Uneven Bars? Looking for Environmental Microaggression Effects in NCAA Women's Gymnastics[*]

Tommy Morgan[†] Seth Cannon[‡]

August 7, 2025

Abstract

We study whether environmental microaggressions, a type of racial microaggression, go unaccounted for by coaches of NCAA Division I women gymnasts of color when competing at any given university hosting women's gymnastics meets between 2015–2024. NCAA gymnastics meets provide an excellent setting to test for such an effect because, while scores are assigned individually, teams compete collectively, giving coaches incentives to set lineups optimally. We hypothesize that environmental microaggression effects would manifest as institution-specific racial gaps in scoring and employ a difference-in-differences model that quantifies such gaps, controlling for ability-, preparation-, and event-level factors that also contribute to scores. Across two specifications – first, comparing Black gymnasts to White gymnasts, and second, comparing White gymnasts to all other gymnasts – we find no convincing evidence that gymnasts' scores are affected by an environmental microaggression effect in any way not accounted for by their coaches.

Keywords: women's gymnastics, college sports, racial microaggression theory

JEL Codes: J15, Z2, Z20

[*] The authors are grateful to Zach Flynn for his help in compiling the original data used in this study and also thank Joe Price and Jeff Denning for helpful direction in the early stages of this project. † Corresponding author. Email: tmorga39@vols.utk.edu. ‡ Email: sethbcan@wharton.upenn.edu.

Introduction

According to Sue et al. (2007), *racial microaggressions* are "brief, everyday exchanges that send denigrating messages to people of color because they belong to a racial minority group." In their seminal paper on the topic, the authors identify nine broad categories of racial microaggressions; one category of interest was referred to by Sue et. al. as *environmental microaggressions*, referencing "[m]acro-level microaggressions which are more apparent on systemic and environmental levels." Two examples of this type of microaggression supplied by the authors include "[a] college or university with buildings that are all named after White heterosexual upper class males" and "[t]elevision shows and movies that feature predominantly White people, without representation of people of color." Given that the history of women's artistic gymnastics in America is well-known to be sparse of women of color (Reid, 2024; Wamsley, 2023) and that some NCAA universities have complicated racial histories – for example, Brigham Young University (Bergara, 2013) and several SEC schools (Berry, 2004) – it may be the case that gymnasts of color participating in a historically unrepresentative sport experience environmental microaggressions at certain universities. If this is the case, a coach setting an optimal lineup for a given meet would want to adjust their lineup accordingly but would only be able to do so if they know of the existence of such effects in advance.

NCAA women's artistic gymnastics competitions provide several mechanical advantages to a study of environmental microaggression effects. First, meet winners are decided at the team level, but scores are assigned to routines at the individual level. This individual scoring element allows us to look for an individual-by-host effect (i.e. a gymnast of a certain race doing worse at a certain place) that would be unidentifiable or obfuscated in pure team scoring contexts. Second, the frequency and prevalence of meets across the country provide an exceptionally large sample size across a broad range of competitors. Third, like in other college sports, NCAA gymnastics meets are usually hosted by a specific university at a consistent venue. This allows us to examine each host university as a consistent environment in which environmental microaggression effects may present themselves. Finally, under reasonable assumptions and restrictions to the sample, we can assume that coaches only have incentive to set their lineups to maximize their team's total expected score. This allows us to argue that lineup selections are solutions to a maximization problem faced by lineup setters, which will help parse judging effects from environmental effects.

In order to identify a performance effect based in a racial microaggression, we need to assign a race prediction to each gymnast in our dataset. The NCAA collects self-reported demographic data for all of its student-athletes, but these "ground-truth" self-perceptions of race are not available to the public at the individual level. Instead, we apply the FairFace machine learning model created by Kärkkäinen and Joo (2021) to a headshot photograph of each gymnast and assign them to one of several race categories. This allows us to make a relatively unbiased and consistent assignment of race using a model with excellent out-of-sample performance.

We combine these predictions with a newly assembled dataset of all NCAA women's gymnastics scores from meets occurring between 2015-2024[1] to examine 1) whether Black gymnasts experience an observable *negative* performance effect relative to their White peers and 2) whether White gymnasts experience an observable *positive* performance effect relative to their non-White peers at each of the 64 D1 universities that hosted an NCAA meet over that time period. For each university, we use a differences-in-differences approach to analyze the performance of NCAA gymnasts on **visiting** teams over the first ten meets of any season(s) in which they performed at a meet hosted by that university at least once. Though we initially find significant differential racial gaps in scoring at five universities, they follow no discernable pattern, and none of them survive corrections for multiple hypothesis testing; we discuss our methods and the implications of our findings below.

[1] Scores for the 2025 season became available late in the revision stages of this paper, so they are excluded from our analysis.

Background & Related Literature

NCAA Women's Artistic Gymnastics

We begin with a description of how NCAA women's gymnastics meets are scored that relies heavily on Grimsley and Wright (2019), which is a thorough explanation of scoring in NCAA women's gymnastics written by experienced journalists.[2] Our summary also shares points about the sensitivity of gymnastics scoring with the "Gymnastics" section of Meissner et al. (2021). While our summary does not cite these articles for specific points, they were very helpful to its creation.

In women's artistic gymnastics, a regular season meet is composed of four events: vault, uneven bars, balance beam, and floor exercise. Each performance is scored out of a maximum of 10 points by two judges whose independent scores are averaged to a final performance score. The typical regular season meet has four total judges – two from in-region and two from out of region – judging two events each, with two judges per event. When there are more than two teams at a given meet, at least eight individual judges judge one event each (still with two judges per event) with no rotation of judges between events. Importantly, within a given meet, the set of judges that score each event is constant.

At the NCAA level, scores are determined by two factors: first is the "start value" of the routine, which is the score a gymnast would receive by performing their prepared routine perfectly, with more difficult routines providing higher start values up to the maximum 10 points. Second, deductions are taken from the start value for technical or execution errors observed by the judges during the performance of the gymnast's routine. Though an individual score could range anywhere from zero to 10 in each event, routines are required by rule to have at least a 9.4 point start value, and they score below 8.0 very

[2] Elizabeth Grimsley is the founder and editor-in-chief of College Gym News. Rebecca Wright is the current CNN Politics Photo Editor and a former Photo Editor for The Red & Black, a news organization that covers the University of Georgia.

rarely.[3] Because the practical range of scores is small, tiny differences in average scores separate elite teams from great and decent teams.

In each event, five or six gymnasts from each team perform; if six gymnasts perform, the lowest of those six scores is dropped when calculating the overall team score for that event. Although the maximum number of scorers for each event is capped at six, teams are not limited to six total gymnasts for the entire meet. This means coaches can craft lineups with specialized gymnasts performing in their best events and only deploying gymnasts as "all-arounders" (competing in all four events in one meet) as desired. When all the scoring gymnasts have competed for each of the four events, their scores are summed to compute each team's final meet score.

