Uneven Bars? Using a New Dataset of Women's Gymnastics Scores to Look for Environmental Microaggression Effects in the NCAA

Tommy Morgan† Seth Cannon‡

July 9, 2024

Link to most recent version

Abstract

We create a new dataset of all NCAA Women's Gymnastics scores from meets occurring from 2015-2024. The scores were aggregated from RoadtoNationals.com, the NCAA's official statistical and ranking site for women's gymnastics since 2015, and they include important details such as host teams, meet dates, and gymnasts' names. This dataset allows researchers to easily use NCAA gymnastics data in their research. To demonstrate the potential of our dataset, we use it to examine whether Black gymnasts experience negative performance effects when competing at specific universities in the NCAA. We estimate a generalized difference-in-differences regression model with fixed effects for gymnasts and meets to analyze full regular seasons of scores for gymnastics teams as they visit a given university. Across the 87 universities to have hosted an NCAA meet from 2015-2024, we find Black gymnasts experiencing significant performance effects at just a few universities with no obvious connections between them.

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.

1. Introduction

Research investigating behavioral effects using women's gymnastics has primarily focused on elite-level gymnasts. These papers most frequently deal with race-agnostic biases present in judges, such as difficulty bias (Rotthoff, 2020) and ordering bias (Damisch et al., 2006; Morgan and Rotthoff, 2014; Rotthoff, 2015; Joustra et al., 2020; Rotthoff, 2020). Other gymnastics-based research has focused on biases held by competitors, like in Meissner et al. (2021), which focuses on Simone Biles' superstar effect on her competitors. These papers make use of the unique context of gymnastics competitions (individual scoring within team competition, tournament settings, judge panels, etc.) to investigate these biases in a convincing way. However, due to the infrequent nature of elite-level tournaments relative to other gymnastics competitions, these papers frequently rely on a few tournaments' worth of data to draw their conclusions (with Meissner et al. (2021) being a notable exception).

We contribute to research based on gymnastics by creating a dataset of all NCAA Women's Gymnastics scores from meets occurring from 2015-2024. To our knowledge, our dataset of scores is the only comprehensive pre-processed source for these scores. While these scores have been made publicly available for many years thanks to the hard work of the people behind RoadtoNationals.com, the data itself is not readily available for download at that source. In total, our dataset includes 230,088 scores across all four women's gymnastics events received by 4,720 gymnasts over 3,581 meets. We make our full dataset of NCAA women's gymnastics scores and all code used for the analysis in this paper available[1].

We use our dataset of scores from NCAA women's gymnastics meets from 2015-2024 to examine whether Black gymnasts experience a negative performance effect relative to their peers at each of the 87 universities that have hosted a meet over that time period. For each university, we analyze the performance of NCAA gymnasts who are **not** on their team over the entire regular season(s) in which they performed at a meet hosted by said team at least once, and we include scores from the full set of non-invitational regular season meets from the season in which they visited that team. For example, Table 1 shows the full set of team-seasons we include when estimating our empirical model using the University of Alabama as a host.

We leverage the unique circumstances of NCAA women's gymnastics competition, in which the score a gymnast receives is given at the individual level; this is important because any negative effect from environmental racial microaggression (explained below) should also manifest itself primarily at the individual level. We estimate a generalized difference-in-

[1]These will be made available on a GitHub repository after publication

differences model of these scores that uses an indicator variable for being a Black gymnast competing at a given host school as the independent variable of greatest interest. We argue that our model controls for all relevant factors that could affect a gymnast's score, including innate ability, preparation, potential judge bias, and so on. We then plot these estimated effects against the percentage of the full set of gymnasts who ever performed for a given university who were Black and see no generalized trend.

2. Background & Related Literature

2.1 Women's Artistic Gymnastics

In women's artistic gymnastics, a regular season meet is composed of four events: vault, uneven bars, balance beam, and floor exercise. Gymnasts are allowed to compete in all four events at any given meet, but typically complete in only one or two of these events. Each performance is scored out of 10 by two to four judges whose independent scores are averaged to a final performance score. The typical regular season meet has four judges – two from in-region and two from out of region – each judging two events, 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 between events. In each event, five to six gymnasts from each team perform, and if six gymnasts perform, the lowest of those six scores is dropped when calculating the overall team score.

At the NCAA level, scores are determined by two factors: the "start value" of the routine, which is the score a gymnast would receive by performing their prepared routine perfectly, and deductions taken from the start value for technical or execution errors. Routines are required by rule to have at least a 9.4 point start value, but gymnasts at the collegiate level are typically sufficiently skilled to begin at the maximum 10 point start value. Though an individual score can theoretically range anywhere from zero to 10 in each event, routines without major errors typically cluster within the 9.6 - 9.9 range. Even routines with major errors can score relatively high; for example, at the 2024 meet between West Virginia and BYU, West Virginia's Emma Wehry received a score of 7.7 for her performance on the uneven bars despite having crashed directly into the lower bar and, as a result, not having been able to complete her routine.

