless, much can be done if, across all schools, there is a large enough minority presence. Matching of minority and white students based on entering qualifications can be undertaken. Fields of study chosen by minority students can be compared to those chosen by matched majority counterparts. A key challenge in comparing student GPAs arises from the difference in grading standards across majors and courses within majors. The innovative methods developed by Arcidiacono, Aucejo, and Spenner (2012) can be used to standardize grades across courses and also to adjust for the decreasing variance in course grades that may occur as students progress from freshman to senior year. Graduating GPAs so adjusted can then be compared between minority and matched majority students.

Funding organizations might provide grants to encourage research on assessments for different types of educational institutions. For example, a given grant might focus on the assessment of a specific type of educational program. Such research would serve two valuable purposes. One would be the design of an assessment approach appropriate to a particular type of educational program. The other would be the implementation to demonstrate the feasibility of obtaining the metrics required to assess the particular type of program.

Where undertaking such assessments is feasible, school administrators may be concerned

about demonstrating the poor relative performance of minorities. However, discovering such an unsatisfactory outcome can help schools to take steps to improve. The lessons from the USMAPS program are valuable, especially for selective colleges and universities that face challenges in attracting minority students who are often not yet sufficiently well-prepared for the academic rigors of these schools. It is not clear whether highly selective colleges can close preexisting gaps without offering a more structured and personalized learning experience that is similar to that provided by USMAPS. A collaborative effort of selective colleges to develop such a preparatory program merits consideration. Showing good relative performance of minorities is something an institution could publicize. Competition here would be meritorious. Of course, as articulated by Donald Campbell, using a quantitative measure such as course grades as a social indicator can create a tendency to alter the measure in ways that both distort the measure and undercut progress toward the intended outcome. This tendency can be mitigated if the assessment of program outcomes is done by examinations, such as SATs, that are written and graded externally.

References

Abadie, A. and G. Imbens (2006). "Large Sample Properties of Matching Estimators for Average Treatment Effects." Econometrica, 74(1), 235-267.

Altonji, J., Elder, T., and C. Taber (2005). "Selection on Observed and Unobserved Variables: Assessing the Effectiveness of Catholic Schools." Journal of Political Economy, 113, 151-184.

Anderson, E. and D. Kim (2016). "Increasing the Success of Minority Students in Science and Technology," American Council on Education, Report.

Arcidiacono, P., Aucejo, E. and V. J. Hotz (2016). "University Differences in the Graduation of Minorities in STEM Fields: Evidence from California," American Economic Review, Vol. 106(3), 525-562.

Arcidiacono, P., Aucejo, E.. and K. Spenner (2012). "What Happens after Enrollment? An Analysis of the Time Path of Racial Differences in GPA and Major Choice," IZA Journal of Labor Economics, (1): 1-24.

Arcidiacono, P. and M. Lovenheim(2016). "Affirmative Action and the Quality-Fit Trade-off," Journal of Economic Literature, Vol. 54(1), 3 – 51.

Bagde, S., Epple, D. and L. Taylor (2016). "Does Affirmative Action Work? Caste, Gender, College Quality, and Academic Success in India," American Economic Review, 106(6): 1495-1521.

Bertrand, M., Goldin, C., and L. Katz (2010). "Dynamics of the Gender Gap for Young

Professionals in the Financial and Corporate Sectors," American Economic Journal: Applied Economics, Vol. 2, 228-255.

Bettinger, E. and B. Long (2005). "Do Faculty Serve as Role Models? The Impact of Instructor Gender on Female Students," The American Economic Review, Vol 95(2), 152-157.

Black, D., Haviland, A., Sanders, S., and L. Taylor (2006). "Why Do Minority Men Earn Less? A Study of Wage Differentials among the Highly Educated."The Review of Economics and Statistics, Vol. 88 (2), 300-313.

Black, D., Haviland, A., Sanders, S., and L, Taylor (2008). "Gender Disparities among the Highly Educated," Journal of Human Resources, Vol. 43 (3), 630-659.

Boozer, M., Krueger, A. and S. Wolkon (1992). "Race and School Quality since Brown v. Board of Education," Brookings Papers on Economic Activity–Microeconomics, 269-326.

Campbell, D. (1979). "Assessing the impact of planned social change". Evaluation and Program Planning, 1979, 2 (1): 67–90.