Teams qualify for playoff competition (in gymnastics, "nationals") based on their Nationals Qualifying Score (NQS). This score is calculated at the end of the regular season (which includes conference championship meets) in four steps: 1) take a team's three highest meet scores at true road meets; 2) take their three highest meet scores from all other meets; 3) drop the highest of these six scores; and 4) average the remaining five scores. At the end of the regular season, every team is ranked according to their NQS, and only a certain number of the top teams qualify for nationals. The qualifying teams are then seeded in their regional tournament based off of their NQS.

The unique attributes of artistic gymnastics meets offer us several key advantages. First, scores are assigned to gymnasts on an individual basis. This allows us to use individual routine scores to look for the presence of an environmental effect that would manifest itself at the individual-by-host level, i.e. a gymnast of a certain race experiencing a drop in performance at a meet hosted by a particular institution. This scoring model makes this individual-by-host analysis straightforward and differentiates this paper from research on other NCAA team sports like basketball and football in which individual

[3] In the set of scores from which we construct each university sample, fewer than 1% are lower than 8.0; the 5th percentile score is 8.9, the 25th percentile 9.575, and the median 9.75.

performance is not always easy to fully isolate from team performance.

Second, our sample size is very large. Previous research investigating behavioral effects using women's gymnastics has primarily focused on elite-level gymnastics competitions, which do not happen as frequently as NCAA meets. These papers most frequently deal with race-agnostic biases present in judges and competitors, finding effects attributable to difficulty bias (Rotthoff, 2020), overall ordering bias (Joustra et al., 2020; Morgan & Rotthoff, 2014; Rotthoff, 2015) and the superstar effect (Meissner et al., 2021) at the highest level of the sport. However, because there are so many fewer elite gymnasts than NCAA gymnasts, these articles can only analyze the performances of hundreds of gymnasts, while our sample includes thousands.

Third, like in other college sports, NCAA gymnastics meets are usually hosted by a specific university at a consistent venue; relatively few are hosted at neutral sites, especially early in the regular season of competition. This also means that institution-hosted NCAA meets differ from elite meets in the consistency of their environment, as elite meets are hosted at various international venues that are not necessarily fixed, with the Summer Olympics being a classic example. This consistency allows us to examine each host university as an environment in which environmental microaggression effects may present themselves.

Finally, using the NQS as a playoff qualification measure offers several incentives to coaches that are relevant to our analysis: most importantly, coaches need their lineups to perform well at home and on the road, making adjusting to competing at other venues a key focus to maximize their score. In addition, any regular season meet could eventually be counted towards a team's NQS, so coaches have an incentive to maximize their expected meet score at every regular season meet. As the end of the regular season approaches, even if a coach is confident their team will qualify with their current NQS, they will still be incentivized to maximize their score to improve their seeding for nationals.

Racial Microaggression Theory

Recent research that investigates the effects of environmental racial microaggressions at the college level is primarily focused on qualitative interviews or surveys of Black students' experiences at predominantly White institutions (PWIs) (Holliday & Squires, 2020; Mills, 2020). This observation is also generally true of literature in this field historically, as evidenced by the many hundreds of papers based on interviewing Black students attending PWIs published from 1965-2013 that are summarized in Willie and Cunnigen (1981), Sedlacek (1987), and Holliday and Squires (2020). Also relevant to research on racial-environmental effects at the college level is Dix's body of work on sports programs at historically Black colleges and universities (or HBCUs) in which he shows teams from HBCUs experiencing negative performance effects while competing against PWIs in football (Dix, 2017, 2021a), men's basketball (Dix, 2022a, 2022b), women's basketball (Dix, 2019, 2020a, 2022b), baseball (Dix, 2020b), softball (Dix, 2021b), and volleyball (Dix, 2023).

Much research also exists on racial biases within the world of professional sports. This research usually focuses on racial biases in referee/judge decisions (Eiserloh et al., 2020; Gallo et al., 2012; Parsons et al., 2011; Pelechrinis, 2023; Price & Wolfers, 2010; Rotthoff, 2020) or in fan/commentator preferences (Andersen & La Croix, 1991; Preston & Szymanski, 2008; Principe & van Ours, 2022; Quansah et al., 2023; Reilly & Witt, 2011). These studies use data from professional sports leagues in many sports and around the world to show that racial biases can affect sports teams and players both in competitive outcomes and perceived value. We contribute to this vein of research on race effects in sports by studying one of its subtypes (environmental microaggression effects) in a novel setting (NCAA gymnastics).

Of particular relevance to this paper is Caselli et al. (2023), in which the authors show that African players in a professional Italian soccer league improved their performance when COVID-19 prevented fans from attending their games. The authors argue that this effect stems from the absence of overtly racist fan behavior, which is common in the league they studied. As in our analysis of gymnasts' performance, the authors evaluated individual-level performance scores (in this case, aggregate performance scores assigned algorithmically to individual soccer players based on in-game contributions) in a generalized fixed effects model that allows them to control for player- and match-based fixed effects. They model the effects of the *removal* of racial aggressions that were directed towards athletes, while we analyze the *introduction* of athletes to a potential environmental microaggression. We add to what Caselli et. al. found for professional athletes by estimating site-specific racial scoring gaps for college-level athletes.

Methods

Data

RoadtoNationals.com has been the official statistical and rankings website of NCAA Women's Artistic Gymnastics since the summer of 2015 (Fredericks & Fredericks, 2013). It has been used as an accessible source for NCAA scoring and team ranking data in existing literature, as in Xiao (2022), Van Dyke et al. (2020), and Law (2020). To our knowledge, our dataset is the first comprehensive pre-processed compilation of these scores, as the data is not readily available for download at its source. Our dataset includes the 230,088 scores received by 4,720 gymnasts over all 3,580 meets across all three NCAA divisions over the 2015-2024 seasons. We make our full dataset of NCAA women's gymnastics scores and all code used for the analysis in this paper available for future use.[4]

In addition to collecting data on scores, we also need to assign a race to each gymnast. Since we do not have access to the *individual-level* data that each gymnast

[4] The data associated with this paper will be made available via ICPSR repository. Our full dataset of scores, including 2025 data, will be made available at a GitHub repository after publication. At that point, this footnote will be updated to reflect that fact.

reports to the NCAA, we use the FairFace race prediction computer vision model created by Kärkkäinen and Joo (2021) to predict each gymnast's race. In order to apply the model, we collect a headshot photograph of each individual gymnast in our dataset; these consist of official photos from their university team's website for the vast majority of gymnasts and comparable photos obtained from news articles or social media when photos were unavailable from official sources. After collecting the photos, we put each of them through the FairFace prediction model to classify each gymnast into one of seven race categories: White, Black, Indian, East Asian, Southeast Asian, Middle Eastern, and Latina. We then align our categories with reported NCAA categories as far as possible, arriving at a final set of four aligned categories: White, Black, Latina, and Other.

We are aware that assignment of race based on visual features, even by computer algorithm, subjects us to the "eye of the beholder" problem discussed in Fort and Gill (2000). For this reason, we compare our classifications to the *aggregated* racial demographics data provided by the NCAA (National Collegiate Athletic Association, 2018) in Table 1.

