Can Colleges Effectively Educate 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 Donna Gilleskie, David Lyle, Andy Postlewaite, Paul Oyer, Petra Todd, Lesley Turner as well as seminar participants at numerous conferences for comments and discussions. Financial support for this project was provided by the NSF grant SES-1658746.

Abstract:

The growing importance of higher education for economic success coupled with rapidly rising tuitions are bringing increasing pressure on colleges to undertake assessments of their effectiveness in educating a diverse student body. Supreme Court decisions have mandated that college actions to promote diversity must be based on clear, measurable, non-discriminatory criteria. These pressures highlight the need for colleges to undertake a systematic collection of data on students' entering credentials, performance in the college, and placement. An assessment strategy is then needed to compare outcomes of majority and minority students with comparable entering qualifications. The proviso"comparable entering qualifications" is of particular importance, given differences in college readiness between minority and majority students. We develop and implement such an assessment strategy for West Point for comparison of outcomes of majority and minority students. The study of West Point is of interest in its own right, but we believe our contribution is of much broader interest. The database maintained by West Point is a model that other colleges and universities might seek to emulate. Our approach, using matching estimators, can fruitfully be employed by other colleges and universities to assess their performance. We find that minority students at West Point have similar graduation rates as their matched white counterparts but black students have significantly lower achievement scores. Despite the difference in achievement scores, we find no difference in early career outcomes between majority and minority students. Encouragingly, we find that the one-year remedial program provided by the West Point preparatory school substantially improves college readiness for minority students.

KEYWORDS: achievement, attainment, career, racial gap, college readiness, preparatory school.

1 Introduction

The growing importance of higher education for economic success has brought increasing pressure on colleges 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.* " An assessment strategy is, therefore, needed to compare outcomes of majority and minority students with "comparable entering qualifications," given the large differences in college readiness between minority and majority students. The objective of this paper is to provide an empirical framework for such an assessment strategy. Our innovative approach uses matching estimators which are well-suited to assess differences in achievement and attainment by race or ethnicity. We then implement our assessment strategy for the U.S. Military Academy at West Point. Finally, we discuss how to extend our approach so that it can be fruitfully employed by colleges and universities to assess whether they are achieving their state goals of educating a diverse student body.

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. This focus is being stimulated by two forces that have shaped the market of higher education in recent decades. One is stiffening public resistance to the rapid rise in the cost of education. The other 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.

It is of much interest to consider what would be required if a selective college or a professional school within a research university wished to undertake an assessment of outcomes of its students by race and ethnicity. To develop such an assessment strategy, we need to have a clear sense of the college's mission. To accomplish its mission a college typically identifies a variety of academic social, leadership, and other skills that it tries to foster among its students. From a research perspective, determining the effectiveness of a college then narrows to the problem of measuring the effectiveness of skill acquisitions by the enrolled students.

Having defined its objectives, including the key skills and capabilities that it wishes its students to acquire, the college needs to design and collect measures of college readiness of students and outcomes realized for those students. For applicants, colleges currently obtain measures of college readiness, including SAT verbal and math scores. Colleges also typically gather information about family income, and about high school activities of applicants including leadership roles, community engagement, and participation in athletics. Measuring outcomes at graduation is more challenging because of differences across academic majors, but no less important. Graduation rates are an important and readily available metric. Student performance in core courses, as well as overall GPA, are important metrics that can be used in combination with information about the choice of major.

Ideally, for evaluating their effectiveness in educating a diverse student both, colleges should assess post-college outcomes, whether labor market placement, attendance at graduate school, or other pursuits. Colleges typically obtain such information for some but not all graduates. Obtaining such information for all graduates could well be a costly undertaking. It is important to recognize that, for assessing effectiveness in educating a diverse student body, it is not necessary to obtain information for all graduates. The matching approach that we advocate and implement requires information only for matched students. Moreover, matching is done utilizing information obtained at the time students enter college. For our West Point analysis, matching leads us to analyze the academic and career outcomes of roughly 30% of entering students. Hence, matching serves to identify the set of students for which such detailed outcomes are needed and thereby greatly reduces the cost of obtaining such information.