After all events are complete, the five highest scores in each event are summed to compute each team's meet score. Because the practical range of scores is so small, tiny differences in average scores separate elite teams from great and decent teams: elite gymnastics teams have the potential to hit a 198.00 meet score, which is obtainable only with an average score of 9.9 from every gymnast in every event across the entire meet; great teams can consistently hit a 197.00 meet score (a 9.85 average performance score); and good teams can consistently hit a 196.00 meet score (a 9.8 average performance score) (Grimsley, 2019). These are differences of 0.05 points on average per routine, so a negative effect that affects some gymnasts even to the extent that they lose one-hundredth of a point (0.01) could be substantially harmful to their team's success.

The unique attributes of artistic gymnastics meets offer us several key advantages. One such advantage is that scores are assigned to gymnasts on an individual basis. Though research has shown that there may be an overall ordering bias in judging (Damisch et al., 2006; Morgan and Rotthoff, 2014; Rotthoff, 2015; Joustra et al., 2020; Rotthoff, 2020), that same research generally shows that a gymnast's performance is not affected by the performances immediately preceding it. Because scores are mostly independent across individual gymnasts, we can use individual routine scores to look for the presence of an environmental effect that would manifest at the individual level.

Another advantage is that each of the four gymnastics events is a high-intensity, high-focus exercise. We hypothesize that any unusual outside pressure could cause gymnasts to make mistakes they would not make absent such pressure, with potential pressures of this kind including a negative environmental microaggression effect as we will soon define. If Black gymnasts were to experience such an effect when performing at a given host university to a degree that it negatively affected their performance relative to their non-Black peers, we would see it in our data. Whether such an effect, if present, would be significant enough to affect competitive performance to a detectable degree is the focus of the study.

2.2 Racial Microaggression Theory

Recent research that investigates the effects of environmental microaggressions on Black individuals is largely based on the model of microaggression theory presented in Sue et al. (2007), which suggests that environmental factors such as the names of buildings or overall racial climates can constitute "[m]acro-level microaggressions, which are more apparent on systemic and environmental levels" than other types of interpersonal microaggressions. Research on this topic is often focused on qualitative interviews or surveys of Black students' experiences at predominantly White institutions (PWIs) (Mills, 2020; Holliday and Squires, 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).

In addition to recent research on racial bias founded on microaggression theory, much research exists on racial biases within the world of sports. Generally, this research focuses on racial biases in referee/judge decisions (as in Price and Wolfers, 2010; Parsons et al., 2011; Gallo et al., 2012; Rotthoff, 2020; Eiserloh et al., 2020; and Pelechrinis, 2023) or in fan/commentator preferences (as in Andersen and La Croix, 1991; Preston and Szymanski, 2008; Reilly and Witt, 2011; Principe and van Ours, 2022; and Quansah et al., 2023). These studies use data from professional sports leagues in many sports and around the world to generally show that racial biases can affect sports teams both in competitive outcomes and perceived value. We contribute to racial microaggression research by trying to find evidence of one of its subtypes in a novel setting. Additionally, we aim to quantify this specific type of microaggression effect within a statistical framework that could uncover causality, which has not been done previously in the context of environmental microaggression theory to our knowledge.

Two more investigations warrant particular note in this paper. The first 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 effects in football (Dix, 2017, Dix, 2021a), men's basketball (Dix, 2022a, Dix, 2022b), women's basketball (Dix, 2019, Dix, 2020a, Dix, 2022b), baseball (Dix, 2020b), softball (Dix, 2021b), and volleyball (Dix, 2023). This line of research helps bridge the gap between previous work on professional sports and research at the college level. Though Dix's work focuses primarily on averaged team results and not on individual-level effects, it is nonetheless useful for contextualizing our work in this paper. We differentiate ourselves from Dix by focusing on individual-level effects as opposed to team-level effects.

The second investigation of note is found in Caselli et al. (2023). In their paper, the authors show that African players in a professional Italian soccer league improved their performance when COVID-19 prevented fans from attending their games. They argue that this effect stems from the absence of overtly racist fan behavior, which is common in that league. This research is relevant to our paper because it is, to our knowledge, the closest existing paper to this one. Like we will in this paper, Caselli, Falco, and Mattera evaluate individual-level performance scores (in this case, those 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 also model the effects of a quasi-environmental removal of direct racial aggressions, which mirrors our empirical strategy in which we exploit the introduction of gymnasts into a potentially indirect environmental microaggression. We differentiate ourselves from Caselli, Falco, and Mattera by focusing on the introduction of a potential racial microaggression to college-level athletes as opposed to the removal of an overt set of racial aggressions from professional athletes.

3. Empirical Strategy

3.1 Model

To be able to attribute causality to any estimate we produce, we must be reasonably sure that we have controlled for as many other factors that could influence a gymnast's score as possible. We suggest a simple model of a gymnast's score in any given event that breaks down influential factors into four principal categories:

$$\text{score} = \text{ability} + \text{preparation} + \text{environment} + \text{event} + \varepsilon \tag{1}$$

In this model, ability-related factors could include, for example, the genetic makeup of a given gymnast, the age at which they began training, and the set of skills they have the physical capacity to perform. Preparation-related factors might include the quality of the team and coaches surrounding a gymnast, the number of years a gymnast has been competing, the types of skills a gymnast chooses to practice, and the number of meets that have already occurred in a season. Environment-related factors could include the relative competency or biases of judges at a given meet; the altitude, location, quality, and name of the venue; the time of the meet; and so on. Event-related factors adjust the score for the nuances of each event; for example, it is possible to fall off of the uneven bars, but not the floor exercise. Finally, because there is bound to be stochastic error in any model of human behavior, we also include an error term in our model.