It should be noted that the NCAA database includes all registered student-athlete gymnasts, whereas our data only includes those who competed and received at least one score in a given year. Even with this caveat, FairFace predicts many more gymnasts in our sample as Latina than are reported in the NCAA database. This is likely because FairFace only identifies gymnasts as a single race when they may identify in the NCAA demographics as Two or More Races or Unknown; this is especially likely to complicate the counts for the Latina category due to the complex race vs. ethnicity issue common to these kinds of classification exercises. The comparisons in Table 1, especially our overprediction of the Latina category, motivate our decision to estimate our model in only two specifications: 1) comparing Black gymnasts to White gymnasts, looking for an environmental microaggression effect; and 2) comparing White gymnasts to all other (i.e. not White) gymnasts, looking for a sort of environmental micro-privilege effect.

Table 1

Comparing predicted & self-reported racial demographics

	Sample Year									
Race	2015	2016	2017	2018	2019	2020	2021	2022	2023	2024
Panel A: Scorers in our sample (predicted race)										
Black	7.6%	7.6%	7.7%	8.4%	9.0%	9.8%	9.6%	9.4%	10.1%	10.6%
	(88)	(89)	(89)	(100)	(107)	(114)	(89)	(119)	(128)	(142)
Latina	4.4%	5.6%	6.2%	6.7%	7.1%	7.2%	8.2%	9.0%	9.6%	8.5%
	(51)	(65)	(72)	(80)	(84)	(84)	(76)	(114)	(122)	(114)
Other	11.21%	11.3%	11.3%	12.1%	12.9%	12.0%	12.0%	13.3%	12.8%	11.5%
	(130)	(132)	(131)	(144)	(153)	(140)	(112)	(169)	(163)	(154)
White	76.8%	75.5%	74.9%	72.8%	71.0%	71.0%	70.3%	68.4%	67.5%	69.4%
	(891)	(882)	(870)	(865)	(843)	(828)	(654)	(868)	(857)	(931)
Unique Gymnasts	1,160	1,168	1,162	1,189	1,187	1,166	931	1,270	1,270	1,341
Panel B: All enrolled gymnasts (self-reported race)										
Black	7.8%	7.9%	7.7%	8.0%	8.2%	8.5%	8.4%	7.7%	8.1%	7.7%
	(116)	(118)	(117)	(123)	(127)	(129)	(133)	(131)	(139)	(137)
Latina	3.9%	4.4%	4.3%	3.8%	3.2%	3.6%	4.0%	4.6%	5.3%	6.2%
	(58)	(66)	(66)	(59)	(50)	(55)	(64)	(78)	(91)	(111)
Other	15.7%	16.4%	17.9%	18.0%	20.4%	20.4%	19.9%	23.8%	22.9%	21.8%
	(234)	(246)	(273)	(279)	(315)	(309)	(316)	(407)	(392)	(390)
White	72.7%	71.4%	70.1%	70.2%	68.1%	67.5%	67.7%	63.9%	63.7%	64.3%
	(1,084)	(1,072)	(1,067)	(1,086)	(1,050)	(1,022)	(1,073)	(1,091)	(1,093)	(1,147)
Unique Gymnasts	1,492	1,502	1,523	1,547	1,542	1,515	1,586	1,707	1,715	1,785

Note: Percentages of the total count of unique gymnasts from a given year with a given predicted/self-reported race are reported, with actual counts in parentheses.

Sample Construction

We begin by narrowing our sample to a subset of scores that meet four criteria: First, we drop scores from meets hosted at neutral sites (i.e. without a specific host university). Second, we drop meets with special titles like "SLC Regional", "John Zuerlein Invite", and "Big 12 Championships". Third, we remove scores received by gymnasts who are not competing for Division I (DI) schools. Fourth, in our Black-White comparisons, we drop scores from gymnasts that FairFace predicts are not Black or White (in our White-not White comparisons, we drop no scores for this step). But why narrow the sample in such a way?

First, we exclude titled meets in order to completely exclude playoff and invitational meets. Playoff meets, even those that can count towards the NQS, can change incentives from the usual score maximization incentive[5] and present a higher pressure environment than regular meets, representing another change that could muddy our key across-university comparisons. Invitational meets are also not conducive to our comparisons, as they are frequently hosted by organizations, not universities, which is a further change to the regular meet environment that we wish to exclude.

Second, we exclude any remaining meet that is hosted at a neutral site in order to make the context surrounding the scores in our sample as similar as possible. Non-playoff neutral site meets are hosted away from any university's usual venue(s) by definition; therefore, they create a different type of environment than a typical regular season meet would have. This would confound the across-university comparison that is the focus of our research question, so we remove these meets.

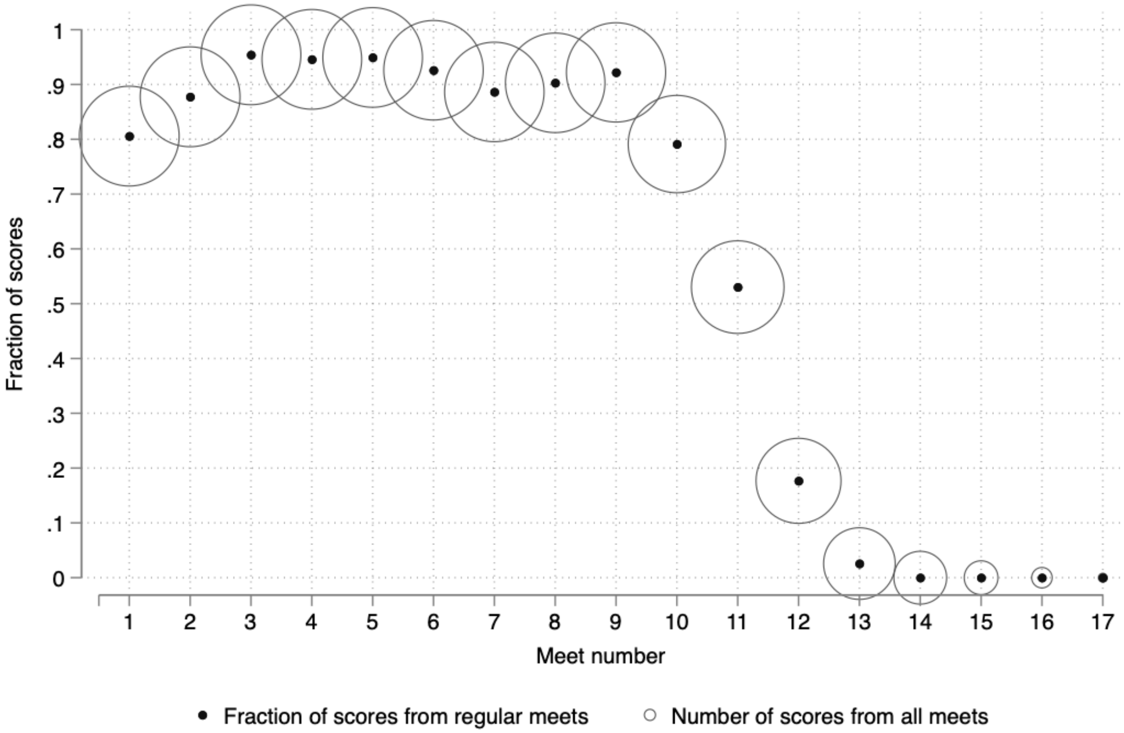
Figure 1 shows the fraction of the total set of scores that are from meets that survive our first and second sample narrowing criteria, with outer rings showing the relative number of scores at that meet number. As seen in the figure, about 80% of teams' first meets (as in meet number one of the season) are ordinary meets whose scores remain in our sample, whereas no team has a single regular meet beyond their 13th meet. Under these criteria, we drop mostly later meets with fewer scores under these criteria, and Figure 1 shows that we keep the vast majority of scores in our sample universe.