We develop and implement an approach that we believe can be applied by many other academic institutions, including many undergraduate institutions as well as graduate professional programs. A 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 US 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 our empirical analysis. While matching estimators are commonly used in program evaluation, there are much more rarely used in assessing differences in achievement and attainment by race, gender or ethnicity.[1] 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 (Rubin, 1973,1974).

As discussed in detail by Diamond and Sekhon (2013) and Imbens (2014), matching requires that important criteria be met. First, matching requires that there be a sufficiently large overlap in the distribution of covariates of the two types being matched, otherwise, the "region of common support" assumption would be violated. Fortunately, our sample is quite large, and, more importantly, the empirical distributions of each of the three main covariates of interest have a common support. Hence, the common-support assumption is well satisfied by our data. In matching, the objective is to achieve full covariate balance. For example, in comparison of outcomes for black and white students, our objective is to choose a matched white sample such that the distribution of covariates used in matching is the same as the distribution for the black students.[2] Finding a match that achieves the best covariance balance obtainable in a given application requires the use of an algorithm that goes beyond

[1] 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).

[2] Rosenbaum and Rubin (1983) establish that a correctly specified propensity score will asymptotically achieve covariate balance. The correct propensity score is, of course, unknown. An incorrectly specified propensity score need not achieve covariate balance, and the attendant results of the analysis may then be invalid.

matching of a propensity score. Several algorithms have been developed.

Matching methods by themselves are not methods of estimation. Every use of matching in the literature involves an estimation step following the matching procedure. We follow the common approach and use simple regressions in the second stage of the analysis. Using the black-white matched sample, we regress the outcomes of interest on an indicator for race. We do the same for Hispanic-white comparisons. All reported regressions coefficients are accompanied by heteroskedasticity-robust standard errors.

We then implement our assessment strategy using data from West Point. Given the range of skills required for a military officer to be effective, the challenge of measuring the effectiveness of the education is arguably more daunting for service academies than for other academic institutions. Because of clarity about their mission and its importance to the nation, service academies are also arguably ahead of their civilian counterparts in focusing on the assessment of the extent to which they are accomplishing their objectives. As a consequence, the U.S. Military Academy at West Point provides an ideal research setting to develop and implement our assessment strategy.

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. [3] 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, QQ plots are particularly useful. Using these criteria, we find that the generic matching algorithm delivers very close matches for all matching variables for all subsamples of interest.

We find small, insignificant differences in graduation rates between black and white students and between Hispanic and white students. For measures of career outcomes, including retention in the Army and early promotion, we find similarly small and insignificant black-white and Hispanic-white differences. In short, we find that matched majority and minority

[3] See Ho, Imai, King, and Stuart (2007, 2011) for detailed discussion.

cadets are equally likely to graduate and have a comparable performance on all career outcome measures.

We also study 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 is in our matched samples. This finding holds for broad measures of academic achievement including the position on the order of merit list, graduating GPA, and GPA in core courses. We find that measures of parental education do not affect these findings.

We also investigate the West Point preparatory school, which provides 10 months of preparatory education. We find that this preparatory school significantly enhances students' college readiness, with somewhat smaller gains for black students than their matched white counterparts. The results concerning achievement establish that there are substantial and significant gains for both black and white students, but smaller gains for black students than their matched white counterparts. That said, the finding of equal career success of matched black and white students bears emphasis. Career outcomes are arguably more important than college achievement measures.

The rest of the paper is organized as follows. Section 2 provides a brief literature review. Section 3 introduces our data set and provides some institutional background that is important to understand the variables used in our analysis. Section 4 presents our analysis of attainment, retention, and early promotion. Section 5 provides our achievement analysis. Finally, Section 6 discusses the policy implications and offers some conclusions.

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. White-black high school graduation rates actually 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; Murnane and Hoffman, 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 (2014) 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. By their estimates, all of the wage gap of so 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), better access to health care (Chay, Guryan and Mazumder, 2009), 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 an admirable summary of this body of research as well as 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 good STEM programs. Some research in higher education has focused on minority participation and graduation in STEM majors. 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 actually 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, AAH below) estimate a discrete choice model of school, major, and persistence-to-graduation using late 1990's data from California's UC-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.[4] 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 the same scale) in the two highest-ranked schools were not drastically different, respectively 52.1 percent and 58.1 percent, but these values dropping 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 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 most institutions, data for multiple cohorts will be required to obtain

[4] 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.

adequate sample sizes. Our analysis for West Point utilizes data for 11 cohorts. Another alternative is assembling of data across colleges. Challenges then arise in addressing both sorting across colleges and heterogeneity of student backgrounds and student choices within and across colleges. These challenges are not insurmountable, and, as our literature review shows, much has been learned by analyses of cross-college data. Our approach can also be applied to large graduate professional programs and seems particularly well suited for large MBA programs.