Having put forth this general framework, we further posit that gymnasts and their coaches behave as score maximizers. 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 has perfect information about the ability- and preparation-related elements of the proposed scoring model. We also assume that a gymnast has no control over the environment-related factors of her scoring.

3.2 Sample Construction

We begin by narrowing our sample to scores from a subset of 3,212 meets in our dataset that meet a certain set of criteria. We want to avoid incorporating playoff meets into our sample, as these meets could induce a different set of incentives than score maximizing under certain circumstances. For example, if a final gymnast only needed a relatively low score on her beam routine to guarantee moving her team to the next round, she might change her routine to minimize the chance of a major mistake instead of trying a riskier, more difficult routine that could maximize her score. In addition, we want to only consider non-invitational meets. Invitational meets often represent a different environment than a typical regular season meet, as they are often hosted by gymnastics organizations instead of specific universities.

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 Arkansas women's gymnastics team performed at Alabama every other year beginning in 2016 through 2024, so we include every score from every Arkansas gymnast at every one of their non-invitational regular season meets from those five years (2016, 2018, 2020, 2022, and 2024) in our sample. In contrast, we include every score from every University of Denver gymnast at every one of their non-invitational regular season meets from only 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 described in Table 1. We then estimate the model we will describe below for Alabama, save the results, and repeat this process for each successive host university.

In order to examine the effect of being Black when performing at a given university, we also need to know which gymnasts are Black and which ones are not. Since we do not have the individual-level data that each gymnast reports to the NCAA, we coded each gymnast as Black or not based on visual inspection of their official photographs from each school's official gymnastics website. Figure 1 shows an example of a typical web page used in this process. We are aware that binary assignment of race based on skin color and other visual features subjects us to the "eye of the beholder" problem discussed in Fort and Gill (2000). For this reason, we compare our classifications to aggregated racial demographics provided by the NCAA (National Collegiate Athletic Association, 2018) in Table 2. It should be noted that their database includes all registered student-athletes, whereas our data only includes those who competed and received at least one score in a given year. In addition, we may identify gymnasts as Black when they identify in the NCAA demographics as Other, including mixed race. Our percentages of Black gymnasts in each year from 2015-2023 (2024 data has yet to be published) are not identical to the NCAA's data, but they do follow similar trends year by year.

3.3 Estimation Strategy

We begin our estimation strategy with a baseline differences-in-differences model:

$$\text{score}_{iem} = \beta_0 + \beta_1 \text{Black}_i + \beta_2 \text{atHost}_m + \beta_3 \text{Black*atHost}_{im} + [\text{event}]_e + u_{iem} \tag{2}$$

where subscripts i, e, and m refer to individual gymnasts, events, and meets, respectively. The dependent variable is the score earned by a gymnast in a given event, and the interaction term $(\text{Black*atHost})_{im}$ takes a value of 1 if the observation is a score received by a Black gymnast competing at a host university and 0 otherwise. Black_i and atHost_m represent binary variables for a gymnast being Black and a meet being held at said university. We also include fixed effects for the event in which a given score was received (i.e. vault, bars, etc.) denoted as $[\text{event}]_e$. The inclusion of this piece is supported by the event score means presented in Table 3, where the events that can be fallen off of (bars and beam) average lower scores than the events that can't (vault and floor), and it broadly controls for the nuances particular to each event. We include a stochastic error term represented by u_{iem}. In this model, β_0 is the regression constant term, and the coefficient of interest is β_3, which we interpret as the differential impact of competing at that university on Black gymnasts relative to non-Black gymnasts and other venues.

This baseline model certainly suffers from omitted variable bias. To combat this, we introduce gymnast- and meet-level fixed effects:

$$\text{score}_{iem} = \beta_0 + \beta_3 (\text{Black*atHost})_{im} + [\text{gymnast}]_i + [\text{meet}]_m + [\text{event}]_e + \epsilon_{iem} \tag{3}$$

where each of the new square bracketed terms denotes a relevant set of fixed effects and the error term is represented by ϵ_{iem}. These new sets of fixed effects control for gymnast- and meet-specific characteristics that influence the score a gymnast receives at a given meet, including (but not limited to) individual judge biases, venue altitude, point-in-time effects (such as the start value a certain skill is worth in the year of a given meet), gymnasts' innate skill and physical qualities, cumulative coach and team effects on gymnasts, and so on. We argue that these fixed effects control for every possible relevant factor to the score a gymnast receives as put forth in our generalized scoring model. Finally, to account for possible correlation between the scores received at a given meet (due to shared judges, etc.), we report standard errors clustered at the meet level.