Card, D. and A. Krueger (1992). "School Quality and Black-White Relative Earnings: A Direct Assessment,"Quarterly Journal of Economics, 107 (1), 151-200.

Carrell, S., Page, M. and J. West (2010). "Sex and Science: How Professor Gender Perpetuates the Gender Gap,"The Quarterly Journal of Economics, Vol. 125(3), 1101 -1144.

Cestau, D., Epple, D. and H. Sieg (2017). "Admitting Students to Selective Education Programs: Merit, Profiling, and Affirmative Action," Journal of Political Economy, 125 (3),

761-797.

Colarusso, M., Heckel, D., Lyle, D. and W. Skimmyhorn (2016). "Starting Strong: Talent-Based Branching of Newly Commissioned U.S. Army Officers." Officer Corps Strategy Monograph Series, Volume 9. U.S. Army War College Press.

Cook, M. and W. Evans (2000). "Families or Schools? Explaining the Convergence in White and Black Academic Performance," Journal of Labor Economics 18, 729-754.

Dale, S. and A. Krueger (2002). "Estimating the Payoff to Attending a More Selective College: An Application of Selection on Observables and Unobservables," Quarterly Journal of Economics, Vol. 117 (4), 1491-1527.

Dale, S. and A. Krueger (2014). "Estimating the Effects of College Characteristics over the Career Using Administrative Earnings Data," Journal of Human Resources, Vol. 49(2), 323-358.

Dehejia, R. and S. Wahba (1999). " Causal Effects in Non-experimental Studies: Reevaluating the Evaluation of Training Programs," Journal of the American Statistical Association, 94:448, 1053-1062.

Dehejia, R. and S. Wahba (2002). "Propensity Score-Matching Methods for Non-experimental Causal Studies," Review of Economics and Statistics, February 2002, 84 (1), 151-61.

Epple, D., Romano R. and H. Sieg (2008), "Diversity and Affirmative Action in Higher Education." Journal of Public Economic Theory, 10 (4), 474-501.

Evans, W., Garthwaite, C. and T. Moore (2016). "The white/black educational gap, stalled progress, and the long-term consequences of the emergence of crack cocaine markets." Review of Economics and Statistics, 98 (5), 832-847.

Fisher, R. A. (1935). Design of Experiments, New York: Hafner.

Fryer, R. and S. Levitt (2010). "An Empirical Analysis of the Gender Gap in Mathematics," American Economic Journal: Applied Economics, Vol. 2 (2), 210 -240.

Griffith, A. (2010). "Persistence of Women and Minorities in STEM Field Majors: Is it the School that Matters?" Economics of Education Review, 29, 911-922.

Gu, X.S. and P. R. Rosenbaum (1993). "Comparison of multivariate matching methods: Structures, distances, and algorithms." Computational and Graphical Statistics, 2, 405-420.

Heckman, J., Ichimura, H., Smith, J. and P. Todd (1998). "Characterizing Selection Bias using Experimental Data." Econometrica, 66 (2), 315-331.

Heckman, J., Ichimura, H. and P. Todd (1997). "Matching As An Econometric Evaluation Estimator," Review of Economic Studies, 65(2), 261-294.

Ho, D., Imai, K., King, G. and E. Stuart (2007). "Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference." Political Analysis, 15(3), 199-236.

Ho, D., Imai, K., King, G. and E. Stuart (2011). "MatchIt: Nonparametric Preprocessing for Parametric Causal Inference." Journal of Statistical Software, Volume 42, Issue 8.

Holzer, H. and S. Baum (2017). "Making Colleges Work: Pathways to Success for Disadvantaged Students. Washington, D.C. Brookings Institutions.

Imbens, G. (2015), "Matching Methods in Practice: Three Examples," J. Human Resources, vol. 50 no. 2 373-419.

Jaynes, G. and R. Williams (1989). "A Common Destiny: Blacks and American Society," Washington, DC: National Academy Press.

Lyle, D. (2007). "Estimating and Interpreting Peer and Role Model Effects from Randomly Assigned Social Groups at West Point." Review of Economic & Statistics, 1-20.

Lyle, D. (2009). "The Effects of Peer Group Heterogeneity on the Production of Human Capital at West Point, " American Economic Journal: Applied Economics, 1:4, 69-84.

Lyle, D. and J. Smith (2014). "The Effect of High-Performing Mentors on Junior Officer Promotion in the U.S. Army," Journal of Labor Economics, 32, 2, pp. 229-58.