Our work complements AAH (2016), who model 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 and gender 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. [5]

Finally, our paper is related to a 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

[5] As we have alluded to and is obvious, it is intuitive that the persistence to graduation at West Point relative to persistence to graduation in STEM at most other universities would be higher because cadets cannot switch to a major that might be easier to complete. It is of interest to research more generally, whether a variety of majors (and ease of switching) reduces very significantly STEM degree persistence.

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 rankings, West Point is ranked number 21 among National Liberal Arts Colleges and number 2 among Top Public Colleges.[6] While West Points has some idiosyncratic features that simplify such an assessment strategy, some extremely valuable lessons can be drawn from our analysis.

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 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, determined by USMA, 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 (CFA). The community leader score (CLS) score is the sum of the first three of the preceding. We observe all these skill measures.

In addition, we observe a variety of student characteristics such as prior-service, attendance at USMAPS, father's and mother's education, as well as the cohort and state of residence of the student. For expositional convenience, we refer to these prior student characteristics as demographics with the understanding that race and gender are not encompassed by this shorthand. The USMAPS primarily serves students who are recruited as athletes and students with prior Army service. We, therefore, also examine interactions between these

[6] The rankings are published at https://www.usnews.com/best-colleges/west-point-2893/overall-rankings.

variables. Note that there is no separate application for USMAPS. 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.

There are several outcomes of interest including college attainment and the subsequent 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. Below the zone promotion is the expression used to denote early promotion (see the previous footnote). We study each outcome separately below.

We also observe several achievement measures for those cadets that 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, physical capabilities, and leadership potential, 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 not only prestigious, but also establishes the order in which candidates choose among the 16 military branches, and hence determines which candidates obtain 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 the 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, and 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.

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

	Variable	Black	Hispanic	White
	academic score	0.551	0.585	0.607
initial	leadership score	0.611	0.602	0.620
skills	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
	promote major	0.026	0.025	0.037
number of observations		840	771	9,892

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 it 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 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 of our analysis is to assess whether attainment and career outcomes of minority students are equivalent to those of majority students with comparable measured entry capabilities. Our main focus is on black-white as well as Hispanic-white comparisons.

We first report our findings that compare black and white cadets. As noted above, we have data for 840 black cadets and 9,892 white cadets. Assessing the effectiveness of West Point training by race and ethnicity requires comparing outcomes of cadets who have comparable skills upon entry. 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 overlap 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.[7]

For each black cadet, the matching algorithm searches for a white cadet with closely matched entering credentials. We match cadets based on 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 as well prior active service, attendance at USMAPS, and both prior service and attendance at USMAPS. We, therefore, restrict attention to those variables that are used by West Point for admission decisions. This 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 there are any systematic differences by race or gender once one controls for the relevant variables that are used in admission.

To assess the quality of the matching algorithm, we begin by comparing the means of 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 a standard difference-in-means test. It reveals that the means in both subsamples match up quite well for all of the variables used in the analysis.

[7] 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.

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

Variable	Black	White	Difference
prior academic score	0.5510	0.5508	0.0003
prior physical fitness	0.5753	0.5742	0.0011
prior leadership 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 score	0.5852	0.5852	0.0000
physical fitness	0.5373	0.5373	0.0000
leadership 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	0.0713	0.0713	0.0000

None of the differences are statistically significant.

We next compare the distributions of the three continuous entry score variables for the matched sample. This is done in the left panel of Figure 2. Each graph in Figure 2 is a quantile-quantile plot. For example, the 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.[8]

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

[8] Looking more closely, we see that the upper-left graph in Figure 2 shows that there is one black cadet with a very low CEER score- noticeably below the 45-degree line. The lower-left shows that there is also a black cadet with a CLS score noticeably below the 45-degree line. To investigate robustness, we repeated the analysis without these two observations. We obtained virtually the same results for all of the comparisons reported in the tables below. Not surprisingly, these 2 out of 840 observations have a negligible effect on our findings.