For a causal interpretation of β_3 to be possible, the gymnasts in our sample must satisfy the parallel trends assumption. Specifically, we assume that Black gymnasts performing at a given host university would experience the same relative change in performance at that venue as their non-Black counterparts would at the same venue in the absence of any extra effect of that venue on Black gymnasts. Since different teams visit different hosts at different points in the season throughout our dataset, we attempt to support our parallel trends assumption with Figure 2, in which we plot average scores in our sample by race and meet number. The figure shows that both Black and non-Black gymnasts see their scores increase over the course of a season on average. Most notably, the lines of best fit we plot in that figure fit comfortably within one standard deviation from the average scores of gymnasts across both groups over the course of a regular season. We suggest that this figure demonstrates the baseline viability of the parallel trends assumption needed for our estimates to be interpreted as causal.

We interpret a given host university's β_3 estimate as a gymnast-at-host-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 university affecting Black gymnasts for some reason. We reason that if we were to see negative β_3 estimates within certain sets of universities - such as those that have never had a Black gymnast on their team or those included in a web article or Twitter thread compiling gymnasts speaking out against racism within their teams (such as in Duffy (2020) or Boswell (2020)) - then we may reasonably conclude that an environmental microaggression effect is present for Black gymnasts at some subset of universities. However, if universities with statistically significant β_3 estimates do not follow a noticeable trend, we might instead attribute those results to statistical noise.

We also considered implementing a triple difference design following Olden and Møen (2022), where the third difference would come from a score being received pre- or post-May 2020. The motivating event behind that difference would have been the aftermath of George Floyd's death and its effect of raising social awareness about racial issues in the United States. However, to satisfy the parallel trends assumption for the triple difference estimator, we would have had to assume that Black gymnasts performing at a given school would have experienced the same relative change in performance pre- and post-2020 as their non-Black counterparts absent the heightened environment of racial awareness. This would mean assuming that the COVID-19 pandemic had the same effect on Black gymnasts as non-Black gymnasts. Knowing of the large body of research against this point (see Goldman et al. (2021) and Vasquez Reyes (2020) for just two examples), we decided that such a design would not be sufficiently useful to include in this paper.

4. Results

Figure 3 shows the results of estimating β_3 for each of the 87 host universities in our sample. Of those 87 universities, only seven return estimates of β_3 that are statistically different from zero, and four of those are positive. There does not appear to be a notable trend in what universities return these statistically significant estimates; they cover a large span of the Black gymnast participation rate captured on the X-axis, 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). For a closer look at these specific universities, we record the exact results of our estimation for each of these seven universities in Table 4. Given these results, we find it unlikely that there is an environmental microaggression effect that affects Black gymnasts in the NCAA. But, then, how should we interpret the results we do see?

By using 95% confidence intervals as our judge for the statistical difference of β_3 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 87 universities in our sample, we would nonetheless expect to see two instances of statistically negative effects and two of statistically positive effects. As such, we argue that the few statistically significant estimates we observe in Figure 3 are more likely to be a result of our choice of statistical inference method than they are to be indicative of any sort of environmental effect among certain universities as meet hosts. In addition, the fact that we see a few more positive estimates than we would expect to see by pure chance may be reflective of a trend visible in Table 3 and Figure 2 in which Black gymnasts score higher on average than their non-Black peers across the NCAA as a whole.

Given that we saw significant estimates across a seemingly random set of schools, 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_3? 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 show in Figure 4 the results of applying the Bonferroni correction[2] to the confidence intervals used in Figure 3 (as established in Dunn (1961) and discussed in Armstrong (2014) and VanderWeele and Mathur (2019)).

We see in Figure 4 that the β_3 estimate for the University of Pittsburgh remains statistically significant even after applying the Bonferroni correction. According to (Armstrong, 2014), our use of the Bonferroni correction in this context can be interpreted as testing the "universal null hypothesis" that every test we ran is not statistically significant. Seeing that one of our results remains significant suggests there is some differential effect for Black gymnasts in at least one place, but we have no knowledge as to why this would be positive generally or found at the University of Pittsburgh specifically. It is possible that the Pittsburgh estimate survives because it has the greatest number of meets and scores of the previously significant estimates, implying this estimate having the greatest statistical power of that set; this would align with the fact that the Bonferroni correction generally reduces the power of a statistical estimate. We emphasize that this is only speculation; we primarily view this result as "hypothesis generating", not hypothesis confirming.

5. Conclusion

We contribute to research that uses sports to research broader effects on agents by making a comprehensive dataset of collegiate women's gymnastics scores available to researchers for the first time. We use that dataset to test whether Black gymnasts experience a change in performance that their non-Black competitors do not experience when competing at 87 NCAA universities. Our dataset allows us to introduce important sets of fixed effects to control for relevant factors that influence scores, which allow us to isolate the interaction between a gymnast's being Black and the host of a meet at which they perform.

We find few significant differences in score distributions between Black and non-Black gymnasts using our model, and we argue that this result provides no evidence that any given university hosting a given meet negatively affects Black gymnasts' performance to a notable degree. We do find that one result survives an extremely conservative statistical adjustment, but we have no solid explanation for why. As such, we call for further research into this effect, both in the context of collegiate sports and beyond.

[2] In our case, this involves creating 99.95% confidence intervals for each estimate, approximating the $\frac{0.05}{87}$ adjustment to our p-values that approximately requires a 99.942% CI.