Maddi, S. R., M. D. Matthews, D.R. Kelly, B. Villarreal, B., and M. White, M. (2012). "The role of hardiness and grit in predicting performance and retention of West Point cadets." Military Psychology, 24(1), 19-28.

Murnane, R., (2013). "U.S. High School Graduation Rates: Patterns and Explanations," Journal of Economic Literature, Vol. 51(2), 370-422.

Neal, D. (2006). "Why Has Black-White Skill Convergence Stopped?" in: E. Hanushek

and F. Welch, eds., Handbook of the Economics of Education, vol. 1. Oxford: Elsevier North-Holland.

Neal, D. and W. Johnson (1996). "The Role of Pre-Market Factors in Black-White Wage Differences," Journal of Political Economy, Vol. 104, 869-895.

Neal, D. and A. Rick (2016). "The Prison Boom and the Lack of Progress after Smith and Welsh." The Journal of Legal Studies, 45 (1), 1-41.

Oreopoulos, P. (2021). "What Limits College Success? A Review and Further Analysis of Holzer and Baum's Making College Work." Journal of Economic Literature, 59 (2): 546-73.

Oster, E. (2019). "Unobservable Selection and Coefficient Stability: Theory and Evidence." Journal of Business & Economic Statistics, 37 (2), 187-204.

Rosenbaum, P. and D. Rubin (1983). "The Central Role of the Propensity Score in Observational Studies for Causal Effects," Biometrika, 70,41-55.

Rubin, D. (1973). "Matching to Remove Bias in Observational Studies". Biometrics. 29 (1): 159-183.

Rubin, D. (1974). "Estimating Causal Effects of Treatments in Randomized and Non- Randomized Studies," Journal of Educational Psychology, 66 (5), 688-701.

Sass, T., Zimmer, R., Gill, B., and T. Booker (2016). "Charter High Schools: Effects on Long-Term Attainment and Earnings," Journal of Policy Analysis and Management, V. 35(3), 683-706.

Sawyer, S. (2005). "Resampling Data: Using a Statistical Jackknife," Washington University Lecture Notes.

Smith, J. and F. Welch (1989). "Black Economic Progress After Myrdal," Journal of Economic Literature, 27(2), 519-564.

Thompson, O. (2018), "The Determinants of Racial Differences in Parenting Practices," Journal of Political Economy 126, 438-449.

Tables

Table 1: Descriptive Statistics by Race and Ethnicity: Full Sample

	Variable	Black	Hispanic	White
	academic score	0.551	0.585	0.607
entry	leadership score	0.611	0.602	0.620
scores	physical fitness	0.574	0.537	0.552
	male	0.766	0.839	0.853
demographics	usmaps	0.419	0.263	0.111
	prior service	0.057	0.088	0.068
	maps & prior service	0.052	0.071	0.052
attainment	graduate	0.785	0.763	0.809
	retain 60	0.602	0.633	0.627
career	retain 96	0.357	0.372	0.378
	early promote major	0.026	0.025	0.037
number of observations		840	771	9,892

Note: This table presents descriptive statistics for 11,503 cadets in the 11 cohorts of cadets that enrolled at West Point between 1998 and 2008.

Table 2: Difference-in-Means Balance Tests: Matched Sample

Variable	Black	White	Difference
academic entry score	0.5510	0.5508	0.0003
physical fitness entry score	0.5753	0.5742	0.0011
leadership entry score	0.6044	0.6060	-0.0016
male	0.7726	0.7750	-0.0024
usmaps	0.4167	0.4179	-0.0012
prior service	0.0679	0.0619	0.0060
usmaps & prior	0.0536	0.0536	0.0000
Variable	Hispanic	White	Difference
academic entry score	0.5852	0.5852	0.0000
physical fitness entry score	0.5373	0.5373	0.0000
leadership score entry score	0.6022	0.6022	0.0000
male	0.8392	0.8392	0.0000
usmaps	0.2633	0.2633	0.0000
prior service	0.0882	0.0882	0.0000
usmaps & prior service	0.0713	0.0713	0.0000
None of the differences are statistically significant.			

Note: This table presents difference-in-means tests for the match of 840 black cadets with 700 white cadets, and for the match of 771 Hispanic cadets with 705 white cadets. We find no significant differences in the means of the key variables.