After applying the first two sample narrowing criteria, we further limit our sample to scores received by DI gymnasts. Table 2 illustrates average scoring by event, division, and predicted race, and it shows that DI gymnasts are much better and more consistent

[5] A gymnast's incentive to maximize her score may change if, for example, she is the sixth to compete on the uneven bars and she knows her team will qualify for the next round as long as her score is 9.75 or higher. In that case, she might adjust her planned routine to make it less likely she commits a major error instead of pushing for her highest possible score, which is the exact change in incentives we avoid by excluding playoff meets from our sample.

Figure 1

Fraction of scores from "regular" meets by meet number.



Note: A "regular" meet is a meet hosted by a specific university that does not have a special title like "invitational" or "regionals".

scorers both within and across events than their non-DI counterparts. We also motivate our decision to include only DI gymnasts' scores with Figure 2. The top panel of Figure 2 shows the season-long trend in scoring when including all gymnasts in our sample across all three NCAA divisions (Divisions I, II, and III), and it clearly shows that White gymnasts catch up to their not-White and Black peers as the season progresses. However, when we limit the sample to only DI gymnasts as in the bottom panel of Figure 2, we see that gymnasts of all predicted races improve at nearly the same rate throughout the season. This same-rate improvement is consistent with the parallel trends assumption central to our differences-in-differences design, so keeping only the DI scores properly enables our identification strategy.

After applying all of the above criteria, we further narrow the sample each time we

Table 2

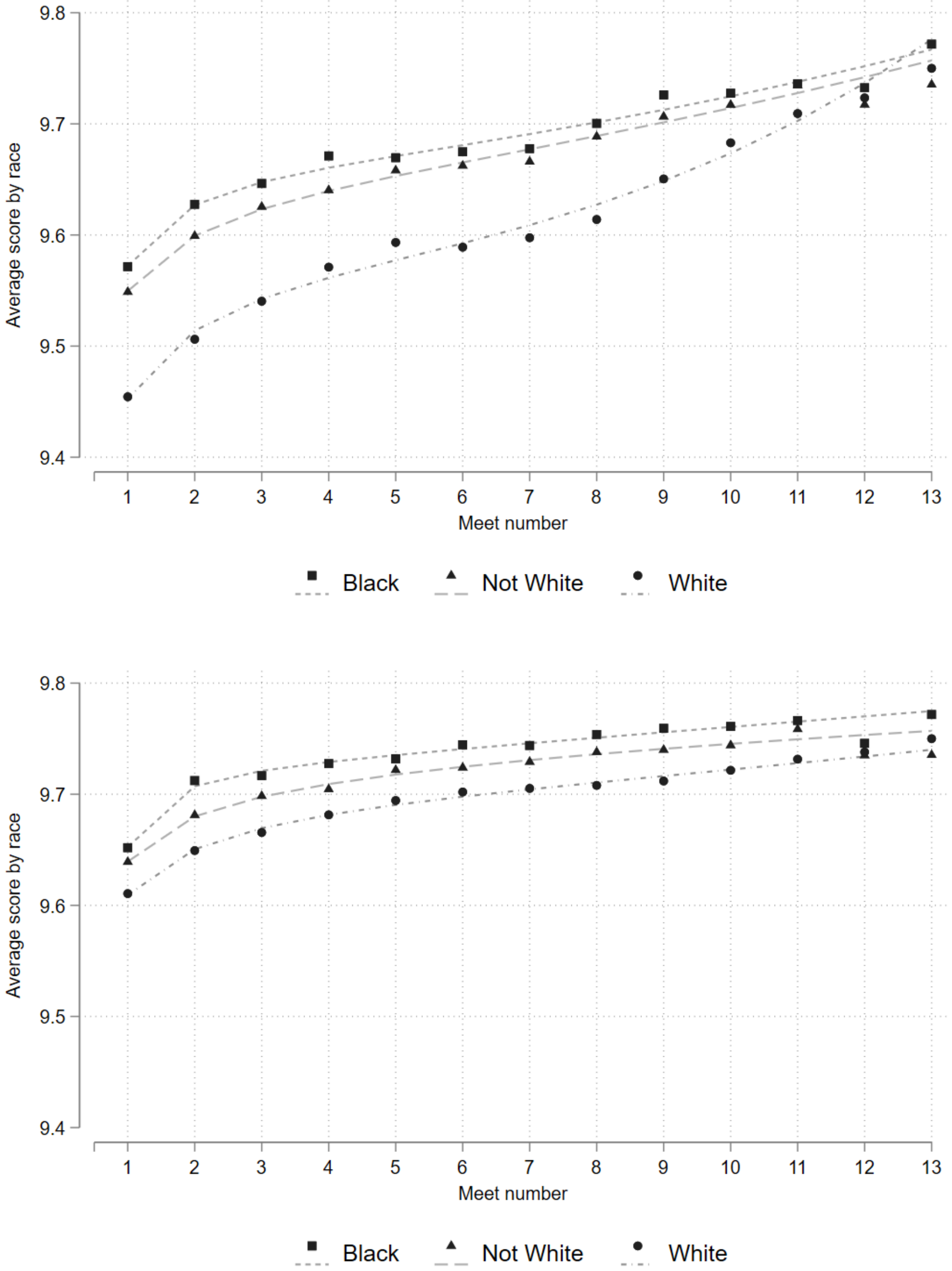
Average score by event, NCAA division, and predicted race

	Vault	Bars	Beam	Floor	Overall
Panel 1: All gymnasts					
White	9.64	9.52	9.55	9.63	9.58
	(0.26)	(0.50)	(0.41)	(0.37)	(0.40)
Black	9.71	9.62	9.61	9.72	9.67
	(0.24)	(0.43)	(0.36)	(0.32)	(0.34)
All gymnasts	9.66	9.54	9.57	9.65	9.61
	(0.25)	(0.48)	(0.39)	(0.36)	(0.38)
Panel 2: Only D1 gymnasts					
White	9.72	9.66	9.66	9.72	9.69
	(0.18)	(0.36)	(0.30)	(0.30)	(0.30)
Black	9.76	9.70	9.68	9.77	9.73
	(0.18)	(0.35)	(0.29)	(0.27)	(0.28)
All gymnasts	9.73	9.67	9.67	9.73	9.70
	(0.18)	(0.36)	(0.30)	(0.30)	(0.29)
Panel 3: Only non-D1 gymnasts					
White	9.39	9.09	9.22	9.38	9.27
	(0.30)	(0.60)	(0.49)	(0.42)	(0.48)
Black	9.44	9.16	9.28	9.46	9.35
	(0.29)	(0.55)	(0.46)	(0.45)	(0.45)
All gymnasts	9.39	9.10	9.23	9.38	9.28
	(0.30)	(0.59)	(0.49)	(0.42)	(0.48)

Note: Scores from titled or neutral meets are not included in calculations for this table.