References

Andersen, T. and La Croix, S. J. (1991). Customer racial discrimination in Major League Baseball. *Economic Inquiry*, 29(4):665–677. DOI: 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. DOI: 10.1111/opo.12131.

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., and Mattera, G. (2023). When the stadium goes silent: How crowds affect the performance of discriminated groups. *Journal of Labor Economics*, 41(2):431–451. DOI: 10.1086/719967.

Damisch, L., Mussweiler, T., and Plessner, H. (2006). Olympic medals as fruits of comparison? Assimilation and contrast in sequential performance judgments. *Journal of Experimental Psychology: Applied*, 12(3):166–178. DOI: 10.1037/1076-898X.12.3.166.

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. DOI: 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 Milford, M. and Reichart Smith, L., editors, *Communication and Contradiction in the NCAA: An Unlevel Playing Field*, chapter 13, pages 213–234. Peter Lang, New York.

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. DOI: 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. DOI: 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. DOI: 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. DOI: 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. DOI: 10.1080/01621459.1961.10482090.

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

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

Gallo, E., Grund, T., and 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. DOI: 10.1111/j.1468-0084.2012.00725.x.

Goldman, N., Pebley, A. R., Lee, K., Andrasfay, T., and Pratt, B. (2021). Racial and ethnic differentials in COVID-19-related job exposures by occupational standing in the US. *PLOS ONE*, 16(9):e0256085. DOI: 10.1371/journal.pone.0256085.

Grimsley, E. (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. and 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. DOI: 10.1111/josl.12438.

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

Meissner, L., Rai, A., and Rotthoff, K. W. (2021). The superstar effect in gymnastics. *Applied Economics*, 53(24):2791–2798. DOI: 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. DOI: 10.1037/dhe0000121.

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

National Collegiate Athletic Association (2018). NCAA demographics database. *NCAA.org*. https://www.ncaa.org/sports/2018/12/13/ncaa-demographics-database.aspx. Accessed 1 July 2024.

Olden, A. and Møen, J. (2022). The triple difference estimator. *The Econometrics Journal*, 25(3):531–553. DOI: 10.1093/ectj/utac010.

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

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

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

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

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

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

Reilly, B. and Witt, R. (2011). Disciplinary sanctions in English Premiership Football: Is there a racial dimension? *Labour Economics*, 18(3):360–370. DOI: 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. DOI: 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. DOI: 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., and Esquilin, M. (2007). Racial microaggressions in everyday life: implications for clinical practice. *American Psychologist*, 62(4):271–286. DOI: 10.1037/0003-066X.62.4.271.

VanderWeele, T. J. and 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. DOI: 10.1093/aje/kwy250.

Vasquez Reyes, M. (2020). The disproportional impact of COVID-19 on African Americans. *Health and Human Rights Journal*, 22(2):299–307. PMCID: PMC7762908.

Willie, C. V. and 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.