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 error are reported in parentheses.

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 results of four regressions for our matched sample of black and white cadets. In each of these regressions, the dependent 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.

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 for 5 years and roughly one third are retained for 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 a variety of robustness checks. 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. Overall, we find the coefficients of black continue to be negligible in magnitude and statistically insignificant. Next, we added some additional variables that we observe and did not include in the matching analysis. Here we focus on variables that capture parent education levels. We omitted these variables from our initial analysis since West Point does not use this demographic information in the admission process. Regression results for binary variables with match variables and parent education variables included show that the estimated black-white differences continue to be small and insignificant. The inclusion of these parental education variables has no consequential effect on any of the other variables. We thus conclude that our main findings are robust to a variety of changes in the specification of the second-stage regression model.

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 middle panel of Figure 2 also indicates that the quality of the match is very good. Hence, we 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 conclude that rates of graduation, retention, and early promotion rates are very similar for comparable Hispanic and white cadets.

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 differences in 5-year and 8-year retention rates and rates of early promotion to major are quantitatively small and statistically insignificant. Finally, there are no systematic attainment or career gaps between comparable Hispanic and white cadets.

5 Achievement and College Readiness

Thus far, we have compared outcomes for the matched sub-samples of admitted cadets. Since we can only measure achievement for those cadets that graduate, we now restrict our attention to matched subsamples of graduating cadets. Note that we do not rematch our samples to obtain a good match. However, the results reported in Table 4 suggest that we can also compare outcomes for the subset of these matched cadets who graduated without rematching. We, therefore, continue our analysis by focusing on differences in achievement among the graduating cadets for which we can measure achievement. The empirical findings are summarized in Table 5.

Table 4: Difference-in-Means Balance Tests: Subsample of Graduates without Rematching

	Academic Score	Leadership Score	Physical Skills
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 error are reported in parentheses.

Our first measure of achievement is the position on the Order of Merit List which is basically 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, translating to a roughly 9 percentage point difference in OML. Next, we focus on academic, physical, and leadership measures. The 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, academic scores in core common areas, leadership scores, and physical scores than their matched white counterparts. We thus conclude that black cadets have significantly lower achievement measured by cumulative GPA scores at graduation and significantly less favorable positions on the order of merit list than comparable white cadets.

Table 5: Achievement Analysis: Subsample of Graduates without Rematching

	OML	Academic	Academic Core	Physical	Leadership
		GPA	GPA	GPA	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 error are reported in parentheses.

Again we also conducted a variety of robustness checks including the variables used in matching as well as parental background variables. Overall, our main conclusions are unchanged. In particular, we find that the parental variables have negligible effects on the coefficient of black.

We next turn to the results for Hispanic and white cadets. Table 4 also shows that the differences in entering academic and leadership scores of those who graduated are also 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. 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.

From a broader policy perspective, we would like to know what colleges can do to close the racial 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. The preparatory school primarily serves minority students, students that are recruited as athletes, and students with prior Army service in the enlisted ranks. Here we focus on the black-white comparison. Table 6 shows that again matching is a promising approach and we can construct matched subsamples with virtually identical observed characteristics.

Table 6: Difference-in-Means Balance Tests: Matched USMAPS Sample

Variable	Black	White	Difference
previous academic	0.493	0.492	0.001
previous physical	0.572	0.573	-0.001
previous leadership	0.584	0.583	0.001
prior sat score	1059.56	1061.53	-1.96
prior service	0.095	0.095	0.00
male	0.792	0.792	0.00
number of obs	346	245	

None of the differences are statistically significant.

Next, we analyze the gains in college readiness scores focusing]on academic, leadership, and physical scores. One nice feature of the analysis is that we have separate measures at the beginning and the end of the preparatory school. Hence we can difference the scores and compute the gains for each student. Table 7 summarizes the empirical results for our analysis of the gains.

