Research Article

Does Teacher Feedback Potentiate a Positive Feedback Loop That Underlies the "Achievement Gap"? The Role of Race and Gender in the Interrelations Between Self-Perceptions, Teacher Feedback, and Academic Performance at West Point

Hillary Schaefer[1]

1. Independent researcher

Dynamic relationships among teacher or supervisor judgments, self-beliefs, and performance are important, especially for Black and female students in White, male majority settings. We examined temporal interrelationships among GPA, self and instructor ratings of classroom competence among military cadets at the United States Military Academy (*n*=8,612). Mixed models run on samples matched on SAT score revealed Black-White differences in GPA, self and instructor ratings, and gender differences in self-rating. Samples matched on freshman GPA revealed Black-White differences in final GPA and instructor ratings. A random intercept cross-lagged panel model, which accounts for between-person average levels of each variable, showed that instructor ratings of classroom competence had a stronger effect on future GPA for Black versus White students, illuminating one mechanism underlying racial disparities in final GPA. This finding did not hold for instructor ratings of social skill, showing the specificity of the effect. Self-assessment of competence solidified in Year 2, whereas GPA and instructor ratings influenced each other over time in a positive feedback loop. Results are discussed in regard to the experiences of Black cadets and impact of superior Officers and supervisors.

Correspondence: papers@team.qeios.com — Qeios will forward to the authors

 doi.org/10.32388/KUANCU

1. Introduction

To truly impact socioeconomic outcomes for underrepresented groups, we must deeply interrogate systems, like higher education, the military, and even psychological science, for actionable reforms[1][2]. Considerable work has investigated the so-called "achievement gap," perhaps more aptly named "educational debt," the difference between Whites and non-Whites in educational/occupational outcomes, due to a mix of systemic factors such as inequalities in opportunity and resources, generations of economic suppression, implicit biases and so on[3][4]. Despite explicit and implicit racism (among other biases) *decreasing* since the mid-2000s (Charlesworth & Banaji, 2022), the achievement gap has continued to, if anything, widen (Hanushek et al., 2019) and likely was worsened by the COVID-19 pandemic, mostly due to socioeconomic factors[5]. In the military, poor racial climate and discrepancies in Black Officer promotion and female retention rates have been consistently noted, despite a wider cultural view that the military is a meritocracy with better race relations than the wider public; further, the most recent report on diversity among Officers noted that General Officers were 94% White and male[6][7]. The current study utilized longitudinal data from the United States Military Academy (USMA) to better understand such gaps in occupational outcomes among military Officers by integrating self-efficacy theory with stereotype threat to understand the integration of feedback, self-beliefs, and performance over time.

The United States Military Academy (USMA) is a collegiate institution with a comprehensive character focus emphasizing, "empathy, respect, and humility that enables an individual to treat others with dignity"[8]. However, USMA is *also* an institution with a White, male majority, robust history of excluding those who are not White men, and has perpetuated a tradition of hegemonic White masculine culture in which women and non-White men face considerable difficulties[9][10][11]. Evidence suggests sexism is worse at military academies than other collegiate environments, and USMA has faced consistent criticism regarding racial issues[9][12][13]. Importantly, USMA has a relatively fixed educational experience[1], multi-rater evaluation system regularly completed by instructors, supervising Officers, and students (self-rating), and clear metrics of success (e.g., cadets are ranked according to 55-30-15 mix of academic, military, and physical performance that determines access to occupational tracks; [14]). The present study utilized this system to delineate longitudinal relationships between GPA, self and instructor ratings that directly relate to occupational outcomes, more precisely than possible in other systems[15]. Enumerating these relationships within USMA may have relevance to other professional and academic evaluations, Officership in general, and inequities in other historically White male settings.

1.1. Metacognitive self-beliefs and achievement

Self-efficacy theory has driven work connecting performance to metacognitive self-beliefs, which are adversely affected by stereotypes and lived experiences of bias for marginalized groups (e.g., [16]). Self-beliefs create confidence (or doubt) that one can be successful, based on attitudes, self-assessments, and experiences[17][18]; academic metacognitive beliefs range from global competencies to judgements of task performance. Generalized self-efficacy, or beliefs about one's capacity to achieve goals and be successful, positively relates to academic performance and professional outcomes[19], but also correlates with other stable, positive attributes including self-esteem and conscientiousness, any one of which often fails to be uniquely predictive or predictive over more context-specific attributes[20]. Global attributes may be less susceptible to influences of feedback and performance than specific self-beliefs, especially for stigmatized students who may discount feedback as stereotype-driven[21]. At the other extreme, judgments on specific tasks illuminate self-appraisal processes; however, they may not generalize across academic context or time[22][23]. Specific domain-level metacognitive judgements, such as USMA's end-of-semester rating system, form an ecologically useful middle ground to evaluate self-assessments that capture information beyond generic positivity, and may have more interrelatedness with performance as they are dynamically calibrated[24][25].

The calibration of metacognitive self-assessments for marginalized groups must consider experiences of bias and stereotyping. The overarching effect can be protective, limiting, or unstable. Positive and nurturing relationships – among family, but also with teachers and peers – may be protective forces and shape self-beliefs[26]. Black students may resist stereotype internalization through leadership and connection with like peers, and a more positive self-concept that can buffer against microaggressions, stereotype-driven feedback, and other experiences of bias[27][28][29]. However, reduced domain-specific self-efficacy and inaccurate or unstable self-assessment has also been found for Black students[16][30]. Relevant too is work on stereotype threat, which demonstrates how stereotypes can influence performance and overarching self-beliefs[31][32]; especially for those with internalized beliefs about their group, exposure to stereotype threat drags down self-beliefs and performance[33][34].

For collegiate women, lower self-beliefs about professional competence limit performance and retention in STEM[35], which may have relevance to the counterstereotypical environment of the military. Additionally, a large literature shows that women judge themselves more harshly than men, which holds over the life span and different levels of self-judgements[36][37]. This paints a conflicting picture as to

how domain-level self-assessments might be calibrated by performance and feedback, especially for Black women. Assessing whether the relationship between self-beliefs and performance is different for Black versus White, and male versus female students is crucial, especially in settings such as USMA where masculine stereotypes and Whiteness are salient, and educational and professional implications are inextricably linked[38][39].

1.2. Teacher feedback and perceptions

Teacher perceptions of students, including judgements of performance and general capacities, are important; teacher judgements influence future academic self-concept and performance, when feedback is given to students and when it is not (e.g., [40][41]). Teacher judgements or expectations (often the two are confounded) can create self-fulfilling prophecies in terms of student performance and academic self-concept, and influence student-teacher dynamics[42]. Anti-Black prejudice has been demonstrated in terms of teacher feedback and confirmation bias that focuses attention on stereotype-consistent behaviors, as well as overall beliefs that some students' abilities are fixed[43][44][45]. Positivity bias, or higher praise for subpar work presumably completed by Black students, has also been found[46][47]. For girls and women, teacher biases may follow gender stereotypes, with a positive bias toward women in subjects such as reading, and negative bias in math (e.g., [48]). Importantly, given USMA's original role as an engineering school, and general science focus that awards all graduates with a Bachelor's of Science, a large literature describes anti-female bias in collegiate STEM[49][14].

Teacher perception is clearly influential for student outcomes, but how does the addition of direct feedback alter the importance of teacher judgements? In-class feedback can significantly impact student self-efficacy and performance[50]. Conversely, written feedback can be discounted, especially in college, as students attend primarily to grades or scores[51]. There is limited work on summary feedback similar to USMA's system; one study demonstrated that student ability level may be influential, such that lower-performing students benefit most from performance summaries[52]. USMA's instructor evaluation is accompanied by an in-person discussion. and we expected this system to more likely approximate classroom teacher-student dynamics and summary feedback (possibly influential) than written task feedback (possibly discounted). This standardized system also allowed examination of teacher assessments across the entire collegiate experience at USMA, extending work from younger ages into a college-age population.

1.3. Intersectionality

The collegiate experiences of Black women and Black men can differ in important ways, especially given that Black misandry and misogyny tend to emphasize different stereotypes (e.g., [38]). The military context is also important, with multiple investigations demonstrating pervasive sexist attitudes and increased gender role salience versus civilian settings[12]. Conversely, fewer resources have been devoted to understand systemic racism in the military[7][53]. There is a relative lack of true intersectional work for our key constructs, especially at this developmental level. Self-assessments have clear gender patterns and racial effects, as noted above, although they tend to be studied separately; teacher judgements of children have shown simultaneous race and gender effects, with the former often being more substantial[54]. Education-focused work, especially in college, often focuses on either race or gender as most salient to a given outcome[55]. The present study aimed to address this gap.