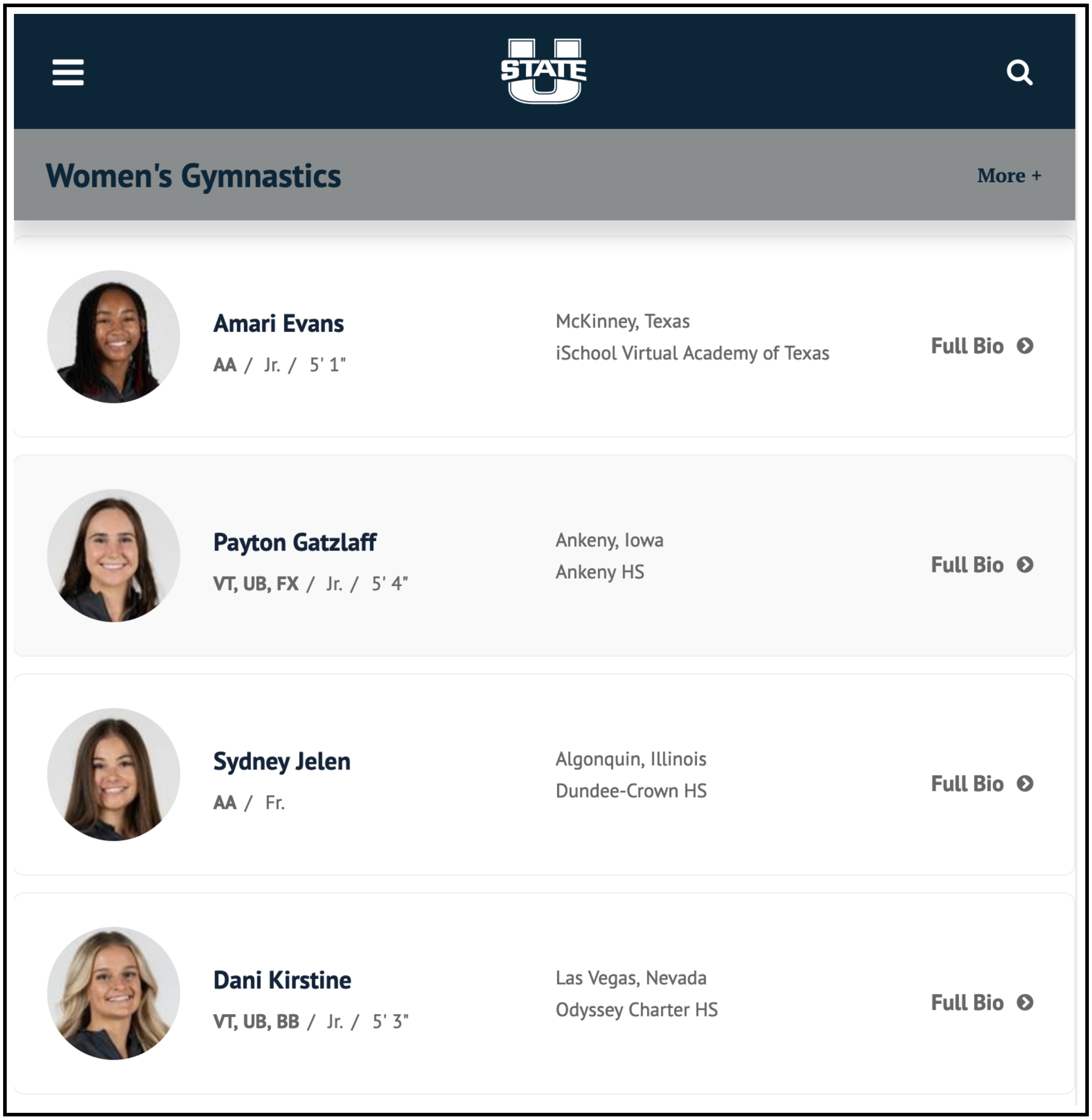
Figures and Tables

Table 1: Team-Seasons Included in the Sample for the University of Alabama

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					
Talladega										X
Temple					X					
West Virginia		X								
Western Michigan								X		
Scores by Year	1,437	1,173	1,078	1,441	2,128	1,079	768	1,318	1,127	1,581
Total Scores	13,130									

Notes. Each X represents a year in which the team on a given row competed at Alabama in a regular season meet. If a team has a check for a given year, every non-invitational regular season meet in which that team competed in that year is included in our dataset. *Asterisks denote conference opponents.

Figure 1: Example Screenshot of a Women's Gymnastics Roster



Notes. This screenshot shows four gymnasts from the 2022 Utah State University women's gymnastics roster as displayed on the official team website. The screenshot was taken by the authors from https://utahstateaggies.com/sports/womens-gymnastics/roster/2022 on 28 December 2023. More photos of each gymnast were available upon clicking their names. Amari Evans was coded as Black, while the other three gymnasts pictured were not.

Table 2: NCAA Demographics Comparison

	Our Sample		NCAA Demo.	
Year	N	Black %	N	Black %
2015	108	9.30%	116	7.77%
2016	108	9.25%	118	7.86%
2017	111	9.54%	117	7.68%
2018	126	10.53%	123	7.95%
2019	131	10.93%	127	8.24%
2020	141	12.03%	129	8.51%
2021	104	11.12%	133	8.39%
2022	141	11.01%	131	7.67%
2023	155	12.08%	139	8.10%

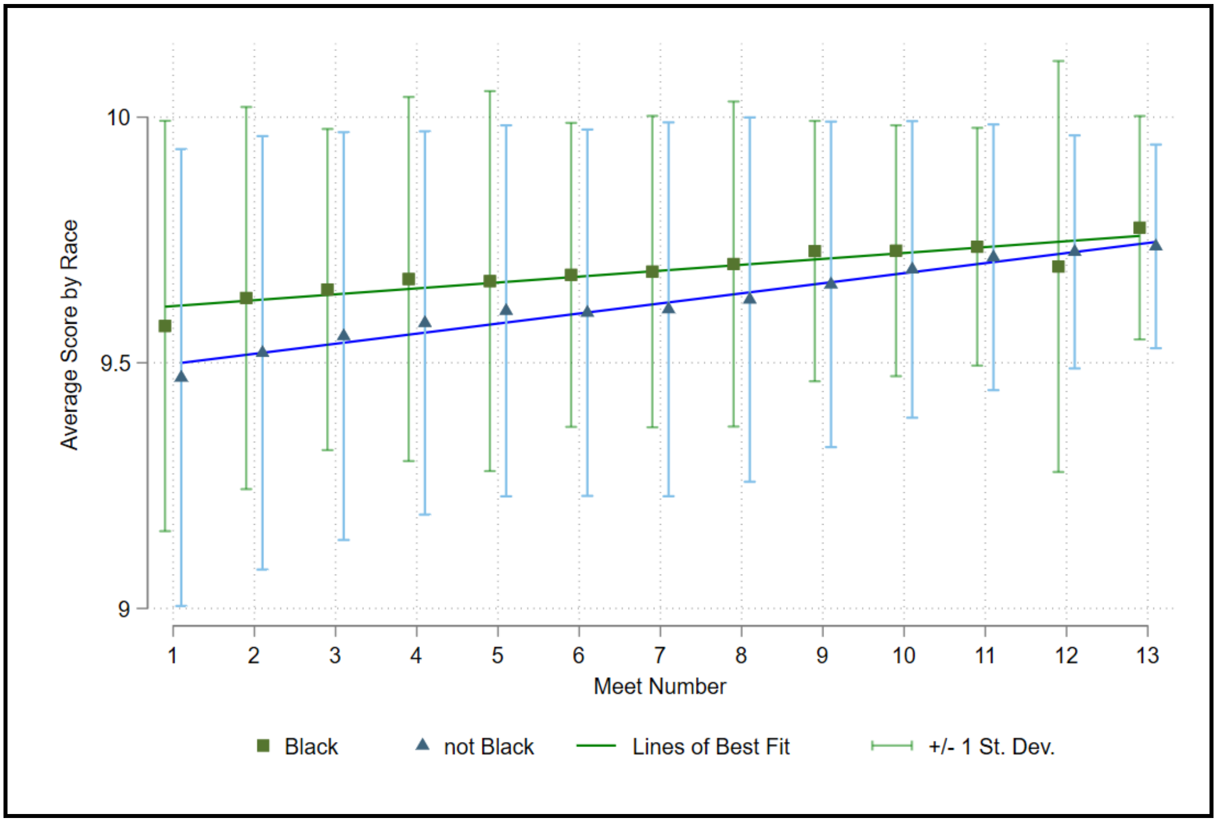
Notes. N is number of Black gymnasts, and Black % is the percentage of all gymnasts in that year who are Black. Our Sample includes only gymnasts who recorded a score at a non-invitational regular season meet in a given year; NCAA Demo. includes all gymnasts registered with the NCAA in a given year whether they compete or not.