Table 3: Attainment and Career Outcomes: Matched Sample

	Graduation	Retention 60 Months	Retention 96 Months	Early Promotion to Major
intercept	0.777	0.590	0.330	0.027
	(0.016)	(0.019)	(0.018)	(0.006)
black	-0.021	-0.011	0.019	-0.002
	(0.022)	(0.025)	(0.024)	(0.008)
N	1,540	1,540	1,540	1,540
intercept	0.779	0.621	0.387	0.036
	(0.016)	(0.018)	(0.018)	(0.007)
Hispanic	-0.023	0.005	-0.015	-0.011
	(0.022)	(0.025)	(0.025)	(0.009)
N	1,476	1,476	1,476	1,476
Standard errors are reported in parentheses.				

Note: This table contains results for the matched samples presented in Table 2. The four outcomes of interest are graduation, retention after 60 months, retention after 96 months and early promotion to major. Overall, we find no significant differences among the cadets by race and ethnicity.

Table 4: Robustness Analysis: Attainment and Career Outcomes

Controlling for Variables Used in Matching				
	Graduation	Retention 60 Months	Retention 96 Months	Early Promotion to Major
black	-0.0169	0.0074	0.0383	0.0001
	(0.0215)	(0.0251)	(0.0241)	(0.0080)
Controlling for Variables Used in Matching and Parental Education				
	Graduation	Retention 60 Months	Retention 96 Months	Early Promotion to Major
black	-0.0171	0.0052	0.0388	0.0005
	(0.0216)	(0.0254)	(0.0245)	(0.0081)
Weighted Least Squares with Matched Variables as Controls				
	Graduation	Retention 60 Months	Retention 96 Months	Early Promotion to Major
black	-0.0201	0.0014	0.0348	0.0013
	(0.0211)	(0.0249)	(0.0240)	(0.0079)
With Rematching				
	Graduation	Retention 60 Months	Retention 96 Months	Early Promotion to Major
black	—	-0.007	0.016	-0.003
		(0.026)	(0.030)	(0.01)
Standard errors are reported in parentheses.				

Note: This table contains additional results for the matched samples presented in Table 2. The first and third panels include as controls the variables used in matching, with two alternative calculations of standard errors: heteroskedasticity robust and weighted least squares respectively. The second panel includes eight indicator variables for parental education. Note that there is no graduation regression in the bottom panel because all students in that subsample graduated from West Point.

Table 5: Achievement Analysis: Subsample of Graduates without Rematching

	OML	Academic GPA	Academic Core GPA	Physical GPA	Leadership GPA
intercept	569.5	2.752	2.621	3.052	2.997
	(10.3)	(0.017)	(0.018)	(0.017)	(0.015)
black	84.6	-0.132	-0.145	-0.120	-0.090
	(13.7)	(0.023)	(0.024)	(0.022)	(0.021)
N	1,179	1,179	1,179	1,179	1,179
intercept	509.8	2.881	2.751	3.019	3.026
	(8.4)	(0.014)	(0.015)	(0.012)	(0.012)
Hispanic	25.8	-0.022	-0.046	-0.015	-0.056
	(21.2)	(0.035)	(0.038)	(0.033)	(0.029)
N	1,132	1,132	1,132	1,132	1,132
Standard errors are reported in parentheses.					

Note: This table contains results for the samples presented in Table 2. The outcomes of interest are the order of merit (OML), as well as academic, physical and leadership cumulative GPA scores at graduation. Overall, we find significant differences among the cadets by race, but no differences by ethnicity.

Table 6: Robustness Analysis: Achievement

Controlling for Variables Used in Matching					
	OML	Academic GPA	Academic Core GPA	Physical GPA	Leadership GPA
black	70.61	-0.103	-0.110	-0.134	-0.072
	(12.25)	(0.019)	(0.020)	(0.020)	(0.021)
Controlling for Variables Used in Matching and Parental Education					
	OML	Academic GPA	Academic Core GPA	Physical GPA	Leadership GPA
black	70.38	-0.103	-0.109	-0.132	-0.072
	(12.36)	(0.019)	(0.020)	(0.021)	(0.021)
Weighted Least Squares with Matched Variables as Controls					
	OML	Academic GPA	Academic Core GPA	Physical GPA	Leadership GPA
black	64.49	-0.090	-0.102	-0.118	-0.073
	(11.92)	(0.019)	(0.020)	(0.020)	(0.020)
With Rematching					
	OML	Academic GPA	Academic Core GPA	Physical GPA	Leadership GPA
black	86.31	-0.118	-0.128	-0.124	-0.081
	(14.32)	(0.024)	(0.024)	(0.022)	(0.022)
Standard errors are reported in parentheses.					