1.4. The specifics of the USMA environment

Retaining Cadets of diverse racial backgrounds and training Officers to be inclusive, self-aware leaders has been a stated objective of the wider Army for some time (e.g., [56]). USMA's mission statement outlines the importance of diversity, and importance of leadership of diverse teams, [8]. Equity, especially by race, is particularly important in terms of Cadet training, given that Cadets will soon become platoon leaders of enlisted Soldiers, who are on average more diverse than Officers[6]. Despite this intent, race[7] (Hopkins & Williams, 2013) and gender (Baldwin, 1996) differences in promotion rates have been noted for decades, and problems of sexual assault/harassment remain in the Army generally (Street, Stafford, Mahan, & Hendricks, 2008; Turchik & Wilson, 2010), and USMA specifically (Arbeit, 2016). Over the last few decades, sparse reports have detailed differences in promotion and retention by race and/or gender (e.g., [6][57]), no study to the author's knowledge has quantitatively examined what metrics might be associated with bias, or the mechanisms through which bias is created in the military. Although studying full military career trajectory was outside of the reach of this study, the current study aimed to add to a scant, if not nonexistent, empirical literature about race, and race *and* gender in terms of performance in this U.S. military setting.

Ratings systems within the military, represent the standards and values of the organization and of the preceding leaders who have shaped the instruments; chain-of-command ratings are a primary component of the United States military's promotion system (Moore & Trout, 1978). Originally based on

the rating system used with Noncommissioned Officers, the USMA's rating system contains 23 items relevant to Cadet leadership, performance, social skill, professionalism, and effectiveness (see supplementary Table 1). While both these ratings and GPA measure different aspect of cadet "success" at USMA, GPA importantly is used to calculate class rank which influences branch selection as well as training and other opportunities while at USMA. Unlike the ratings cadets will encounter later as Officers, USMA's rating system is not graded per se, does not factor into class rank, and is intended as a developmental counseling tool. Overall, understanding demographic discrepancies in GPA tests whether bias influences the early trajectories of these young Officers, and builds on a larger literature about bias in college achievement. Conversely, interrogating USMA's rating system for bias is important to understand how these military trainees are perceived, has relevance to the ratings systems used in the greater Army, and also reveals information about the instructor ***raters***, most of whom are Officers themselves who either come in from a longer career and end their service at USMA, or serve a 2-3 year terms at a USMA before returning to a more typical billet in their branch. In the present study, 63% of ratings were given by active duty Officers, 4.5% by active duty Enlisted Soldiers, and 32% by civilians, many of which are retired or previously have served. Finding evidence of race or gender bias at USMA is, therefore, highly suggestive of similar biases in the military at large.

Descriptive Statistics							
		Instructor rating		**Self rating**		**GPA**	
Freshmen	*N*	Mean	SD	Mean	SD	Mean	SD
White men	5842	3.10	0.63	3.25	0.73	3.02	0.65
White women	1375	3.12	0.63	3.01	0.70	3.04	0.64
Black men	1027	2.90	0.58	3.18	0.72	2.37	0.62
Black women	323	2.96	0.59	3.00	0.67	2.59	0.63
Seniors							
White men	5842	3.99	0.59	3.88	0.57	3.16	0.50
White women	1375	4.06	0.57	3.78	0.54	3.19	0.48
Black men	1027	3.77	0.59	3.78	0.56	2.70	0.44
Black women	323	3.89	0.6	3.71	0.57	2.85	0.47
SAT-matched sample							
		Instructor rating		**Self rating**		**GPA**	
Mixed model terms		*F*	*p*	*F*	*p*	*F*	*p*
Time		**1500**	**<.0001**	**660**	**<.0001**	**141**	**<.0001**
Race		**12**	**<.0001**	3.6	.057	**40**	**<.0001**
Gender		**13**	**<.0001**	**36**	**<.0001**	**21**	**<.0001**
Time x race		2.2	.14	.08	.77	1.5	.23
Time x gender		2.3	.13	**7.6**	**.0057**	1.2	.28
Race x gender		.042	.84	1.9	.16	**8.7**	**.0032**
Time x race x gender		.075	.78	.24	.62	.42	.52
Simple effects				$\chi 2$	*p*	$\chi 2$	*p*
M-F difference: freshmen				**50**	**<.0001**		
seniors				4.3	.039		
B-W difference: women						1.3	.26

Descriptive Statistics							
		Instructor rating		**Self rating**		**GPA**	
men						**28**	**<.0001**
Freshman GPA-matched sample							
		Instructor rating		**Self rating**		**GPA**	
Mixed model terms		*F*	*p*	*F*	*p*	*F*	*p*
Freshman year GPA		.28	0.60	**9.9**	**.0017**	**3100**	**<.0001**
Senior year GPA		**78**	**<.0001**	2.3	.13	n/a	n/a
Race		**7.8**	**0.0053**	**7.1**	**.0077**	**22**	**<.0001**
Gender		**12**	**<.0001**	.23	.63	**95**	**<.0001**
Race x Gender		.37	.55	1.8	.19	.18	.67

Table 1. Full sample descriptive statistics and matched sample mixed model results for classroom competence ratings and GPA

Notes: Linear mixed models were run with samples matched by race on SAT or freshman GPA and gender, and were run separately for instructor rating, self-rating, and GPA. Ns for the SAT-matched sample were: 323 Black women, 338 White women, 1065 Black men, 1050 White men; for the freshman-year-GPA-matched sample the same ns were: 305, 287, 805, and 823. Mixed model F-tests had 1,3578 df for instructor rating, 1,3718 for self-rating, and 1,3587 for GPA for the SAT-matched sample, and 1,872 df, 1,872 df, 1,873 for the freshman-GPA match, respectively.

1.5. The present study

Our primary aim was to examine dynamic interrelations between grade point average (GPA), self-assessments, and instructor judgements across the four years of academics as a function of race and gender. Utilizing USMA's institutionally-derived multi-rater evaluation system, we selected items conceptualizing classroom participation and academic competence. One item indexed perceived expertise to capture anti-Black and anti-female intelligence stereotypes[58][59]; the other measured classroom communication, given racial stereotypes about elocution and lack of authority or credibility

when speaking for women, and the pressure put on both groups to conform their communication to White male standards (e.g., [60][61][62]). Together these represented common gender and race stereotypes and, importantly, frequently-reported microaggressions in education[63][64].

First, linear mixed models on samples matched by SAT and another matched on freshman GPA established whether racial and/or gender discrepancies existed in GPA and classroom competence ratings. Then, a random intercept cross-lagged panel model (RI-CLPM; Figure 1) evaluated dynamic relationships occurring above static, between-person effects. Race and gender effects were tested for the overarching relationships between GPA, self-rating, and instructor rating, and also for their dynamic influences from year to year. We expected linear nixed models to show group-level differences in GPA and ratings that favored men for self-assessments and White students for GPA and instructor ratings within our matched samples. Although there is limited prior work using teacher and self-assessments as well as GPA over time that also considered race and gender, for all students we expected both stable and dynamic interrelationships between our three variables.