Table 3: Score Summary Statistics

	Vault	Bars	Beam	Floor
Black				
Mean	9.707	9.626	9.617	9.721
(SD)	(0.227)	(0.439)	(0.349)	(0.334)
N	6,667	5,306	4,489	5,960
Not Black				
Mean	9.646	9.533	9.566	9.642
(SD)	(0.256)	(0.487)	(0.396)	(0.364)
N	37,628	39,174	40,129	38,417

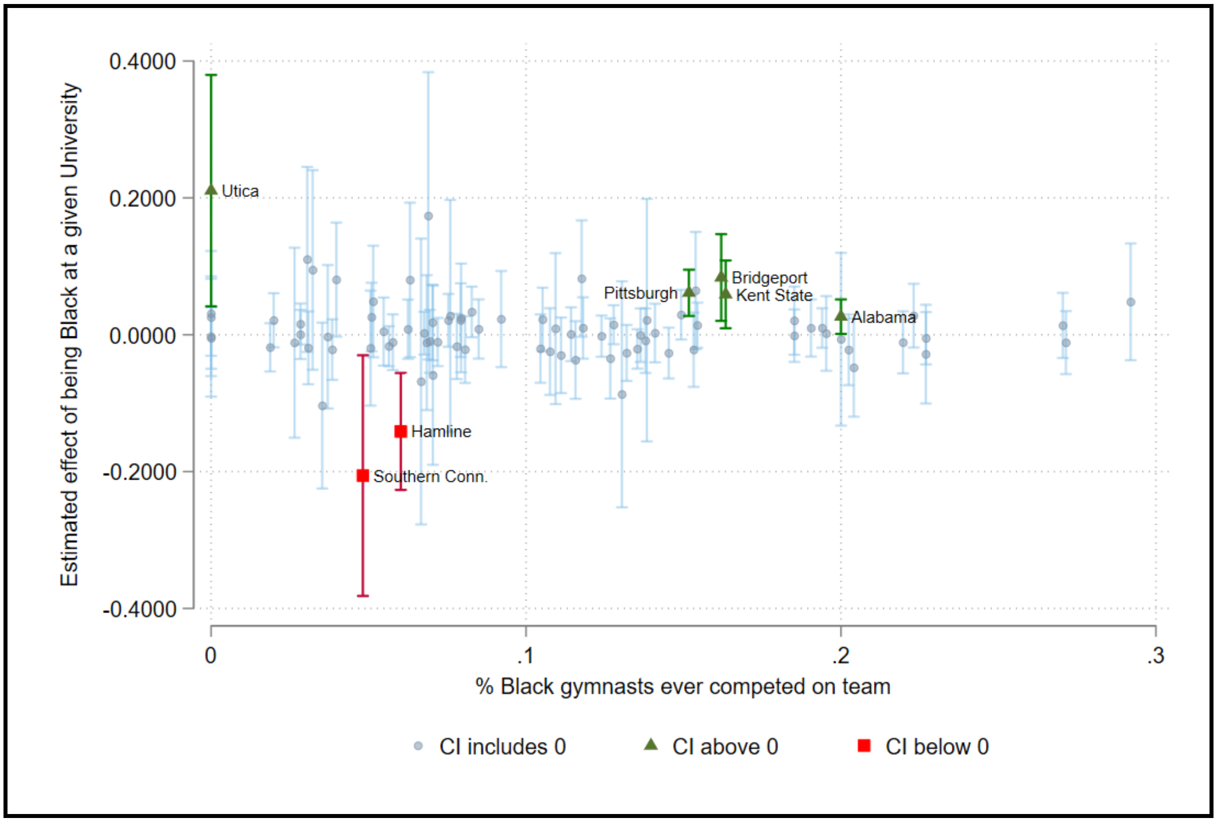
Notes. This table includes all of the scores we use in our sample; it contains no scores from invitational or playoff meets.