Table 7: The Effectiveness of the USMAPS

	Difference Academic Score	Difference Physical Fitness	Difference SAT Score	Difference Leadership Score
intercept	0.034	-0.014	50.69	0.027
	(0.002)	(0.005)	(3.90)	(0.003)
black	-0.007	0.019	-12.19	-0.005
	(0.003)	(0.006)	(4.96)	(0.004)
N	591	591	591	532
Standard error are reported in parentheses.				

Our analysis of the data reveals that cadets who attended West Point, USMAPS significantly improved their academic and leadership skills during that year. The gains are relatively large and reflect 5 to 6 percentage point improvement in academic and leadership scores. Academic and leadership gains are somewhat smaller for black students than white students, but still substantial. Black students also have larger gains in physical fitness than comparable white students. 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.

6 Conclusions

There are large initial differences in college readiness among enrolling cadets at West Point. In particular, minority students have, on average, significantly lower academic and leadership scores than majority students. Of course, the same is true for most selective colleges and universities in the US. We develop an approach for making meaningful comparisons of outcomes across demographic groups, employing matching estimators. We have a data set well suited to assessing the viability of our approach. Utilizing these data, we obtain exceptionally good matches for black-white and Hispanic-white comparisons. These well-balanced samples in turn permit precise comparisons of treatment effects of college attendance by race and ethnicity.

We find small, insignificant differences in graduation rates between black and white students and between Hispanic and white students. For retention and early promotion to major, we find similarly small and insignificant black-white and Hispanic-white differences. We also studied achievement among the subsample of cadets who graduate from West Point. Our analysis finds that there are significant black-white achievement gaps in college. This finding holds for broad measures of academic achievement including position on the order of merit list, graduating GPA, and GPA in core courses. These findings contrast with the findings on attainment, retention in the military following graduation, and early promotion to the rank of Major. We find no Hispanic-white achievement, attainment, or career gaps.

As we noted at the outset, a fundamental challenge for West Point and other 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 capabilities required for admission to West Point. Selective colleges and universities can potentially benefit from the experiences of USMAPS 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 not clear whether highly selective colleges can close these preexisting gaps without offering a more structured and personalized preparatory learning experience that is similar to the one provided by USMAPS. A collaborative effort of selective colleges might to develop such a preparatory program might merit consideration.

A broader implication of this work goes beyond the educational topics that we discussed in this paper. Intuitively, the U.S. Army should have an officer corps that reflects the underlying population of America or the enlisted force to be most effective, achieve its objectives, and promote its legitimacy. These goals are harder to attain if there are significant and unattended achievement gaps within certain groups.

In our introductory discussion, we outlined the ways in which colleges and professional schools might employ the approach we have developed. As we noted there, the data for which colleges will likely have incomplete information regards performance after graduation (e.g., earnings, attending graduate school). Obtaining such information for all graduates could well be a daunting and costly undertaking. Hence, as we noted in the introduction, it is very important to recognize that, for evaluating racial gaps, it is not necessary to gather information for all graduates. Consider assessing the black-white gap. With data obtained at the time students enter the college (e.g., SAT verbal and math), matching can be used to determine which entering white students are matched to the entering black students. Then a survey of graduates would only require information about black students and their matched white counterparts. Given the relatively low representation of minority students in most selective colleges, this would appear to be a manageable undertaking. Of course, information for multiple cohorts will likely be needed, and that will take time. In the interim, colleges can undertake, with data they have in hand, an assessment of their effectiveness with respect to achievement and graduation in educating a diverse student.

References

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

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

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

Arcidiacono, P. and Lovenheim, M., "Affirmative Action and the Quality-Fit Trade Off," Journal of Economic Literature, Vol. 54(1), March 2016, 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, Marianne, Goldin, Claudia, and Katz, Lawrence, "Dynamics of the Gender Gap for Young Professionals in the Financial and Corporate Sectors," American Economic Journal: Applied Economics, Vol. 2, July 2010, 228-255.

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

Black, Dan, Haviland, Amelia, Sanders, Seth, and Taylor, Lowell, "Why Do Minority Men

Earn Less? A Study of Wage Differentials among the Highly Educated,"The Review of Economics and Statistics, Vol. 88 (2), May 2006, 300-313.