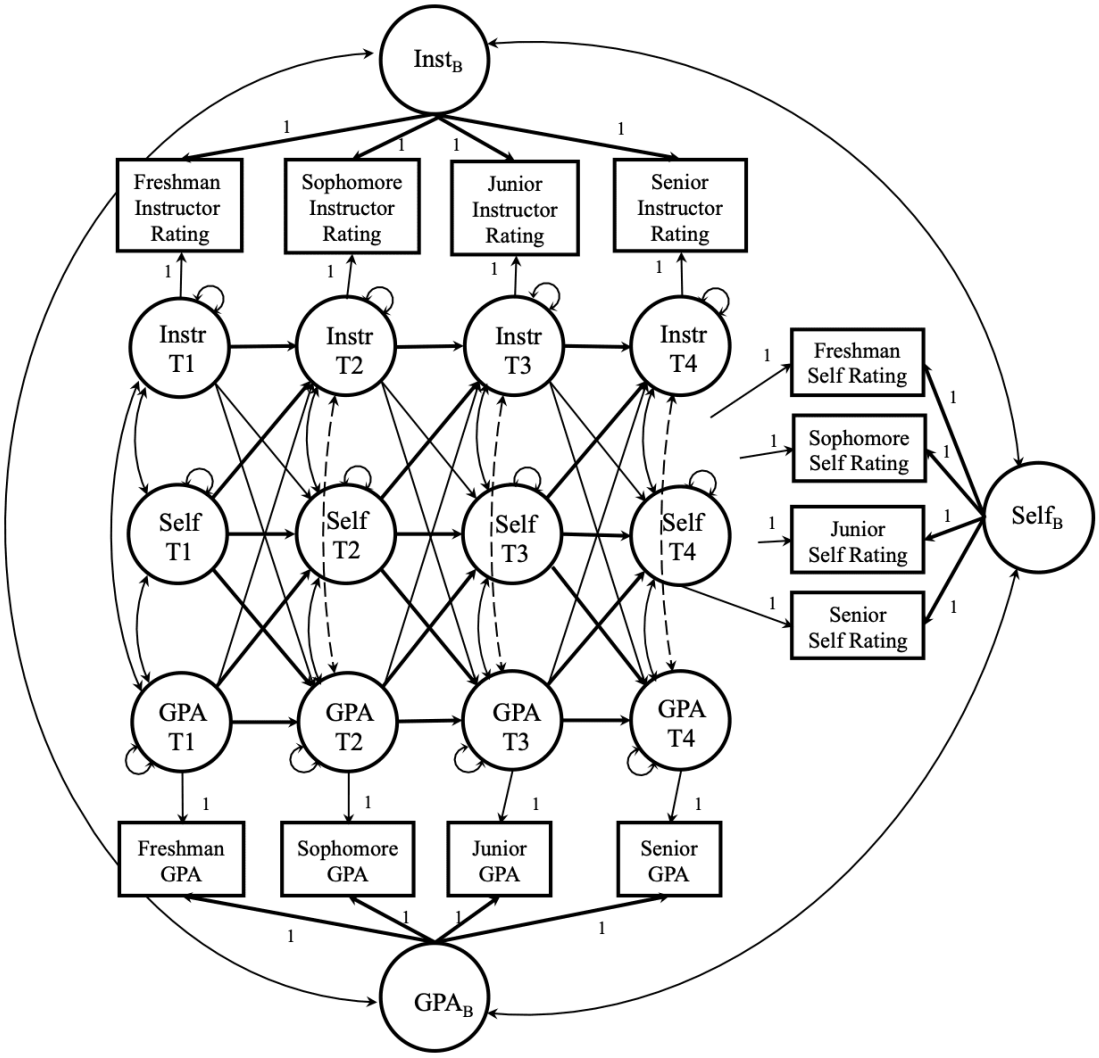

Figure 1. Full RI-CLPM evaluating temporal effects of instructor and self-ratings of classroom participation/knowledge, and GPA. $Instr_B$, $Self_B$, and GPA_B are the between-person random intercepts.

2. Method

2.1. Participants and Data Sources

Our full sample included *N*=5,842 White men, 1,375 White women, 1,054 Black men, and 341 Black women from the graduating classes of 2017 to 2023. Student ages were unavailable but are comparable to other collegiate institutions[2]. Models were estimated via full information maximum likelihood (FIML), which allows for missing values, so students who did not graduate or paused their education were included. At

USMA, a self-reported entry defines race and ethnicity with options: White, Black, Hispanic, Asian, and other, and only binary gender options. Only Black and White-identifying students were included due to the size of other groups among women. The study was approved by USMA's Institutional Review Board (IRB) system; all data were obtained from USMA's data warehouse and coded by an anonymous ID in accordance with IRB procedures.

2.2. Measures

2.2.1. Instructor and Self Ratings

As noted above, USMA utilizes a rating system to provide students regular feedback on the 23 attributes important to USMA and the U.S. Army[65]. For each of the two semesters per academic year, students complete the 23-item scale themselves and are assessed by a randomly selected instructor from one of their courses; self and instructor versions are identical (Supplementary Table 1). Instructor ratings capture personal classroom observations, whereas student self-rating covers all courses in the semester. Feedback sessions are completed with the instructor, where instructors review ratings and provide space for discussion. Here, USMA emphasizes self-reflection and notes, "When receiving feedback from others, cadets often see their perceptions of their actions or intentions have been perceived differently … and can take action to close that perception-reality gap."[8]. Ratings are archived into the student's portal, follow the student until graduation, and can be accessed by supervising officers at will.

We used two items from this instrument relevant to classroom participation and academic performance. "Expertise" rates the student on whether they, "Possess facts, beliefs, and logical assumptions in relevant areas." "Communication" reads: "Clearly expresses ideas to ensure understanding and employs effective communication techniques." Other items, like leadership, are military-relevant and not conceptually as related to grades, or are possibly related to grades, like "Sound Judgement," but were more intangible (see Supplementary Table 1 for full scale). Although all items on this instrument are somewhat conceptual, we were most interested in (ratings of) behaviors observable in the classroom and not abstract constructs, as these may be differentially impacted by stereotypes[66]. The two items were averaged across semesters (two self-ratings and two from different instructors) to index perceived classroom participation and competence; correlation between the two items within a semester and rater was r=.58. We also selected two items hypothesized to be *least* related to academics and classroom performance, Empathy and Tact, and used an analogous average of these socially-relevant items in a final model to ensure that results

from the classroom competence metric were not instead generic to the rating instrument; correlation between these two items within rater was r=.70. In the spring semester of 2018, the rating scale was changed from 1-4 to 1-5 for instructors, and the self-rating scale was changed similarly the next fall. Scores on the 1-4 format were scaled (minimum and maximum preserved) into 1-5 for analyses. Instructor ratings of classroom competence correlated moderately-to-weakly with GPA and self-ratings (r=.32 and .17, respectively, p's<.001), and self-ratings correlated similarly with instructor (r=.17, p<.001), suggesting each variable captured a unique perspective on student performance. For the social metric, the correlation with GPA was r=.09 and .10 for self and instructor rating, respectively (p<.001).

2.2.2. Grade-Point Averages

A student's total GPA at USMA is composed of three components: academic (course grades), military (military course grades and military leadership score based on each semester's duty assignment), and physical GPA (physical education course grades, sportsmanship score). The full GPA determines class rank, while ultimately influences branch selection and other professional opportunities at USMA[14]. For the current study, we were most interested in academic GPA given that it is the largest component of the full GPA and, importantly, the source of academic GPA (coursework) is more consistent across students as compared to military GPA, which is heavily influenced by factors such as leadership detail placement, and company assignment as there is a forced distribution among groups of cadets[67]. Further, the use of academic GPA allows better generalization to other universities, including other commissioning sources. This "academic GPA" (henceforth referred to as just GPA) at USMA is a traditional computation of course grades in academic areas with a maximum of 4.33[67]. The first two years are composed nearly entirely of required courses, and the final years fulfill major requirements. GPAs, as well as other data were obtained from the data warehouse following IRB approval and were linked to an anonymous ID.

2.3. Analytic Approach

The analytic approach contained two components: A. linear mixed models tested for differences by demographic group (Black, White x Male, Female) in GPA, self, and instructor ratings for the freshman and senior year to contextualize the greater system and test for **average discrepancies** among groups at the start and end of cadet education, and B. an RI-CLPM that accounts for "baseline" or between-person effects which evaluated, after accounting for any "baseline" differences, whether demographics influenced how GPA, self, and instructor ratings **dynamically influenced** each other over time.

For the linear mixed models, baseline differences in outcomes (e.g., GPA) by race a were expected across the entire Corps of Cadets, especially given a 128-point discrepancy in average SAT by race in this sample, and likely reflects the generational "Education Debt" incurred over generations of oppression[4]. Therefore, matching processes were used to understand whether performance differed consistently by race and/or gender for cadets when specific characteristics were accounted for, statistically speaking. Two sets of matched participants were created. First, subsamples of White men and women were matched to Black men and women by SAT; from this, linear mixed models run separately for GPA, self, and instructor rating tested for effects of race and gender. This established whether groups with comparable test scores at entry had comparable performance at USMA, addressing concerns that baseline racial differences were merely due to secondary education or related pre-USMA factors[68]. A second set of mixed models used samples matched by freshman year GPA and tested whether graduating senior GPA, self, and instructor ratings of classroom competence differed by race among graduating seniors; these tests helped illuminate whether cadets performing well during the first year maintained equitable performance over time.

A stepwise process and AICc-based selection criteria then evaluated the model shown in Figure 1, an RI-CLPM that tested how GPA, self-ratings, and instructor ratings influenced the next year's scores (autoregressive and cross-lag effects), accounting for individual variable stability (between-person random intercept). This model accounts for background differences and tests how variables influence each other over time, permitting better causal inferences than models without baseline terms[69]. First, for all students in a single group, we established whether the between-person random intercepts significantly increased the model fit as compared to a traditional CLPM, and then whether the relationships between GPA, self, and instructor ratings were equivalent over time (time invariance). This established the appropriateness of the model shown in Figure 2 across *all* students; next we tested whether parameters differed by race and gender, i.e., tests of group invariance[70]. Race and gender differences were evaluated for: manifest intercepts (scalar invariance), variance/covariance, and regressions/cross-lag parameters. Notably, some common types of invariance, i.e., metric, are not applicable to an RI-CLPM. Akaike weights determined the best fit at each step[71]. This approach avoided excessive significance testing, should we have freed parameters individually or created race by gender confidence intervals around all parameters, while also allowing vital group differences to emerge[72].