Note: The first three panels of this table use the sample presented in Table 5. The fourth panel contains results for Black and White graduates matched using the three entry examination scores and indicator variables for the following: USMAPS, prior service, the interaction of indicators for USMAPS and prior service, gender, recruited athlete, eight mother and father education variables, and region indicators for eight Census regions denoting cadets' home location. Match requires equal numbers of black and white cadets. There are 553 of each race. The lower number of black graduates than in Table 5 results from missing information about home location for some black cadets.

Table 7: Achievement in Selective Courses

	Cannot Be Tested Out				
	EV 203 Environment Geography	PL 100 Psychology	MS 100 Military Science I	MS 200 Military Science II	MS 400 Military Science IV
black	-0.411	0.011	-0.183	-0.149	-0.066
	(0.057).	(0.058)	(0.058)	(0.059)	(0.062)
N	1,167	1,158	1,167	1,167	1,048
	Rarely Tested Out			Commonly Tested Out	
	MS 300 Military Sciences III	PL 300 Leadership	MA 206 Probability Statistics	SS 202 American Politics	PY 201 Physics II
black	-0.206	-0.185	0.018	-0.161	-0.227
	(0.059)	(0.059)	(0.058)	(0.060)	(0.058)
N	1,152	1,149	1,165	1,126	1,167
	Commonly Tested Out				
	MA 104 Calculus I	SS 307 International Relations	SS 201 Economics	EN 101 English Literature	MA 103 Math Modeling
black	0.055	-0.143	-0.198	-0.237	-0.092
	(0.061)	(0.060)	(0.059)	(0.058)	(0.062)
N	1,095	1,111	1,127	1,155	1,047
Estimated standard errors are reported in parentheses.					

Note: This table compares GPA scores for black and white students from the matched sample. Differences in sample sizes arise from variation in the number of students taking the courses.

Table 8: The Effectiveness of the USMAPS

	Difference Academic Entry Score	Difference Fitness Entry Score	Difference SAT Score	Difference Leadership Entry Score
intercept	0.037	-0.018	50.41	0.025
	(0.002)	(0.004)	(3.55)	(0.003)
black	-0.009	0.028	-11.33	-0.002
	(0.003)	(0.006)	(2.33)	(0.003)
N	630	630	630	630
Standard errors are reported in parentheses.				

Note: This table presents an analysis of differences in score gains at the US Military Academy Preparatory School for a sample of black and white cadets. The key variables of interest are the SAT score and the three entry scores (academic, leadership, and physical fitness).

Table 9: Descriptive Statistics by Gender: Full Sample

	female		male	
Variable	mean	std dev	mean	std dev
academic entry score	612	58	606	58
leadership entry score	628	46	623	50
physical fitness entry score	555	70	551	71
usmaps	0.091	0.288	0.114	0.317
prior service	0.031	0.172	0.072	0.259
usmaps & prior service	0.023	0.150	0.057	0.232
graduate	0.764	0.424	0.816	0.387
retain 60	0.544	0.498	0.641	0.479
retain 96	0.278	0.448	0.395	0.489
promote major	0.033	0.178	0.037	0.189

Note: This table presents descriptive statistics for 11,503 cadets in the 11 cohorts of cadets that enrolled at West Point between 1998 and 2008 by gender.

Table 10: Difference-in-Means Balance Tests: Matched Sample by Gender

Variable	Treatment	Control	Difference
academic entry score	0.6120	0.6118	0.0001
physical fitness entry score	0.5547	0.5550	-0.0004
leadership entry score	0.6238	0.6238	0.0000
usmaps	0.0924	0.0924	0.0000
prior service	0.0290	0.0290	0.0000
usmaps & prior service	0.0214	0.0214	0.0000

Note: This table presents difference-in-means tests for the matched sample of male and female cadets. We find no significant differences in the means of the key variables.