Figure 2

Average Score by Race and Meet Week Number



Note: Not White also includes Black. Top figure includes all gymnasts; bottom includes only D1 gymnasts. Lines represent observation count-weighted fractional polynomials of best fit. Scores from titled or neutral meets are not included.

estimate our model for a given university. Explaining which meets we include in our sample for each university is easiest via an example, so suppose we take the University of Alabama as a host university. The University of Georgia women's gymnastics team performed at Alabama every other year beginning in 2016 through 2024, so we include in the University of Alabama sample every score from every Georgia gymnast at every one of their "regular" meets from those five years (2016, 2018, 2020, 2022, and 2024) in our sample. In contrast, the Alabama sample only includes University of Denver scores from 2019, as this is the only season in which Denver visited Alabama over our 2015-2024 time span. We repeat this process for every team that visited Alabama over that time period, collecting scores from meets in seasons in which they visited Alabama to eventually build the full Alabama sample as depicted in Table 3. We then estimate the models we describe below for Alabama and then repeat this sample building and estimation process until we have done so for each of the 64 DI universities to have hosted a meet in that time span.

Table 3

Team-seasons included in the University of Alabama sample

Team	2015	2016	2017	2018	2019	2020	2021	2022	2023	2024
Arizona	X									
Arkansas		X		X		X		X		X
Auburn	X		X		X		X		X	
Boise State	X		X						X	
Bowling Green					X					
Denver					X					
Florida	X		X		X		X		X	
Georgia		X		X		X		X		X
Illinois										X
Iowa State			X							
Kentucky		X		X		X	X	X		X
LSU	X		X		X		X		X	
Michigan					X					
Michigan State									X	
Minnesota										X
Missouri		X		X		X		X		X
North Carolina				X				X		
Northern Illinois					X					
Oklahoma	X			X		X				
S.E. Missouri					X					
Temple					X					
West Virginia		X								
Western Michigan								X		
Scores by Year	1,414	1,173	1,078	1,441	2,104	1,079	768	1,318	1,127	1,390
Total Sample Size	12,892									

Note: X denotes a year in which the scores from a team on a given row are included in the Alabama sample.

Model

We know the actual process by which a gymnast receives a score: 1) she works with her coaches to prepare a routine that can maximize the quality of her expected performance based on her ability and preferences; 2) she is selected by her coach to perform in the meet lineup; 3) she submits her routine to the judges (receiving the "start value" of her routine); 4) she performs her routine for the judges (for which they may issue deductions); and 5) she receives her final score from the judges. We can model these decisions with the following set of functions:

$$\text{routine}_{gu} = r(\text{ability}_g,\ \text{preferences}_g,\ \text{coaching}_u) \tag{1}$$

where gymnast g and the coaching staff at her university u select and prepare for a routine (which will have a certain start value) based on her ability and preferences and the coaching staff's preferences/specialties;

$$\text{performance}_{gutv} = g(\text{routine}_{gu},\ \text{experience}_{gt},\ \text{venue}_v,\ \text{diff. venue effects}_{gv}) \tag{2}$$

where gymnast g's overall level of performance at the event during the meet is a function of her selected routine gu, her level of experience gt both within the season and across her career, any venue effects v that are consistent across gymnasts and meets (such as venue altitude), and any differential effects of the venue on her gv (such as if she has a particularly adverse reaction to a higher altitude); and

$$\text{score}_{gutvj} = j(\text{performance}_{gutv},\ \text{judge tendencies}_j,\ \text{judge bias}_{gj}) \tag{3}$$

where the final score is a function of the gymnast's performance $gutv$, any gymnast- and meet-constant tendencies of judge j (such as one judge always being stricter than another), and any gymnast-specific judge biases gj (such as one judge normally giving higher scores to shorter gymnasts).

The above three equations collectively cover every step of the scoring process except for step two, lineup selection. Given our sample restrictions as described above, we assume that coaches have no other incentive in the meets in our sample than to maximize their potential NQS by maximizing their expected team score at each meet. This means that, for each meet, they dynamically adjust their lineup by selecting routines (via assigning lineup slots to gymnasts) to solve the following problem:

$$\max_{\{\text{routines}\}} \sum_{events} \left(\mathbb{E}\left[\sum_{i=(n-4)}^{n} \left(\text{score}_{(i)}\right) \right] \right) \quad \text{s.t.} \quad n \in \{5, 6\} \tag{4}$$

where $\text{score}_{(i)}$ is the ith order statistic, or the ith lowest score received in a given event at a given meet, with the sum from the $(n-4)$th to the nth lowest scores representing the sum of the team's five highest scores in each event. In simpler terms, coaches select gymnasts and routines across each event to maximize the total score of each event's top five scorers out of either five or six participants.

By assuming that coaches solve Equation 4 when setting their lineups and coaching their gymnasts, we assume that these agents aim to maximize the score a gymnast will receive in a given event by selecting routines of appropriate difficulty and practicing them adequately; this implies that a gymnast and her coach have perfect information about the ability-, preparation-, and event-related elements of our proposed scoring model, and can therefore select the optimal routines and lineup. We also assume that, although they have no control over the venue- and judge-related factors of scoring, coaches account for and respond optimally to their expectations of the effects of those factors.

If we assume that coaches have a reasonable knowledge of salient factors that might affect their gymnasts (such as a venue's altitude or a judge's bias) but no knowledge of more obscure factors like potential race-venue interaction effects, then we can estimate a model of scoring with a race-venue interaction term that is unconfounded by any other gymnast-by-venue effects, including judge effects and biases. This would be a natural result of optimal lineup setting behavior in which coaches correctly adjust lineups based on things they know about (judges, venues, their gymnasts, etc.) but fail to adjust optimally to a more obscure type of effect they do not expect. Determining whether lineup setters fail to account for such an effect at the race-by-venue level is the focus of this paper.