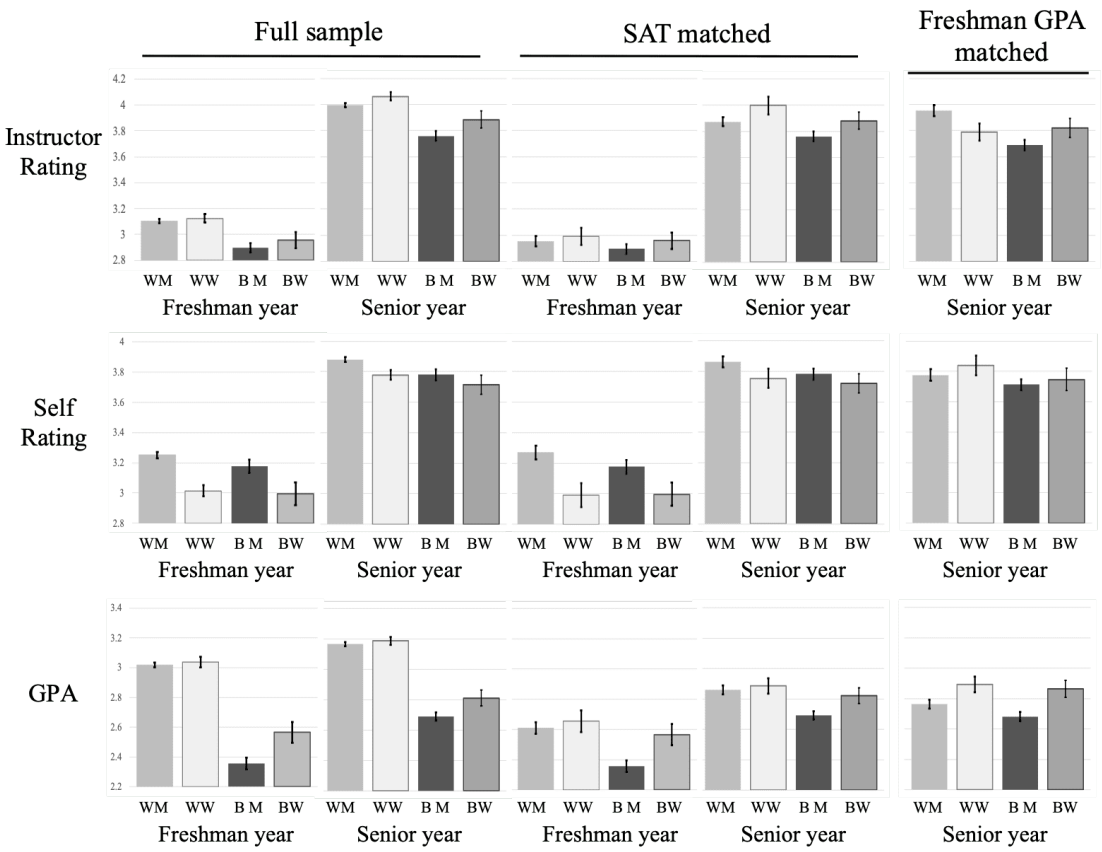

Figure 2. Mean classroom competency ratings and GPA for the Freshman and Senior year. Subsamples of White students were selected to match across gender and SAT, or gender and freshman GPA. Error bars are +/- 2 standard errors. WM = White men, WW = White women, BM = Black men, BW = Black women. See Table 1 for sample ns.

To correct for imbalanced group sizes when invariance testing, which could mask differences in smaller groups[73], we tested both the SAT-matched sample and a randomly subsampled n = 1054 (2nd-smallest subgroup) for each White men and women. The random sample assured that matching process did not introduce secondary problems due to differences in distributions or impacts of the matching variable, especially when using standardized tests[74][75]. The final model was run with the full sample. Finally, although ratings of classroom competence were of primary interest, the final model was run with the full sample and socially-relevant ratings, constructing confidence intervals around key parameters to test whether results were unique to classroom competence or generic to the rating scale.

R version 4.0.2 was used throughout[76]. Matched samples were created with package mmsample, mixed models run with lmer, and simple effects with phia[77][78][79]. RI-CLPM code was initially generated with

riclpmr[80]; however, this program did not support multi-group constraints, so syntax was edited. Models were fit with the lavaan package[81]; AICcmodavg generated AICc and Akaike weights from the resultant fits[82].

3. Results

3.3.1. Mixed Models with Matched Groups

Table 1 shows average GPA, self and instructor ratings for freshman and at graduation, for the full and matched samples. of For samples matched by SAT, mixed models tested the gender by race by time interaction for GPA, self and instructor rating; significant main effects were found for year, race, and gender for all three (except for a p=.057 effect for race for the self-rating). A significant race by gender interaction emerged for GPA and a time by gender interaction for self-rating. Black students were given lower ratings by instructors and had lower GPAs than White students with comparable SATs; simple effects demonstrated that the effect was stronger for men versus women. Men rated themselves more highly than women as freshmen and seniors, and the effect was stronger for freshmen; however, women earned higher GPAs and instructor ratings than men.

Mixed models were also run for the socially-relevant rating items within the SAT-matched sample, and uncovered a significant gender effect for instructor but not self-rating, such that men were rated lower as compared to women by instructors but rated themselves on par with women (instructor $F(1,3084) = 33.4$, $p<.0001$, self $F(1,3212) = 0.008$, $p=.93$; see Supplementary Table 1). For both ratings, seniors were rated higher than freshmen (instructor $F(1,3084)=954$, $p<.0001$, self $F(1,3212)=879$, $p<.0001$). This indicated that the race effects were limited to the classroom competence ratings and not reflective of more general anti-Black bias or a function of the rating instrument or system itself.

Finally, a sample was created that was matched by gender and freshman-year GPA, and models tested graduating senior GPA, instructor classroom competence rating, and self-rating of classroom competence. In order to maximally evaluate the aforementioned race and gender effects, models predicted senior outcomes (GPA, instructor and self-rating) from race, gender, race x gender, freshman GPA, and senior GPA for the two ratings. For the ratings, models tested whether race and/or gender discrepancies existed among samples matched on freshman year performance, *even when* past academic performance *and* current academic performance were accounted for. For instructor rating, Black students were rated slightly, but significantly lower than White students with equitable performance (again,

accounting for past and current GPA), whereas women were rated slightly but consistently higher than men. For the self-rating, Black students rated themselves lower than White students. For GPA, the group of Black students with comparable freshman GPAs to White students achieved a lower senior GPA, and men with comparable freshman GPAs graduated with a lower GPA than the group of women.

3.3.2. RI-CLPM Stepwise Procedure Results

A stepwise procedure built a RI-CLPM that evaluated how GPA, self, and instructor ratings of classroom competence influenced each other over time and whether race and gender influenced relationships. The first step tested, for the random subsample, whether between-person random intercepts ($Inst_B$, $Self_B$, GPA_B) significantly increased the model fit compared to a traditional CLPM, and whether relationships between constructs were consistent over time (time invariance). The RI-CLPM fit better than traditional CLPM, and time invariance was rejected (AICc weight = 1.0, AICc > 300 for both). This indicated that GPA, self and instructor ratings contain stable between-person variance, but dynamic relationships among them vary year-to-year. This model was then tested for race and gender effects.

Scalar invariance tests evaluated whether the intercepts for GPA, self and instructor ratings at each timepoint were different by race, gender, race and gender, or were equivalent across group. The model freely estimating manifest intercepts by race and gender was the best fit, which aligns with the *t*-test results above that demonstrated both race and gender differences (AICc weight = 1; Table 2). Next, we tested for variance/covariance invariance by race and gender, and found that the AICc favored the constrained model, i.e., the GPA and rating interrelations were equivalent across groups.

	AICc	Delta	AIC wt	RMSEA
Manifest intercepts (GPA, Self, Inst T1-T4)				
Constrained model (single group)	29797		0.0	.099
Estimate by gender	29571	-226	0.0	.090
Estimate by race	29079	-718	0.0	.054
***Free across four groups**	**28965**	**-832**	**1.0**	**.037**
Variance and covariances				
Estimate by race	28861	-104	0.0	.036
Estimate by gender	28836	-129	.04	.030
***Constrained model (single group)**	**28830**	**-135**	**.96**	**.033**
GPA autoregressive terms and lags				
Constrained model (single group)	28959	+129	0.0	.045
Estimate by gender	28946	+116	0.0	.044
***Estimate by race**	**28823**	**-7**	**.96**	**.033**
Self-rating autoregressive terms and lags				
Estimate by gender	28795	-28	0.0	.033
Estimate by race	28794	-29	0.0	.033
***Constrain across four groups**	**28780**	**-43**	**1.0**	**.032**
Instructor rating autoregressive terms and lags				
Estimate by gender	28752	-28	0.0	.031
Estimate by race	28750	-30	0.0	.031
***Constrain across four groups**	**28740**	**-40**	**.99**	**.030**

Table 2. Results of the RI-CLPM comparison steps, which began with all parameters estimated freely by race/gender group. Entries in grey are the best for that group and are the comparison model for the next set.

doi.org/10.32388/KUANCU

In a similar manner as above, autoregressive parameters and cross-lags for GPA, self- and instructor ratings were tested sequentially for race and gender effects. For GPA, the best model estimated parameters by race, but for instructor feedback and self-rating estimates, the constrained model estimating all cadets together was the best fit (Table 1); steps were conducted in this order, but all orders yielded the same final model. Contrary to predictions, the stepwise procedure uncovered no significant gender differences in directional effects; in the final model only intercepts were estimated by gender (and race).