Black, Dan, Haviland, Amelia, Sanders, Seth, and Taylor, Lowell, "Gender Disparities among the Highly Educated," Journal of Human Resources, Vol. 43 (3), Summer 2008, 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.

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, Scott, Marianne Page, and James West, "Sex and Science: How Professor Gender Perpetuates the Gender Gap,"The Quarterly Journal of Economics, Vol. 125(3), August 2010, 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, forthcoming.

Chay, K., Guryan, J., and B. Mazumder, "Birth Cohort and the Black-White Achievement Gap: The Roles of Access and Health Soon after Birth," NBER working paper 15078 (2009).

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, Michael D., and William N. Evans (2000). “Families or Schools? Explaining the Convergence in White and Black Academic Performance,” Journal of Labor Economics 18, 729-754.

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

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

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

Fryer, Roland and Levitt, Steven, “An Empirical Analysis of the Gender Gap in Mathematics,” American Economic Journal: Applied Economics, Vol. 2 (2), April 2010, 210 -240.

Griffith, Amanda, “Persistence of Women and Minorities in STEM Field Majors: Is it the School that Matters?” Economics of Education Review, Vol. 29, 2010, 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., H. Ichimura, J. Smith and P. Todd (1998). “Characterizing Selection Bias using Experimental Data.” Econometrica, 66 (2), 315-331.

Heckman, James, Hidehiko Ichimura and Petra Todd (1997). “Matching As An Econometric Evaluation Estimator," Review of Economic Studies, 65(2), 261-294.

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

Imbens, Guido W., "Matching Methods in Practice: Three Examples," J. Human Resources Spring 2015 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, Richard J., “U.S. High School Graduation Rates: Patterns and Explanations,” Journal of Economic Literature, Vol. 51(2), June 2013, 370-422.

Murnane, Richard J. and Hoffman, Stephen, “Graduations on the Rise,” Education Next, Fall 2013.

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 Johnson, W.R., “The Role of Pre-Market Factors in Black-White Wage Differences,” Journal of Political Economy, Vol. 104, 1996, 869-895.

Neal, D. and A. Rick (2014). “The Prison Boom and the Lack of Progress after Smith and Welsh.” NBER Working Paper 20283.

Rajeev H. Dehejia and Sadek Wahba (1999).“ Causal Effects in Non-experimental Studies: Reevaluating the Evaluation of Training Programs,” Journal of the American Statistical Association, 94:448, 1053-1062.

Rajeev H. Dehejia and Sadek Wahba (2002). “Propensity Score-Matching Methods for Nonexperimental Causal Studies,” Review of Economics and Statistics, February 2002, 84 (1), 151-61.

Rosenbaum, Paul and Donald 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.