Estimation Strategy

We estimate our full model of a gymnast's score using a difference-in-difference design augmented with fixed effects. We have two across-race comparison specifications: first, we compare Black gymnasts to White gymnasts, and then we compare White gymnasts to all non-White gymnasts. For our Black gymnasts to White gymnasts model, it takes the following form:

$$\begin{aligned}\text{score}_{gmue} = \alpha &+ \beta(\text{Black*atHost})_{gm} \\ &+ \gamma(\text{season meet number})_{um} + \mu(\text{career meet number})_{gm} \\ &+ [\text{team}]_u + [\text{gymnast}]_g + [\text{event-by-meet}]_{em} + u_{gmue}\end{aligned} \tag{5}$$

where subscripts g, m, u, and e refer to individual gymnasts, meets, teams, and events (vault, bars, beam, and floor), respectively. When we estimate our White-not White specification, we replace the Black indicator with an indicator for being White and adjust other variables accordingly. The dependent variable is the score earned by a gymnast performing in a given event at a given meet, and the interaction term $(\text{Black*atHost})_{gm}$ takes a value of 1 if the observation is a score received by a Black gymnast competing at a given host university and 0 otherwise. Black_g and atHost_m represent binary variables for a gymnast being Black and a meet being held at the focus university. We include a control for a team's season meet number ($\text{season meet number}_{um}$) to control for scoring trends observed in Figure 2, and a control for a gymnast's career meet number $(\text{career meet number})_{gm}$ to control for increasing experience over the course of a career. We also include fixed effects for 1) the team a given gymnast is on, denoted as $[\text{team}]_u$; 2) each individual gymnast herself, denoted as $[\text{gymnast}]_g$; and 3) each individual event at each individual meet, denoted as $[\text{event-by-meet}]_{em}$. We cluster our standard error u_{gmue} at the event-by-meet level. In this model, α is the regression constant term, and the coefficient of interest is β, which can be interpreted as a difference-in-differences estimate, or the differential impact of competing at the focus university on Black gymnasts relative to other venues compared to the same difference for White gymnasts.

The gymnast fixed effects control for every event-, meet-, and time-invariant characteristic unique to a given gymnast, including her physical characteristics such as strength and height, her training before university, and her innate talent for gymnastics. The event-by-meet fixed effects – four indicators for each unique meet, i.e. $\mathbb{1}$(Vault at Utah State, 9 January 2015), $\mathbb{1}$(Balance Beam at Utah State, 9 January 2015), etc. – control for unchanging venue characteristics like geography, climate, and altitude, but they also control for every gymnast-, meet-, and time-invariant tendency of that event's assigned judges. Within a given meet, the judges giving scores in a single event are constant: all scores in the floor exercise from the meet hosted by Utah State on 9 January 2015 are assessed by the same pair of judges. As a result, not only do the event-by-meet fixed effects capture event- and meet-constant characteristics, such as the date of the meet or the particular nuances of the floor exercise compared to the vault; they also control for the constant characteristics of the assigned judge pair, because the judges giving the scores are an event-by-meet-constant characteristic.

For a causal interpretation of β to be possible, we would need to assume that Black (White) gymnasts performing at a given host university would experience the same relative change in performance – measured by a change in score – at that venue as their White (not White) counterparts would in the absence of any environmental effect of that venue on Black (White) gymnasts' scoring. Since different teams visit different hosts at different points in the season throughout our dataset, we scrutinize our assumption with the bottom panel of Figure 2, in which we plot average scores in our sample by race and meet number. The figure shows that both Black, White, and non-White gymnasts see their scores increase over the course of a season on average, and that this trend is parallel across race groups. This figure demonstrates the baseline viability of the assumption needed for our estimates to be interpreted as the causal effect of race-venue factors on gymnasts' scores.

We follow this baseline inspection by imposing some additional assumptions that make this condition more likely to hold. First, we assume that coaches setting optimal lineups can account for judge biases. This would allow us to parse out the venue-race effect from a potential judge-race effect by assuming any judge-race effect has been accounted for in the lineup setter's maximization problem. We believe this to be a reasonable assumption; our assumed coach would be well-informed enough to know which judges would likely be present at a given meet and understand the reputations of those judges as it relates to their scoring tendencies[6]. Second, we assume that these kind of judge effects are minimally influential on gymnasts' scores. We

a;lsdkfj;laksjdf;lkjas;ldkfj;askj;dfl

Third, we assume that all other relevant factors affecting the differences described above are controlled for by including the covariates we list above in our estimating equation. These conditions feel reasonable and make a conditional parallel trends condition more likely to hold.

We interpret a given host university's β estimate as a gymnast-at-venue-level effect. If that effect is statistically significant at a given university, it could be due to an environmental microaggression factor like the name of a gymnasium or a predominantly non-Black student body, or it could be some other factor at a given venue affecting Black gymnasts for some reason. If we were to see negative β estimates in our Black-White comparison (or positive estimates in our White-not White comparison) within certain sets of universities – such as those included in a web article or Twitter thread compiling gymnasts speaking out against racism within their teams (e.g. Duffy (2020) or Boswell (2020)) – then we may reasonably conclude that an environmental microaggression effect is present for Black gymnasts at that subset of universities. However, if universities with

[6] i.e. "Judge A is strict on handstands", "Judge B is loose on artistic requirements", "Judge C might be biased against a certain type of routine", and so on

statistically significant β estimates do not follow a noticeable trend, we might instead attribute those results to statistical noise, especially if those estimates appear to follow predictable probability distributions and do not survive corrections for multiple hypothesis testing.

Results

Table 4 lists the universities in our sample for which estimating Equation 5 returned estimates of β significant at the 95% confidence level. By estimating Equation 5 for each individual university, we test whether gymnasts who were chosen to perform at meets at that venue perform at a different level than the level at which they would otherwise be expected. Due to our comprehensive set of fixed effects, any estimates of β we obtain will be *ceterus paribus* in terms of judge, gymnast, and team influences. As such, any estimate of β that is statistically different than zero would imply only that gymnasts of a given race are affected at the given venue in a way that they and their coaches did not forsee when lineups for that meet were set.

Of the 64 DI host universities in each specification, only two have a significant Black-White coefficient, while three have significant White-not White coefficients. There does not appear to be any pattern to which universities return these statistically significant estimates; they are not concentrated in any one geographic region, and most of them are not contained in Boswell's compilation Twitter thread or Duffy's article (Boswell, 2020; Duffy, 2020). In addition, only two estimates in the entire table (UC Davis and Towson) show an estimate with a sign opposite the direction of the gap that Table 2 shows already exists: White gymnasts are already the lowest average scorers, and Black gymnasts already the highest. Given these results, we find it unlikely that there is an environmental microaggression effect that goes unaccounted for by lineup setters in the NCAA. But how should we interpret the few statistically significant results we do find?

By using 95% confidence intervals as our judge for the statistical difference of β

Table 4

Universities with significant interaction term estimates

University	Estimate	(St. Error)	p-value	Sample Size
Comparing Black gymnasts to White				
Pittsburgh	0.058	(0.024)	0.018	11,643
UC Davis	-0.073*	(0.034)	0.032	8,304
Comparing White gymnasts to not White				
Alabama	-0.032	(0.014)	0.022	12,892
Towson	0.042*	(0.021)	0.040	15,089
Washington	-0.037	(0.017)	0.028	11,338

Note: Teams with significant estimates for each specification are included in alphabetical order.
**Sign of estimate indicates potential race-venue effect (opposite of expected gap).*

from zero, we necessarily subject ourselves a 5% Type I error rate, meaning we expect to estimate a truly zero effect as statistically different from zero once in twenty tries. Since our testing is also two-sided, we would expect to estimate a truly zero effect as statistically greater than zero once in forty tries, and likewise in the opposite direction. If the null hypothesis of a true-zero effect held across all 64 universities in our sample, we would nonetheless expect to see one or two instances of statistically negative effects and one or two schools with statistically positive effects. In addition, because our sample already has an existing scoring gap, we would expect to see many more significant positive (negative) estimates in our Black-White (White-not White) specification, and fewer in the opposite direction; indeed, this is the case, with only two of the five significant estimates being in a direction opposite of the existing gap. As such, we argue that the few statistically significant estimates we observe in Table 4 are nothing more than what we would expect from random chance.