2.3.3. RI-CLPM Final Model

The final model was an excellent fit to the full dataset (Figure 3; CFI = .993, RMSEA = .028, SRMR = .033). The stable, between-person factors for GPA, instructor, and self-rating were significantly related to each other. Instructor rating and GPA were strongly interrelated over time, both for the stable, between-person estimates and dynamic year-to-year relationships. Self-rating in Year 2 was significantly predicted by all freshman-year constructs (instructor and self-rating, GPA), and thereafter mostly related to previous self-rating. Two inverse relationships emerged in the senior year: junior year self-rating negatively predicted senior year instructor ratings, and junior GPA negatively related to senior self-rating.

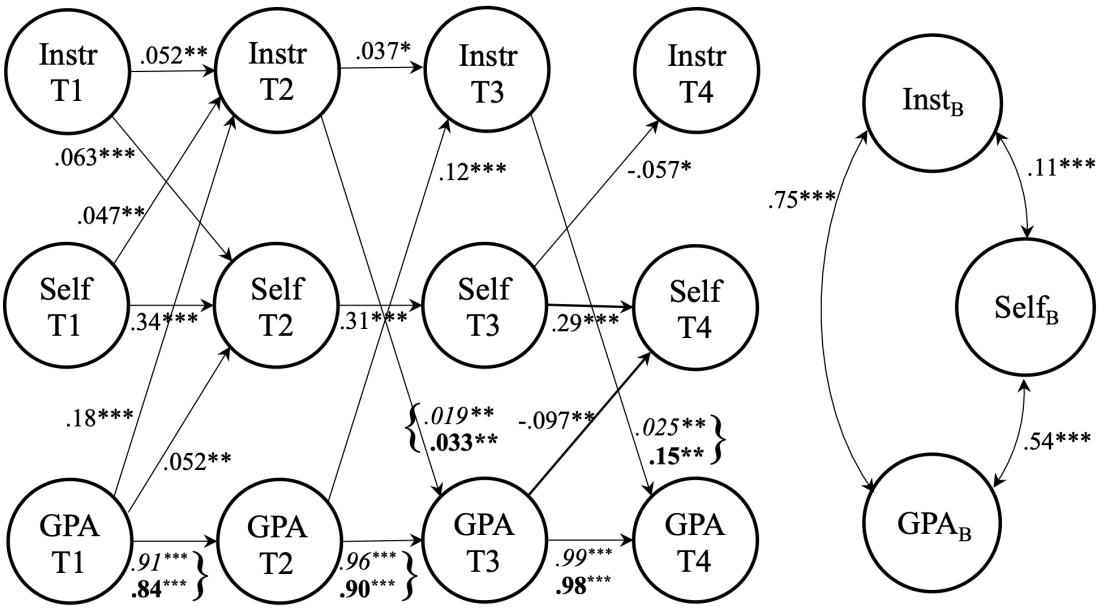

Figure 3. Results of the final RI-CLPM, which was invariant for variance/covariance, but had reliably different lags by race (regressions) for GPA. GPA coefficients for White students are on top in italics, and for Black students below and bolded; parameter pairs with parentheses have non-overlapping 95% confidence intervals. $Instr_B$, $Self_B$, and GPA_B are the between-person random intercepts, which were equivalent across groups. Note that the instructor raters are different each year. *** $p \leq .001$ ** $p < .01$ * $p < .05$

The stepwise tests above demonstrated a race effect for GPA prediction, in that estimating the lags predicting GPA separately by race improved model fit; confidence intervals (CI) around significant parameters showed that the effect of the previous year's instructor feedback was significantly higher for Black versus White students in years three and four (Black junior year 95% CI=.023-.043, White =.012-.023; Black senior year =.058-.23, White = .002-.049). This indicated that the influence of instructor rating on future GPA was higher for Black students for these years. Autoregressive effects on GPA were stronger for White students in years two and three (Black sophomore year=.82-.87, White =.90-.92; Black junior year =.88-.92, White = .99-1.01).

The final model was also run with the socially-relevant items, and was a good fit to the data (CFI = .993, RMSEA = .026, SRMR = .028). The stable, between-person estimates of instructor and self-ratings were significantly related to each other (β= .54, p<.001), but as compared to the classroom competence ratings, the social items were less positively related to GPA (social items self-rating to GPA β= -.19, 95% CI = -.35=.-033, classroom competence β= .11, 95% CI .067-.16; instructor to GPA β= .29, 95% CI .23-.34,

classroom competence β= .75, 95% CI = .71-.79; all *p*s<.001; See Supplementary Figure 1). Additionally, the directional effects from instructor rating to GPA were not significant for either Black or White students from year 3 to 4 (White β=- .02, *p*=.10; Black β=- .04, *p*=.38), and the effect from year 2 to 3 was significant only for White students, with confidence intervals overlapping the Black students (White student β=.009, *p*=.032, 95% CI = .001-.017; Black student β=- .032, *p*=.054, 95% CI= -.001-.065). This suggested that the race-based effect of instructor classroom competence rating influencing future GPA was not generic to the rating instrument.

4. Discussion

This study illuminated processes underlying achievement in a military setting, how these differed by demographics, and demonstrated how, in the context of USMA, self-assessment, academic performance (GPA), and (predominantly) military Officer instructor ratings are interrelated. Black students, especially men, receive lower GPAs than expected from their SATs, and Black students are consistently rated by instructors as having lower classroom competence. As compared to women, men over-rate themselves across capabilities despite lower GPAs and instructor ratings. GPA and instructor judgments about classroom competence influenced each other positively over time; vitally, the effect of instructor judgments on the next year's GPA was stronger for Black as compared to White students. This formed a positive feedback loop wherein students who receive higher grades are perceived as having greater classroom competence the next year, and then receive higher grades the following year, and so on; this loop operates identically for poorer performance, which begets even poorer future feedback. This relationship is stronger among Black students who, on average, start with lower feedback and performance. These longitudinal differences cannot be explained by SAT or even first-year GPA, in that Black, as compared to White, graduating seniors earn lower GPAs than predicted by their freshman-year performance, and are given lower ratings of classroom competence even accounting for their own current and past grades.

Findings agree with and bring together previous work, often in younger students, examining academic performance, teacher judgments, and self-assessments (e.g., [24]), and for the first time explain longitudinal quantitative performance in a military Officer setting. We replicated the general-population finding that self-beliefs are formed from past performance, but the reverse effect of self-beliefs on future performance is weaker[24]. Current results also agree with work on longitudinal interrelationships of teacher judgements and student grades, and that teacher and student reports have positive but relatively

low agreement[83][40]. Two negative relationships in the senior year, from junior self-rating to senior instructor rating and from junior GPA to senior self-rating, might be associated with under-calibration, or under-estimating one's performance or knowledge, which can be beneficial for future performance[25].

A Black-White difference was found in GPA and instructor ratings of classroom ability, even for the matched groups; this agrees with other work in White-majority institutions showing that racial disparities in college cannot be offset by "background"[84][68]. This is concerning regardless, but at USMA final GPA determines class rank, which influences various career options including selection of Army branch[67]; the difference between class ranks for the average senior Black versus White man in the SAT matched sample was 0.47, which would transform a cadet at the median rank (about 525) to a rank of about 200; in other words, a graduating Black cadet picks his branch *over 300 people later* as compared to a White cadet with the same entering SAT.

Teacher judgements of Black students can be either overly generous or negatively biased, and self-concepts of Black students can in some ways be buffered against negative feedback and in other cases adversely impacted (e.g., [46][26][85]). Here, both baseline and dynamic model results suggested that, for this level of self- and instructor-assessment, a negative bias prevailed in that Black cadets receive lower grades and ratings compared to White cadets matched on SAT, and these ratings are more strongly internalized by Black students, impacting the GPA as well as self-concept. Protective factors, such as peer support, found at other institutions[27], did not appear to sufficiently counteract forces facing Black cadets at USMA.

Stable gender effects were found for the self-ratings of both classroom competence and social ability, replicating previous work showing men assess themselves more positively[36]. Contrary to predictions, no significant gender effects emerged for *interrelations* between self-assessment, teacher judgements, and performance, although intercepts were estimated by race and gender. Thus, lower self-assessment of women was not differentially impacted, as compared to men, by either GPA or teacher effects. Although White women appear generally successful in regard to their GPA and instructor ratings, greater attention could be paid to *all* women's self-concepts, particularly with regard to leadership abilities in order to maintain motivation in a stereotype laden career such as the military[86][87]. Despite relatively poorer self-assessment, instructors rated the classroom competence of women higher than men (of the same race), which might simply reflect the higher GPAs earned by women with equitable SATs. Such a finding

should be contrasted with the generally negative views about women in the military, even at this specific institution[12]. Vitally, further study is warranted to determine whether these results translate into counterstereotypical military domains, such as leadership, that are important to how these cadets are perceived and rated as Officers[88].