Table 11: Attainment and Career Outcomes

	Graduation	Retention after 60 Months	Retention after 96 Months	Early Promotion to Major
intercept	0.815	0.640	0.394	0.041
	(0.011)	(0.013)	(0.014)	(0.006)
female	-0.047	-0.093	-0.115	-0.009
	(0.016)	(0.019)	(0.018)	(0.007)
N	2,745	2,745	2,745	2,745
Standard errors are reported in parentheses.				

Note: This table contains results for the matched sample of male and female cadets. The four outcomes of interest are graduation, retention after 60 months, retention after 96 months and early promotion to major. Overall, we find significant differences among the cadets by gender in graduation and retention rates.

Table 12: Difference-in-Means Balance Tests: Graduating Sample without Rematching

	Academic Entry Score	Leadership Entry Score	Physical Fitness Entry Score
intercept	0.614	0.624	0.557
	(0.002)	(0.002)	(0.002)
female	0.001	0.001	-0.002
	(0.003)	(0.002)	(0.003)
N	2,168	2,168	2,168
Standard errors are reported in parentheses.			

Note: This table reports some difference-in-means tests for the graduating subsample of female and male cadets. Overall there are no significant differences in entry scores.

Table 13: Achievement Analysis by Gender

	OML	Academic GPA	Academic Core GPA	Physical Fitness GPA	Leadership GPA
intercept	428.1	3.033	2.915	3.082	3.077
	(8.0)	(0.014)	(0.014)	(0.011)	(0.011)
female	-3.00	-0.001	-0.013	0.01	0.018
	(11.2)	(0.019)	(0.020)	(0.017)	(0.015)
N	2,168	2,168	2,168	2,168	2,168
Standard errors are reported in parentheses.					

Note: This table contains results for the matched sample of male and female cadets. The outcomes of interest are the order of merit (OML), as well as academic, physical fitness and leadership cumulative GPA scores at graduation. Overall, we find no significant differences among the cadets by gender.

Table 14: Difference-in-Means Balance Tests: Subsample of Graduates without Rematching

	Academic Entry Score	Leadership Entry Score	Physical Skills Entry Score
intercept	0.560	0.607	0.569
	(0.002)	(0.002)	(0.003)
black	-0.005	-0.001	0.006
	(0.003)	(0.003)	(0.005)
N	1,179	1,179	1,179
intercept	0.589	0.605	0.541
	(0.002)	(0.002)	(0.002)
Hispanic	-0.003	0.006	-0.012
	(0.004)	(0.004)	(0.005)
N	1,132	1,132	1,132
Standard errors are reported in parentheses.			

Note: This table contains additional results for the subset of graduates from the matched samples in Table 2. The top panel has 635 black cadets and 544 white cadets. The bottom panel has 584 Hispanic cadets and 548 white cadets.