Given that we think our significant estimates are the result of random chance, we might ask a natural follow-up question: would a conservative adjustment to our standard errors designed to account for the multiple hypothesis tests we perform leave any school with a statistically significant estimate for β? After all, if significant results are the result of our method of statistical inference allowing too high of a Type I error rate (false rejection of the no effect null hypothesis), then a correction or adjustment designed to prevent Type 1 errors may help clarify our results. Following this logic, we apply the Bonferroni correction as established in Dunn (1961) and discussed in Armstrong (2014) and VanderWeele and Mathur (2019), in which a p-value is seen as significant only if it remains below 0.05 after being multiplied by the number of tests performed. Our model is estimated for 64 schools, so any p-value that survives the correction would have to be smaller than $\frac{0.05}{64} = 0.00079$. Seeing that there are no p-values below 0.017 in Table 4 makes it clear that none of our estimates survive the correction, supporting our claim that the estimates are significant only by chance.

Discussion & Conclusion

We test whether Black gymnasts experience a change in performance that their White competitors do not experience that goes unforeseen by their coaches when competing at 64 Division I NCAA universities. Our comprehensive dataset of women's gymnastics scores allows us to isolate racial scoring gaps at a given host university by controlling for the influences of individual events, gymnasts, meets, and judges through a series of fixed effects and reasonable model assumptions. While we do initially find a few significant differences in score distributions using our model that could be attributed to environmental interaction effects, they do not follow any predictable pattern, and none of them remain significant after adjusting for multiple hypothesis testing. This leads us to conclude that the evidence is inconsistent with the idea that any specific host university negatively affects Black or positively affects White gymnasts' performance to a notable degree beyond what lineup setters expect.

Our findings challenge expectations that systemic or environmental factors associated with universities with complicated racial histories might lead to measurable disparities in agents' performance at those universities. Instead, it may indicate that NCAA gymnastics environments, characterized by standardized scoring practices and strict competition protocols, limit the avenues through which such effects could manifest. They could also suggest that, in the case that such effects actually do exist, agents with knowledge of the effects can adjust to them optimally.

We acknowledge that measuring the possible impact of environmental microaggressions only by examining performance scores is incomplete and overlooks the lived experiences the athletes might face by visiting these universities. We stress that the absence of measurable performance effects does not eliminate the possibility that experiencing environmental microaggressions influences gymnasts in other ways. They may affect other dimensions of gymnasts' experiences, such as mental well-being or feelings of inclusion, which are beyond the scope of this study. Continuing to explore these aspects through qualitative methods could provide valuable context to complement our quantitative analysis.

Within this analysis, we also rely heavily on the FairFace computer vision model for race predictions, which imposes additional limitations. Though FairFace is known for its strong out-of-sample performance, it still relies on visual cues to assign race, which may not align with how individuals self-identify. With access to that self-reported data, future research could examine this question using self-identified race variables to better isolate differential environmental effects on athletes of many different races.

References

Andersen, T., & La Croix, S. J. (1991). Customer racial discrimination in Major League Baseball. *Economic Inquiry*, *29*(4), 665–677. https://doi.org/10.1111/j.1465-7295.1991.tb00853.x.

Armstrong, R. A. (2014). When to use the Bonferroni correction. *The Journal of the College of Optometrists*, *34*, 502–508. https://doi.org/10.1111/opo.12131.

Bergara, G. J. (2013). "This time of crisis": The race-based anti-BYU athletic protests of 1968-1971. *Utah Historical Quarterly*, *81*(2), 204–229. https://doi.org/10.2307/45063320.

Berry, M. S. (2004). Leveling the playing field: African-Americans and collegiate athletics. *EIU Historia*, *13*, 113–128. https://www.eiu.edu/historia/Berry.pdf.

Boswell, S. F. (2020). Black gymnasts sharing their experiences of racism in the sport, a thread. *Twitter: @sf_boswell.* https://web.archive.org/web/20220128164941 /https://twitter.com/sf_boswell/status/1270158886381764610.

Caselli, M., Falco, P., & Mattera, G. (2023). When the stadium goes silent: How crowds affect the performance of discriminated groups. *Journal of Labor Economics*, *41*(2), 431–451. https://doi.org/10.1086/719967.

Dix, A. (2017). A decade of referee bias against college football programs from historically Black colleges and universities. *International Journal of Science Culture and Sport*, *5*(3), 197–212.

Dix, A. (2019). "And 1" more piece of evidence of discrimination against Black basketball players. *Howard Journal of Communications*, *30*(3), 211–229. https://doi.org/10.1080/10646175.2018.1491434.

Dix, A. (2020a). And 10 more years of bias against HBCU female basketball players. *Texas Speech Communication Journal*, *44*, 1–18.

Dix, A. (2020b). Critical race theory, the NCAA, and college baseball: Contradiction on the diamond. In M. Milford & L. Reichart Smith (Eds.), *Communication and contradiction in the NCAA: An unlevel playing field* (pp. 213–234). Peter Lang.

Dix, A. (2021a). Referee judgments of communication in the field of play: A study on historically Black colleges and universities in Division II college football. *International Journal of Sport Communication*, *14*(4), 554–573. https://doi.org/10.1123/ijsc.2021-0032.

Dix, A. (2021b). Softball umpires call more walks than strikeouts when the pitcher plays for a historically Black college and university. *International Journal of Sport Culture and Science*, *9*(1), 91–103. https://web.archive.org/web/20231214222019/https://dergipark.org.tr/en/download/article-file/1412460.

Dix, A. (2022a). The non-Sweet Sixteen: Referee bias against historically Black colleges and universities in men's college basketball. *Sociology of Sport Journal*, *39*(1), 118–124. https://doi.org/10.1123/ssj.2020-0187.

Dix, A. (2022b). Stay woke: An analysis of how referees evaluate the in-game communication of an HBCU that competes in a PWI conference. *Communication and Sport*, *OnlineFirst*. https://doi.org/10.1177/21674795221103407.

Dix, A. (2023). Indications of referee bias in Division I women's college volleyball: Testing expectancy violations and examining nonverbal communication. *International Journal of Sport Communication*, *16*(4), 414–422. https://doi.org/10.1123/ijsc.2023-0050.

Duffy, P. (2020). Multiple NCAA gymnastics teams accused of racism by former athletes. *Gymnastics Now*. https://web.archive.org/web/20240310002706/https://gymnastics-now.com/multiple-ncaa-gymnastics-programs-accused-of-racism.