The design of the current study does not allow for conclusive statements about underlying mechanisms for race and gender effects; however, results are consistent with similar, non-experimental field studies in which an intellectual competence context is sufficient to induce stereotype threat, one possible underlying mechanism[33]. Although the sizes of lagged effects of teacher feedback on future GPA were much smaller than autoregressive effects, the ability of instructor judgements *on a single course* to have any impact on subsequent GPA speaks to the power of these instructors and this setting. These processes could be better understood via measures of classroom dynamics, growth mindset, or bias measures (e.g., [43]). Importantly, we can only speculate about the relative importance of the feedback session, rating itself, instructor rank, and/or how instructor perception might impact classroom dynamics, as all are likely to be impactful (e.g., [89][40][85]).

Despite mechanistic imprecision, instructor judgements appeared to be another avenue through which students might experience disparities. Other work has demonstrated that teacher feedback is impactful, and summary feedback is more influential for some students in particular[50][52]. Given the ability of instructors to influence student outcomes in general, and Black students' grades in particular, extra care should be taken to make sure feedback is congenial, actively avoids racial/gendered stereotypes (e.g., treating resistance as agitation), and is specific and goal-driven[90][91]. USMA specifically designed this rating system to promote professional development; any institution invested in professional and character assessment should ensure that such evaluation systems do not contribute to marginalization of underrepresented groups.

It is important to emphasize that these results should *not* be interpreted in ways that add to the deficit-based narrative of Black underachievement and academic disengagement, even if that disengagement is brought on by racism[92][93], but instead potentiate thorough interrogation of the conceptualization and measurement of "success" by the very people who have benefitted from the current conceptualization[94]. We cannot say whether the processes underlying poorer scores given equivalent entry tests and stronger internalization of teacher feedback for Black students were predominantly student-teacher interactions, implicit beliefs, explicit racism, structural/institutional elements, and/or

factors within students (e.g., stereotype threat). However, real-life racial injustices contain mechanistic uncertainty, and openly considering the source of large-scale racial disparities is vital to sustaining conversations that can disrupt racist practices[95]. In education and the military, racial or "equal opportunity" discussions often focus on small-group climate and individual acts of bias (e.g., [56]) while avoiding structural contributions; such treatments give lip service to awareness but in fact support White supremacy by ignoring systems of power that have great potential for change[96][97]. Academic institutions, military academies, and militaries at large should seek to understand systemic influences on performance and the factors unique to marginalized students that occur within institutions, including outcome measurement, group dynamics, and provide true anti-racist leader education[98]. Meaningfully addressing systemic racism is key for the military given its historical position as mechanism of social mobility for disadvantaged groups and how it presents itself as a meritocracy[99][100], contrasted with the lived experiences of Black Soldiers who may eventually feel, "resignation and acceptance that they could not change the system … and gave up fighting to do so,"[101].

Several limitations of this study warrant discussion. First is the military academic context, which limits generalizability to the larger miliary to some degree. Our utilization of classroom competence ratings and academic GPA was an attempt to maximize generalization of the study across contexts, contribute to the body of knowledge (and often, debate) on stereotype threat and related mechanisms, and avoid some nuisance variance due to the specifics of the USMA system. However, USMA presents students with several competing identities across military, academic, and physical fitness programs, as well as athletic, social, and service opportunities[102]; experiences might be impacted differently and students might find support and growth in other areas when faced with negative academic feedback. Additional measures, such as stigma consciousness, belonging, or peer support might refine relationships we found, including illumination of protective factors found in other, often qualitative, studies[103][27]. Ratings captured self-assessments and instructor reports of classroom competence, and the perspectives therein were our primary interest, but evaluation of participation and grading are vital future directions. We were also unable to evaluate instructor demographics beyond a general active duty – civilian split and did not examine course content/selection, other important influences[104][105]. Results nonetheless suggested that immediate attention is needed to educate instructors as to the pervasive and structural systems maintaining racial inequalities within education and beyond[106][107].

In sum, anti-Black disparities are pervasive in historically White institutions, including USMA, as part of the Education Debt that has accumulated over generations[4]. Disparities may be maintained by positive feedback loops whereby early discrepancies are exacerbated over time selectively for the same groups who may be told along the way that everyone has an equal chance of success[63]. Military instructors and supervisors are integral to this process, and their perceptions show clear evidence of anti-Black bias. In this context, White women perform well and are perceived positively, although more attention could be paid to their self-concept. Black students may more intensely internalize feedback, which is more negative on average than their White counterparts. Immediate and genuine attention needs to be paid to addressing the obstacles to success for Black Soldiers.

Footnotes

[1] Classroom sizes are consistent, courses/exams are often standardized, students have few opportunities to select instructors, and a larger percentage of courses are required.

[2] Students must be between 17 and 23 upon entry, see https://www.westpoint.edu/admissions/steps-to-admission

References

1. ^*Buchanan NT, Perez M, Prinstein MJ, Thurston IB (2021). "Upending Racism in Psychological Science: Strategies to Change How Science Is Conducted, Reported, Reviewed, and Disseminated." Am Psychol. 76(7):1097–1112. doi:10.1037/amp0000905.*
2. ^*Kempf A (2020). "If We Are Going to Talk About Implicit Race Bias, We Need to Talk About Structural Racism: Moving Beyond Ubiquity and Inevitability in Teaching and Learning About Race." Taboo J Cult Educ. 19 (2):19.*
3. ^*Bensimon EM (2005). "Closing the Achievement Gap in Higher Education: An Organizational Learning Perspective." New Dir High Educ. 2005(131):99–111. doi:10.1002/he.190.*
4. a, b, c*Ladson-Billings G (2006). "From the Achievement Gap to the Education Debt: Understanding Achievement in U.S. Schools." Educ Res. 35(7):3–12.*
5. ^*Grewenig E, Lergetporer P, Werner K, Woessmann L, Zierow L (2021). "COVID-19 and Educational Inequality: How School Closures Affect Low- and High-Achieving Students." Eur Econ Rev. 140:103920. doi:10.1016/j.euroecorev.2021.103920.*

6. a, b, c *Asch BJ, Miller T, Malchiodi A (2012). A New Look at Gender and Minority Differences in Officer Career Progression in the Military. RAND.*
7. a, b, c *Burk J, Espinoza E (2012). "Race Relations Within the US Military." Annu Rev Sociol. 38(1):401–422. doi: 10.1146/annurev-soc-071811-145501.*
8. a, b, c *United States Military Academy (2018). Developing Leaders of Character.*
9. a, b *Glenn H (2020). "West Point Graduates' Letter Calls For Academy To Address Racism." NPR.Org. https://www.npr.org/sections/live-updates-protests-for-racial-justice/2020/07/06/887540591/west-point-graduates-letter-calls-for-academy-to-address-racism.*
10. ^ *Rosenfeld P, Newell CE, Le S (1998). "Equal Opportunity Climate of Women and Minorities in the Navy: Results From the Navy Equal Opportunity/Sexual Harassment (NEOSH) Survey." Mil Psychol. 10(2):69–85.*
11. ^ *Rosenstein JE, Angelis KD, McCone DR, Carroll MH (2018). "Sexual Assault and Sexual Harassment at the US Military Service Academies." Mil Psychol. 30(3):206–218.*
12. a, b, c *Matthews MD, Ender MG, Laurence JH, Rohall DE (2009). "Role of Group Affiliation and Gender on Attitudes Toward Women in the Military." Mil Psychol. 21:241–251.*
13. ^ *Seidule T (2017). "From Slavery to Black Power." In Intolerance: Political Animals and Their Prey. Vol. 1. p. 69.*
14. a, b, c *United States Military Academy (2016b). Academic Program (RedBook).*
15. ^ *Carrell SE, Page ME, West JE (2010). "Sex and Science: How Professor Gender Perpetuates the Gender Gap *." Q J Econ. 125(3):1101–1144.*
16. a, b *Aronson J, Inzlicht M (2004). "The Ups and Downs of Attributional Ambiguity: Stereotype Vulnerability and the Academic Self-Knowledge of African American College Students." Psychol Sci. 15(12):829–836.*
17. ^ *Bandura A (1977). "Self-Efficacy: Toward a Unifying Theory of Behavioral Change." Psychol Rev. 84(2):191–215.*
18. ^ *Klassen RM, Usher EL (2010). "Self-Efficacy in Educational Settings: Recent Research and Emerging Directions." In Urdan TC, Karabenick SA (Eds.), The Decade Ahead: Theoretical Perspectives on Motivation and Achievement: Vol. 16 Part A. Emerald Group Publishing Limited. pp. 1–33.*
19. ^ *Manzano-Sanchez H, Outley C, Gonzalez JE, Matarrita-Cascante D (2018). "The Influence of Self-Efficacy Beliefs in the Academic Performance of Latina/o Students in the United States: A Systematic Literature Review." Hisp J Behav Sci. 40(2):176–209.*
20. ^ *Dixson DD, Worrell FC, Olszewski-Kubilius P, Subotnik RF (2016). "Beyond Perceived Ability: The Contribution of Psychosocial Factors to Academic Performance." Ann N Y Acad Sci. 1377(1):67–77.*

21. ^ Morgan SL, Mehta JD (2004). "Beyond the Laboratory: Evaluating the Survey Evidence for the Disidentification Explanation of Black-White Differences in Achievement." Sociol Educ. 77(1):82–101.