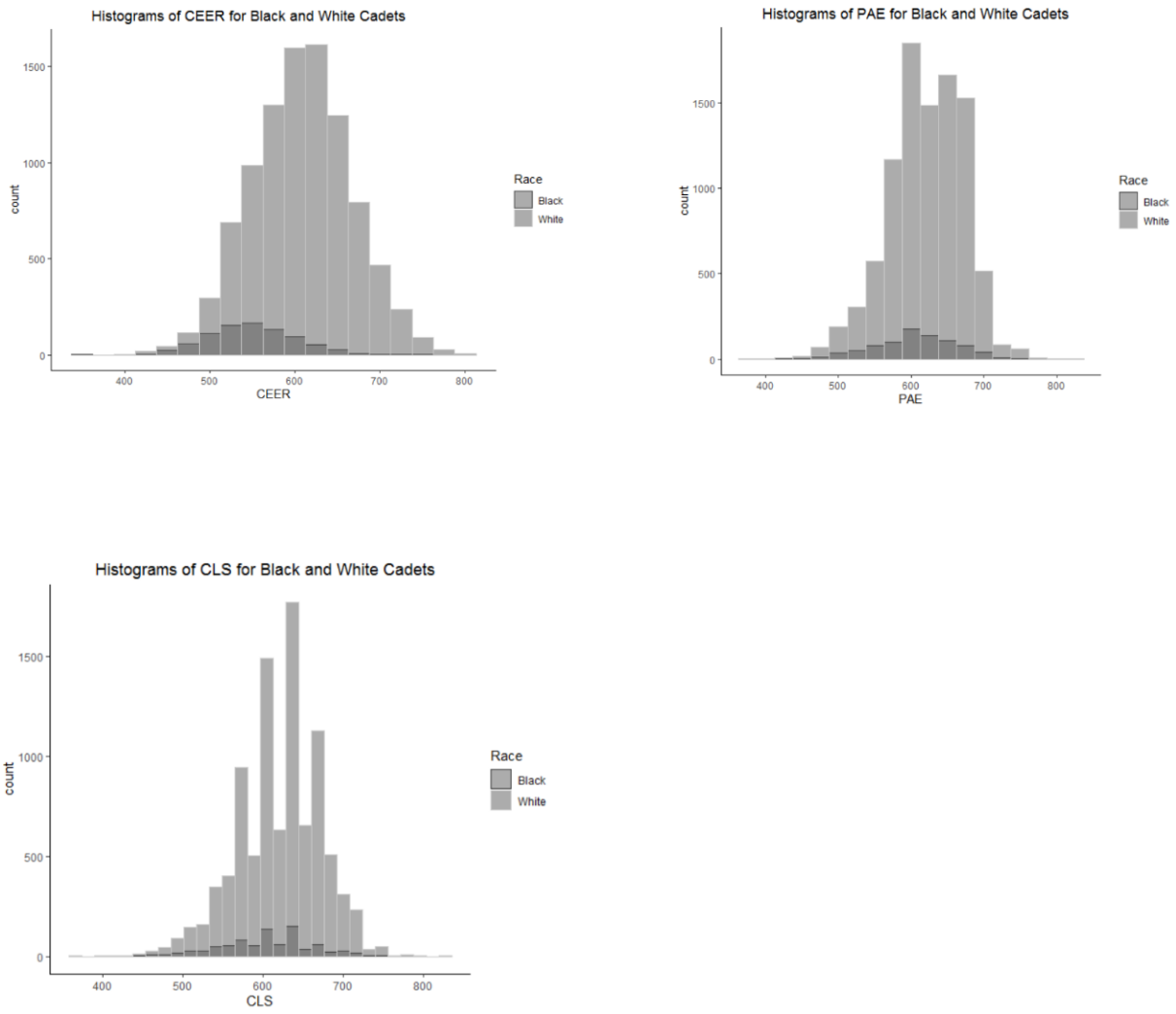
Table 15: Difference-in-Means Balance Tests: Matched USMAPS Sample

Variable	Black	White	Difference
academic entry score	0.493	0.497	-0.004
physical fitness entry score	0.569	0.565	0.004
leadership entry score	0.583	0.594	-0.011
sat score	1059	1078	-19
prior service	0.105	0.156	0.051
male	0.806	0.806	0.00
number of obs	315	315	
None of the differences are statistically significant.			

Note This table presents differences in means tests for the matched black and white cadets who attended the US Military Academy Preparatory School.