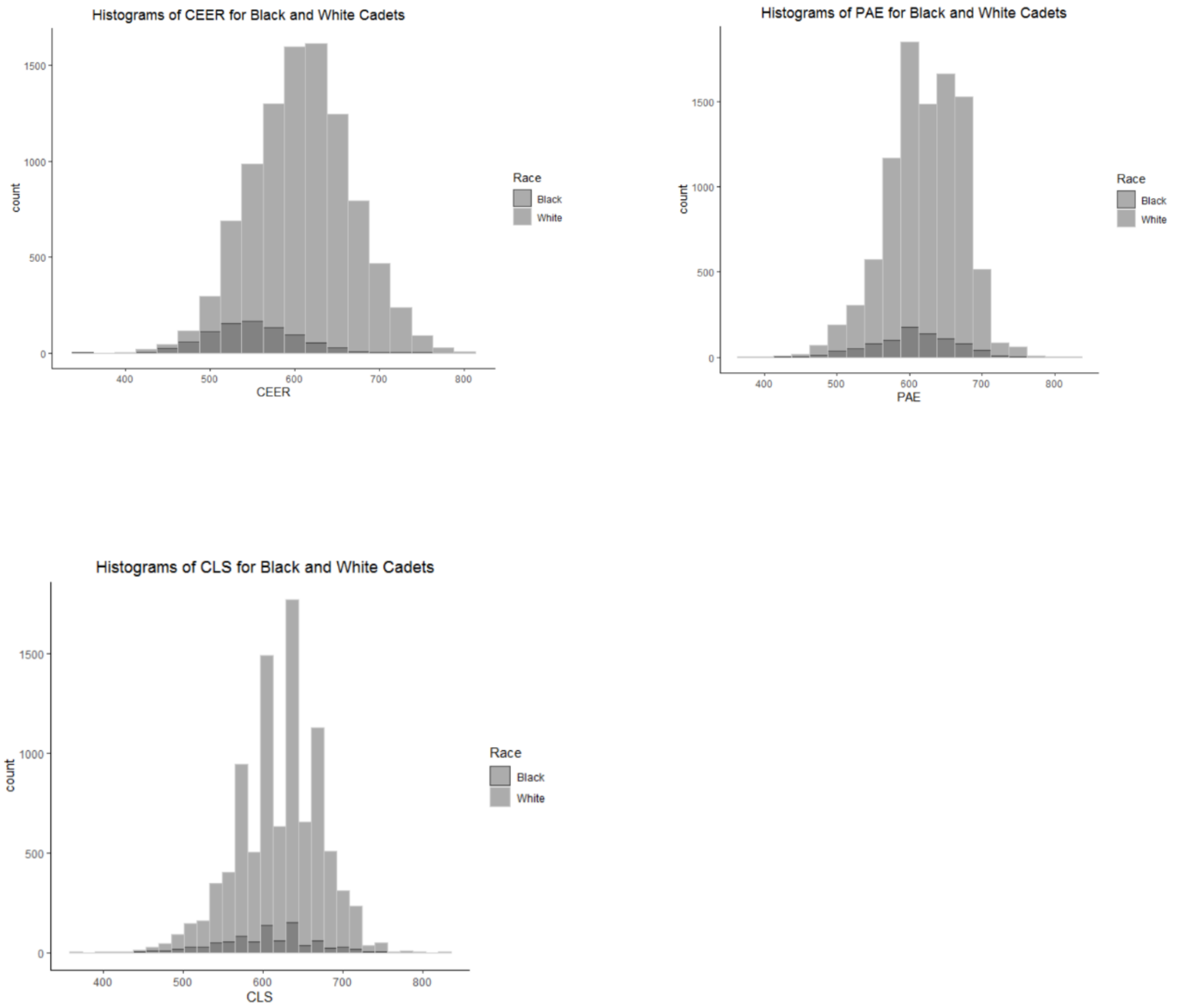
Rubin, D. (1974). “Estimating Causal Effects of Treatments in Randomized and Non- Randomized Studies,” Journal of Educational Psychology, 66 (5), 688-701.

Smith, J. P., and F. R. 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.

A Figures

Figure 1: Histograms

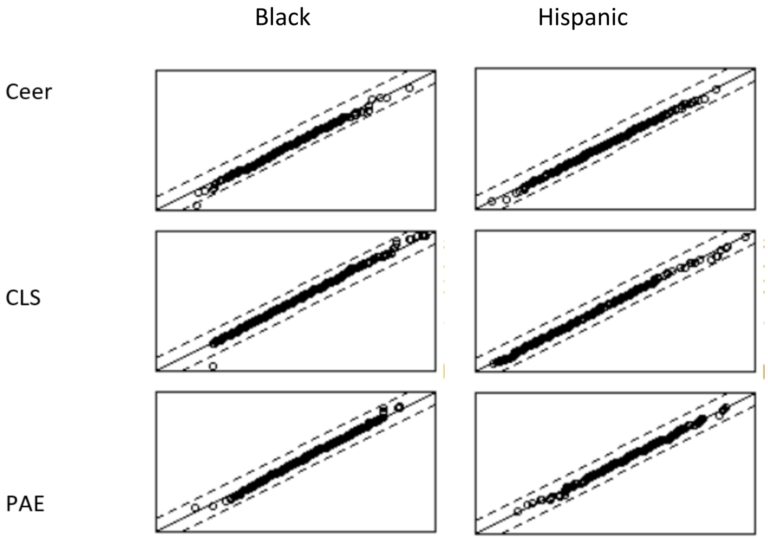


Histograms of CEER for Black and White Cadets

Histograms of PAE for Black and White Cadets

Histograms of CLS for Black and White Cadets

Figure 2: QQ Plots


Black
Hispanic
Ceer
CLS
PAE