22. ^ Pieschl S (2009). "Metacognitive Calibration—An Extended Conceptualization and Potential Applications." Metacognition Learn. 4(1):3–31.

23. ^ Zabrucky KM (2010). "Knowing What We Know and Do Not Know: Educational and Real World Implications." Procedia Soc Behav Sci. 2(2):1266–1269.

24. a, b, c Talsma K, Schüz B, Schwarzer R, Norris K (2018). "I Believe, Therefore I Achieve (and Vice Versa): A Meta-Analytic Cross-Lagged Panel Analysis of Self-Efficacy and Academic Performance." Learn Individ Differ. 61:136–150.

25. a, b Talsma K, Schüz B, Norris K (2019). "Miscalibration of Self-Efficacy and Academic Performance: Self-Efficacy≠Self-Fulfilling Prophecy." Learn Individ Differ. 69:182–195.

26. a, b Murry VM, Berkel C, Simons RL, Simons LG, Gibbons FX (2014). "A Twelve-Year Longitudinal Analysis of Positive Youth Development Among Rural African American Males." J Res Adolesc. 24(3):512–525.

27. a, b, c Harper SR (2007). "Peer Support for African American Male College Achievement: Beyond Internalized Racism and the Burden of "Acting White."" J Men's Stud. 14(3):337–358. doi:10.3149/jms.1403.337.

28. ^ Harper SR (2015). "Black Male College Achievers and Resistant Responses to Racist Stereotypes at Predominantly White Colleges and Universities." Harv Educ Rev. 85(4):646–674. doi:10.17763/0017-8055.85.4.646.

29. ^ O'Callaghan KW, Bryant C (1990). "Noncognitive Variables: A Key to Black-American Academic Success at a Military Academy?" J Coll Stud Dev. 31(2):121–126.

30. ^ Wasserberg MJ (2014). "Stereotype Threat Effects on African American Children in an Urban Elementary School." J Exp Educ. 82(4):502–517.

31. ^ Steele CM (1997). "A Threat in the Air: How Stereotypes Shape Intellectual Identity and Performance." Am Psychol. 52(6):613–629.

32. ^ Walton GM, Spencer SJ (2009). "Latent Ability: Grades and Test Scores Systematically Underestimate the Intellectual Ability of Negatively Stereotyped Students." Psychol Sci. 20(9):1132–1139.

33. a, b Chung BG, Ehrhart MG, Holcombe Ehrhart K, Hattrup K, Solamon J (2010). "Stereotype Threat, State Anxiety, and Specific Self-Efficacy as Predictors of Promotion Exam Performance." Group Organ Manag. 35(1):77–107.

34. ^ Franceschini G, Galli S, Chiesi F, Primi C (2014). "Implicit Gender–Math Stereotype and Women's Susceptibility to Stereotype Threat and Stereotype Lift." Learn Individ Differ. 32:273–277.

35. ^Cech E, Rubineau B, Silbey S, Seron C (2011). "Professional Role Confidence and Gendered Persistence in Engineering." Am Sociol Rev. 76(5):641–666.

36. a, bPallier G (2003). "Gender Differences in the Self-Assessment of Accuracy on Cognitive Tasks." Sex Roles. 48(5):265–276. doi:10.1023/A:1022877405718.

37. ^Torres-Guijarro S, Bengoechea M (2017). "Gender Differential in Self-Assessment: A Fact Neglected in Higher Education Peer and Self-Assessment Techniques." High Educ Res Dev. 36(5):1072–1084.

38. a, bArcher EM (2013). "The Power of Gendered Stereotypes in the US Marine Corps." Armed Forces Soc. 39 (2):359–391.

39. ^Byrn J, Royal G (2020). "What Should West Point Do About Its Robert E. Lee Problem?" Modern War Institute. https://mwi.usma.edu/west-point-robert-e-lee-problem/.

40. a, b, cKriegbaum K, Steinmayr R, Spinath B (2019). "Longitudinal Reciprocal Effects Between Teachers' Judgments of Students' Aptitude, Students' Motivation, and Grades in Math." Contemp Educ Psychol. 59:101807.

41. ^Vattøy K-D (2020). "Teachers' Beliefs About Feedback Practice as Related to Student Self-Regulation, Self-Efficacy, and Language Skills in Teaching English as a Foreign Language." Stud Educ Eval. 64:100828.

42. ^Zhu M, Urhahne D, Rubie-Davies CM (2018). "The Longitudinal Effects of Teacher Judgement and Different Teacher Treatment on Students' Academic Outcomes." Educ Psychol. 38(5):648–668.

43. a, bCanning EA, Muenks K, Green DJ, Murphy MC (2019). "STEM Faculty Who Believe Ability Is Fixed Have Larger Racial Achievement Gaps and Inspire Less Student Motivation in Their Classes." Sci Adv. 5(2):eaau4734.

44. ^Downey DB, Pribesh S (2004). "When Race Matters: Teachers' Evaluations of Students' Classroom Behavior." Sociol Educ. 77(4):267–282.

45. ^Morris EW, Perry BL (2017). "Girls Behaving Badly? Race, Gender, and Subjective Evaluation in the Discipline of African American Girls." Sociol Educ. 90(2):127–148.

46. a, bHarber K, Gorman J, Gengaro F, Butisingh S, Tsang W, Ouellette R (2012). "Students' Race and Teachers' Social Support Affect the Positive Feedback Bias in Public Schools." J Educ Psychol. 104:1149.

47. ^van den Bergh L, Denessen E, Hornstra L, Voeten M, Holland RW (2010). "The Implicit Prejudiced Attitudes of Teachers: Relations to Teacher Expectations and the Ethnic Achievement Gap." Am Educ Res J. 47(2):497–527.

48. ^Campbell T (2015). "Stereotyped at Seven? Biases in Teacher Judgement of Pupils' Ability and Attainment." J Soc Policy. 44(3):517–547.

49. ^Moss-Racusin CA, Dovidio JF, Brescoll VL, Graham MJ, Handelsman J (2012). "Science Faculty's Subtle Gender Biases Favor Male Students." Proc Natl Acad Sci U S A. 109(41):16474–16479.

50. a, bHattie J, Timperley H (2007). "The Power of Feedback." Rev Educ Res. 77(1):81–112. doi:10.3102/00346540298487.

51. ^MacDonald RB (1991). "Developmental Students' Processing of Teacher Feedback in Composition Instruction." Rev Res Dev Educ. 8(5). https://eric.ed.gov/?id=ED354965.

52. a, bPhillips F, Wolcott S (2014). "Effects of Interspersed Versus Summary Feedback on the Quality of Students' Case Report Revisions." Account Educ. 23(2):174–190.

53. ^Davis L, Klahr DA, Tercha J, Hill A, Klauberg MWX, White A, Owen B, Rehlberg K (2019). 2019 Military Service Gender Relations Focus Groups. Office of People Analytics (OPA). https://media.defense.gov/2020/Apr/30/2002291696/-1/-1/1/15-ANNEX-1-2019-MILITARY-SERVICE-GENDER-RELATIONS-FOCUS-GROUPS-OVERVIEW-REPORT.PDF.

54. ^Riegle-Crumb C, Humphries M (2012). "Exploring Bias in Math Teachers' Perceptions of Students' Ability by Gender and Race/Ethnicity." Gender Soc. 26(2):290–322.

55. ^Choo HY, Ferree MM (2010). "Practicing Intersectionality in Sociological Research: A Critical Analysis of Inclusions, Interactions, and Institutions in the Study of Inequalities." Sociol Theory. 28(2):129–149.

56. a, bWalsh BM, Matthews RA, Tuller MD, Parks KM, McDonald DP (2010). "A Multilevel Model of the Effects of Equal Opportunity Climate on Job Satisfaction in the Military." J Occup Health Psychol. 15(2):191–207. doi:10.1037/a0018756.

57. ^Stewart JB, Firestone JM (1992). "Looking for a Few Good Men: Predicting Patterns of Retention, Promotion, and Accession of Minority and Women Officers." Am J Econ Sociol. 51(4):435–458.

58. ^von Hippel W, Hawkins C, Schooler JW (2001). "Stereotype Distinctiveness: How Counterstereotypic Behavior Shapes the Self-Concept." J Pers Soc Psychol. 81(2):193.

59. ^Walzer AS, Czopp AM (2011). "Able But Unintelligent: Including Positively Stereotyped Black Subgroups in the Stereotype Content Model." J Soc Psychol. 151(5):527-530.

60. ^Alim HS, Smitherman G (2012). Articulate While Black: Barack Obama, Language, and Race in the U.S. Oxford University Press.