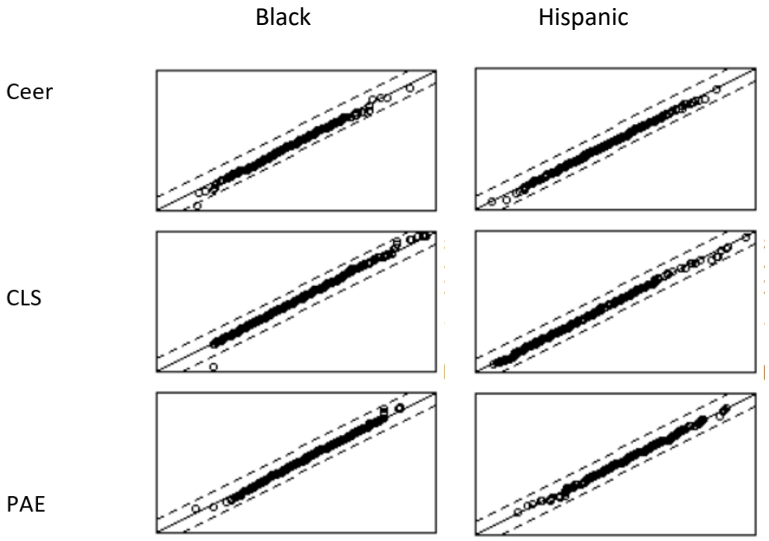
Figures

Figure 1: Histograms



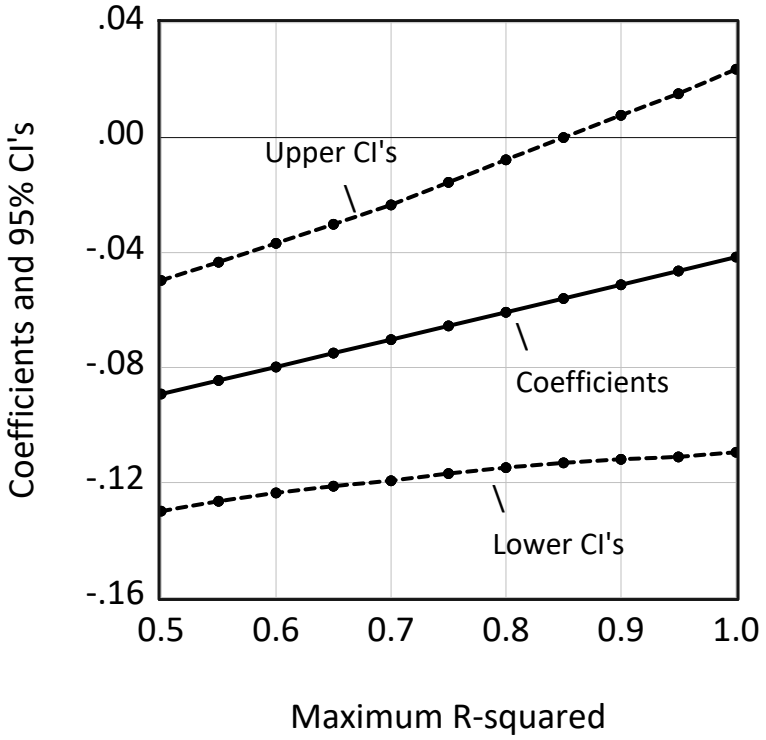
Note: This figure illustrates the histograms of the empirical distributions of key covariates for black and white cadets. We focus on the three entry scores that are used in our matching procedure.

Figure 2: Quantile-Quantile Plots



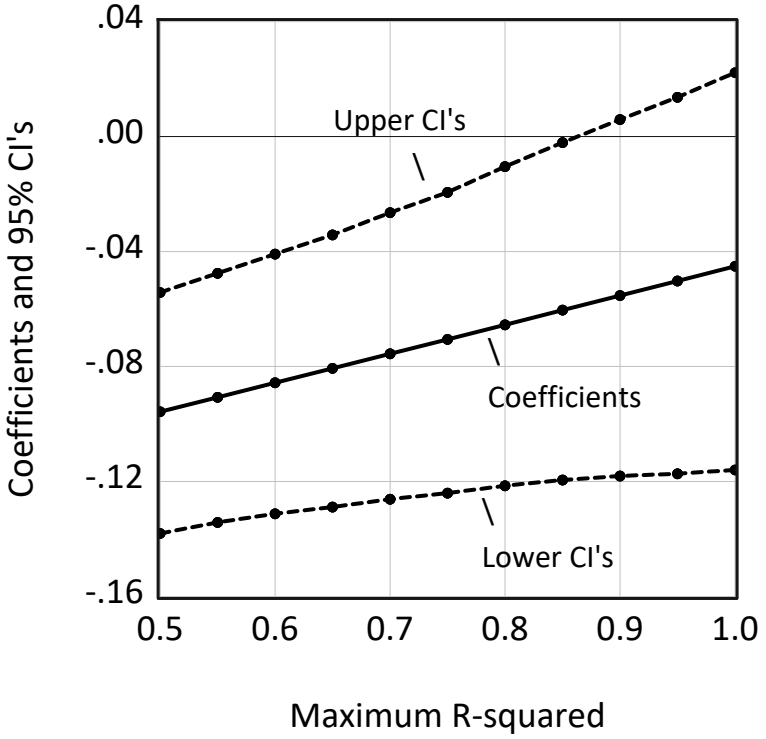
Note: This figure shows the Q-Q plots of the three entry scores that are used to compare the differences in the distributions in our matched samples. Overall, the quality of the match is very high.

Figure 3: Omitted Variable Test: GPA



Note: This figure shows the estimates of the coefficient of black adjusted for bias from omitted variables and the 95% confidence intervals as a function of R^2_{max}.

Figure 4: Omitted Variable Test: Core Courses



Note: This figure shows the estimates of the coefficient of black adjusted for bias from omitted variables and the 95% confidence intervals as a function of R^2_{max}.