Dunn, O. J. (1961). Multiple comparisons among means. *Journal of the American Statistical Association*, *56*(293), 52–64. https://doi.org/10.1080/01621459.1961.10482090.

Eiserloh, D. G., Foreman, J. J., & Heintz, E. C. (2020). Racial bias in National Football League officiating. *Frontiers in Sociology*, *5*(48). https://doi.org/10.3389/fsoc.2020.00048.

Fort, R., & Gill, A. (2000). Race and ethnicity assessment in baseball card markets. *Journal of Sports Economics*, *1*(1), 21–38. https://doi.org/10.1177/152700250000100103.

Fredericks, M., & Fredericks, J. (2013). Road to Nationals - NCAA gymnastics rankings. Accessed 12 June 2024 at https://roadtonationals.com.

Gallo, E., Grund, T., & Reade, J. J. (2012). Punishing the foreigner: Implicit discrimination in the Premier League based on oppositional identity. *Oxford Bulletin of Economics and Statistics*, *75*(1), 136–156. https://doi.org/10.1111/j.1468-0084.2012.00725.x.

Grimsley, E., & Wright, R. (2019). Gymnastics 101: What to know about scoring, rankings and more before the next GymDog meet. *The Red and Black*. https://web.archive.org/web/20230527011633/https://www.redandblack.com/sports/gymnastics-101-what-to-know-about-scoring-rankings-and-more-before-the-next-gymdog-meet/article_1008352e-56bd-11e2-b46e-0019bb30f31a.html.

Holliday, N. R., & Squires, L. (2020). Sociolinguistic labor, linguistic climate, and race(ism) on campus: Black college students' experiences with language at predominantly White institutions. *Journal of Sociolinguistics*, *25*(3), 418–437. https://doi.org/10.1111/josl.12438.

Joustra, S. J., Koning, R. H., & Krumer, A. (2020). Order effects in elite gymnastics. *De Economist*, *169*, 21–35. https://doi.org/10.1007/s10645-020-09371-0.

Kärkkäinen, K., & Joo, J. (2021). Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 1548–1558.

Law, J. (2020). A comparison of team effectiveness and perception of team success in academic and athletic teams. *Oregon State University Honors College Theses.*

Meissner, L., Rai, A., & Rotthoff, K. W. (2021). The superstar effect in gymnastics. *Applied Economics*, *53*(24), 2791–2798. https://doi.org/10.1080/00036846.2020.1869170.

Mills, K. J. (2020). "It's systemic": Environmental racial microaggressions experienced by Black undergraduates at a predominantly White institution. *Journal of Diversity in Higher Education*, *13*(1), 44–55. https://doi.org/10.1037/dhe0000121.

Morgan, H. N., & Rotthoff, K. W. (2014). The harder the task, the higher the score: Findings of a difficulty bias. *Economic Inquiry*, *52*(3), 1014–1026. https://doi.org/10.1111/ecin.12074.

National Collegiate Athletic Association. (2018). NCAA demographics database. [https://www.ncaa.org/sports/2018/12/13/ncaa-demographics-database.aspx. Accessed 15 December 2024, archived at http://archive.today/K7QD4.]. *NCAA.org*.

Parsons, C. A., Sulaeman, J., Yates, M. C., & Hamermesh, D. S. (2011). Strike three: Discrimination, incentives, and evaluation. *American Economic Review*, *101*(4), 1410–1435. https://doi.org/10.1257/aer.101.4.1410.

Pelechrinis, K. (2023). Quantifying implicit biases in refereeing using NBA referees as a testbed. *Scientific Reports*, *13*. https://doi.org/10.1038/s41598-023-31799-y.

Preston, I., & Szymanski, S. (2008). Racial discrimination in English football. *Scottish Journal of Political Economy*, *47*(4), 342–363. https://doi.org/10.1111/1467-9485.00168.

Price, J., & Wolfers, J. (2010). Racial discrimination among NBA referees. *The Quarterly Journal of Economics*, *125*(4), 1859–1887. https://doi.org/10.1162/qjec.2010.125.4.1859.

Principe, F., & van Ours, J. C. (2022). Racial bias in newspaper ratings of professional football players. *European Economic Review*, *141*. https://doi.org/10.1016/j.euroecorev.2021.103980.

Quansah, T. K., Lang, M., & Frick, B. (2023). Color blind - Investigating customer-based discrimination in European soccer. *Current Issues in Sport Science*, *8*(2), 007. https://doi.org/10.36950/2023.2ciss007.

Reid, T. (2024). Making Black history in NCAA gymnastics. *College Gym News*. https://collegegymnews.com/2024/02/20/making-black-history-in-ncaa-gymnastics.

Reilly, B., & Witt, R. (2011). Disciplinary sanctions in English Premiership Football: Is there a racial dimension? *Labour Economics*, *18*(3), 360–370. https://doi.org/10.1016/ j.labeco.2010.12.006.

Rotthoff, K. W. (2015). (Not finding a) sequential order bias in elite level gymnastics. *Southern Economic Journal*, *81*(3), 724–741. https://doi.org/10.4284/0038-4038-2013.052.

Rotthoff, K. W. (2020). Revisiting difficulty bias, and other forms of bias, in elite level gymnastics. *Journal of Sports Analytics*, *6*(1), 1–11. https://doi.org/10.3233/JSA-200272.

Sedlacek, W. E. (1987). Black students on White campuses: 20 years of research. *Journal of College Student Personnel*, *28*(6), 484–495. https://psycnet.apa.org/record/1988-37333-001.

Sue, D. W., Capodilupo, C. M., Torino, G. C., Bucceri, J. M., Holder, A. M. B., Nadal, K. L., & Esquilin, M. (2007). Racial microaggressions in everyday life: Implications for clinical practice. *American Psychologist*, *62*(4), 271–286. https://doi.org/10.1037/0003-066X.62.4.271.

Van Dyke, E. D., Metzger, A., & Zizzi, S. J. (2020). Being mindful of perfectionism and performance among collegiate gymnasts: A person-centered approach. *Journal of Clinical Sport Psychology*, *15*(2), 143–161. https://doi.org/10.1123/jcsp.2019-0100.

VanderWeele, T. J., & Mathur, M. B. (2019). Some desirable properties of the Bonferroni correction: Is the Bonferroni correction really so bad? *American Journal of Epidemiology*, *188*(3), 617–618. https://doi.org/10.1093/aje/kwy250.

Wamsley, L. (2023). Women's gymnastics is changing in more ways than one. *NPR Sports*. https://www.npr.org/2023/09/02/1197287064/womens-gymnastics-is-changing-in-more-ways-than-one.

Willie, C. V., & Cunnigen, D. (1981). Black students in higher education: A review of studies, 1965-1980. *Annual Review of Sociology*, *7*, 177–198. https://www.jstor.org/stable/2946027.

Xiao, Y. J. (2022). Examining how region and individual competency affect team performance in Artistic Gymnastics using RStudio. *Scholarly Review Journal*, *Fall 2022*(4). https://doi.org/10.70121/001c.121677.