61. ^Myers TK (2020). "Can You Hear Me Now? An Autoethnographic Analysis of Code-Switching." Cult Stud Crit Methodol. 20(2):113–123.

62. ^Rudman LA, Phelan JE (2008). "Backlash Effects for Disconfirming Gender Stereotypes in Organizations." Res Organ Behav. 28:61–79.

63. ^a, ^b *Midgette AJ, Mulvey KL (2021). "Unpacking Young Adults' Experiences of Race- and Gender-Based Micro aggressions." J Soc Pers Relat. 38(4):1350–1370. doi:10.1177/0265407521988947.*

64. ^ *Steketee A, Williams MT, Valencia BT, Printz D, Hooper LM (2021). "Racial and Language Microaggression s in the School Ecology." Perspect Psychol Sci. 16(5):1075–1098.*

65. ^ *U.S. Department of the Army (2012). ADRP 6-22 Army Leadership.*

66. ^ *McCrea SM, Wieber F, Myers AL (2012). "Construal Level Mind-Sets Moderate Self- and Social Stereotypin g." J Pers Soc Psychol. 102(1):51–68. doi:10.1037/a0026108.*

67. ^a, ^b, ^c *United States Military Academy (2016a). Military Program (GreenBook).*

68. ^a, ^b *Martin ND, Spenner KI, Mustillo SA (2017). "A Test of Leading Explanations for the College Racial-Ethnic Achievement Gap: Evidence From a Longitudinal Case Study." Res High Educ. 58(6):617–645.*

69. ^ *Hamaker EL, Kuiper RM, Grasman RPPP (2015). "A Critique of the Cross-Lagged Panel Model." Psychol Me thods. 20(1):102–116.*

70. ^ *Mulder JD, Hamaker EL (2021). "Three Extensions of the Random Intercept Cross-Lagged Panel Model." St ruct Equ Model Multidiscip J. 28(4):638–648.*

71. ^ *Burnham KP, Anderson DR (2002). "A Practical Information-Theoretic Approach." In Model Selection and Multimodel Inference. 2nd ed. Springer.*

72. ^ *Symonds MRE, Moussalli A (2011). "A Brief Guide to Model Selection, Multimodel Inference and Model Ave raging in Behavioural Ecology Using Akaike's Information Criterion." Behav Ecol Sociobiol. 65(1):13–21.*

73. ^ *Yoon M, Lai MHC (2018). "Testing Factorial Invariance With Unbalanced Samples." Struct Equ Model Mult idiscip J. 25(2):201–213.*

74. ^ *Mervis CB, Klein-Tasman BP (2004). "Methodological Issues in Group-Matching Designs: α Levels for Cont rol Variable Comparisons and Measurement Characteristics of Control and Target Variables." J Autism Dev Disord. 34(1):7–17.*

75. ^ *Okazaki S, Sue S, Okazaki S, Sue S (1995). "Methodological Issues in Assessment Research With Ethnic Min orities." Psychol Assess. 7(3):367–375.*

76. ^ *R Core Team (2013). R: A Language and Environment for Statistical Computing. R Foundation for Statistic al Computing.*

77. ^ *Bates D, Maechler M, Bolker B, Walker S (2015). "Fitting Linear Mixed-Effects Models Using lme4." J Stat So ftw. 67(1):1–48.*

78. ^ *O'Connell E (2018). "mmsample: Multivariate Matched Sampling." R package version 0.1.*

79. ^Rosario-Martinez HD (2015). "phia: Post-Hoc Interaction Analysis." R package version 0.2-1.

80. ^Flournoy J (2021). "riclpmr: Generate lavaan Syntax for RI-CLPM." R package version 0.1.0.9000.

81. ^Rosseel R (2012). "lavaan: An R Package for Structural Equation Modeling." J Stat Softw. 48(2):1–36.

82. ^Mazerolle MJ (2020). "AICcmodavg: Model Selection and Multimodel Inference Based on (Q)AIC(c)." R package version 2.3-1.

83. ^Bardach L, Yanagida T, Schober B, Lüftenegger M (2019). "Students' and Teachers' Perceptions of Goal Structures – Will They Ever Converge? Exploring Changes in Student-Teacher Agreement and Reciprocal Relations to Self-Concept and Achievement." Contemp Educ Psychol. 59:101799.

84. ^Fleming J (2002). "Who Will Succeed in College? When The SAT Predicts Black Students' Performance." Rev High Educ. 25(3):281–296.

85. a, bScott TM, Gage N, Hirn R, Han H (2019). "Teacher and Student Race as a Predictor for Negative Feedback During Instruction." Sch Psychol. 34(1):22–31.

86. ^Correll SJ (2001). "Gender and the Career Choice Process: The Role of Biased Self-Assessments." Am J Sociol. 106(6):1691–1730.

87. ^Zenger J, Folkman J (2019). "Research: Women Score Higher Than Men in Most Leadership Skills." 8.

88. ^Braun S, Stegmann S, Bark ASH, Junker NM, Dick R van (2017). "Think Manager—Think Male, Think Follower—Think Female: Gender Bias in Implicit Followership Theories." J Appl Soc Psychol. 47(7):377–388.

89. ^Greer W, Clark-Louque A, Balogun A, Clay A (2018). "Race-Neutral Doesn't Work: Black Males' Achievement, Engagement, and School Climate Perceptions." Urban Educ. 0042085918804015.

90. ^Correll S, Simard C (2016). "Vague Feedback Is Holding Women Back." Harvard Business Review.

91. ^Yeager DS, Purdie-Vaughns V, Garcia J, Apfel N, Brzustoski P, Master A, Hessert WT, Williams ME, Cohen GL (2014). "Breaking the Cycle of Mistrust: Wise Interventions to Provide Critical Feedback Across the Racial Divide." J Exp Psychol Gen. 143(2):804–824.

92. ^Burt BA, Williams KL, Smith WA (2018). "Into the Storm: Ecological and Sociological Impediments to Black Males' Persistence in Engineering Graduate Programs." Am Educ Res J. 55(5):965–1006.

93. ^Harper SR (2009). "Niggers No More: A Critical Race Counternarrative on Black Male Student Achievement at Predominantly White Colleges and Universities." Int J Qual Stud Educ. 22(6):697–712.

94. ^Thorp HH (2020). "Time to Look in the Mirror." Science. 368(6496):1161–1161. doi:10.1126/science.abd1896.

95. ^DiAngelo DR (2018). White Fragility: Why It's So Hard for White People to Talk About Racism. Beacon Press.

96. ^de Saxe JG (2021). "Unpacking and Interrogating White Supremacy Educating for Critical Consciousness a nd Praxis." Whiteness Educ. 6(1):60–74.

97. ^Young EY (2011). "The Four Personae of Racism: Educators' (Mis)Understanding of Individual Vs. Systemic Racism." Urban Educ. 46(6):1433–1460.

98. ^Hoffman JL, Lowitzki KE (2005). "Predicting College Success With High School Grades and Test Scores: Li mitations for Minority Students." Rev High Educ. 28(4):455–474. doi:10.1353/rhe.2005.0042.

99. ^Martinez E, Huerta AH (2020). "Deferred Enrollment: Chicano/Latino Males, Social Mobility and Military Enlistment." Educ Urban Soc. 52(1):117–142. doi:10.1177/0013124518785021.

100. ^Segal DR, Bachman JG, Dowdell F (1978). "Military Service for Female and Black Youth: A Perceived Mobili ty Opportunity." Youth Soc. 10(2):127–134.

101. ^Davis BH (2018). Perceptions of Barriers to Leadership Appointment and Promotion of African American F emale Commissioned Officers in the United States Military. University of San Francisco.

102. ^Schaefer HS, Farina AG, Cotting DI, Proctor ES, Cook CL, Lerner RM (2020). "The Benefits and Liabilities of Risk-Taking Propensity and Confidence at the U.S. Military Academy." Armed Forces Soc. 0095327X209733 73.

103. ^Brown RP, Lee MN (2005). "Stigma Consciousness and the Race Gap in College Academic Achievement." S elf Identity. 4(2):149–157.

104. ^Howard J, Zoeller A, Pratt Y (2006). "Students' Race and Participation in Sociology Classroom Discussion: A Preliminary Investigation." J Scholarsh Teach Learn. 6(1):14–38.

105. ^Karpowitz CF, Mendelberg T, Mattioli L (2015). "Why Women's Numbers Elevate Women's Influence, and When They Do Not: Rules, Norms, and Authority in Political Discussion." Polit Groups Identities. 3(1):149–17 7.

106. ^Alvaré MA (2018). "Addressing Racial Inequalities Within Schools: Exploring the Potential of Teacher Educ ation." Sociol Compass. 12(10):e12628.

107. ^Wilson MB, Kumar T (2017). "Long Ago and Far Away: Preservice Teachers' (Mis)Conceptions Surrounding Racism." Int J Multicult Educ. 19(2):182–198.

Declarations

Funding: No specific funding was received for this work.

Potential competing interests: No potential competing interests to declare.