Figure 2: Average Score by Race and Meet Week Number



Notes. This table includes all of the scores we use in our sample; it contains no scores from invitational or playoff meets.

Figure 3: Dif-in-Dif Estimates Plotted by % Black Gymnasts Ever Attending



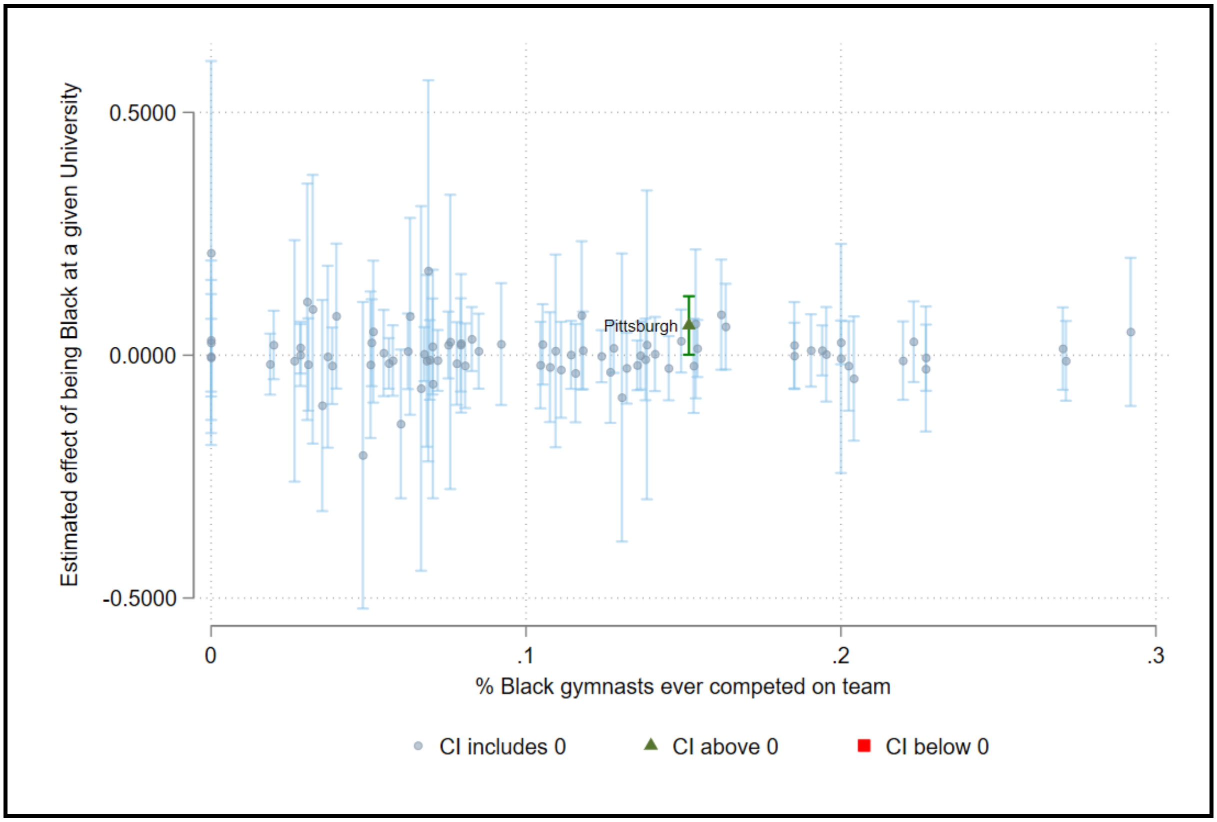
Notes. This figure plots the β_3 estimate for all 87 host universities in our sample along with their corresponding 95% confidence intervals. Estimates with confidence intervals above 0 are colored green and marked with triangles; those below 0 are colored red and marked with squares. % Black gymnasts ever competed on team is calculated using gymnasts performing for a given host university from 2015-2024. Standard errors are clustered at the meet level.

Table 4: Score Summary Statistics

University	Black %	Estimate (St. Err.)	Bonferroni Significant?	N	R^2
Utica	0.00%	0.211** (0.0747)	No	239	0.543
Southern Conn.	4.82%	-0.206** (0.0892)	No	6,374	0.514
Hamline	6.02%	-0.141*** (0.0434)	No	7,175	0.440
Pittsburgh	15.17%	0.061*** (0.0172)	Yes	15,013	0.223
Bridgeport	16.20%	0.084*** (0.0322)	No	7,272	0.407
Kent State	16.34%	0.059** (0.0252)	No	12,353	0.298
Alabama	20.00%	0.026** (0.0128)	No	13,130	0.289

Notes. These universities are labeled in Figure 3 as the only universities in our sample with statistically different-from-zero estimates for β_3. Observations (N) are individual event scores received by gymnasts, selected for inclusion in each row as described in the Sample Construction subsection of our Empirical Strategy section. Standard errors are clustered at meet level. 5%**, 1%***.

Figure 4: Dif-in-Dif Estimates With Bonferroni-Adjusted Confidence Intervals



Notes. This figure plots the β_3 estimate for all 87 host universities in our sample along with their corresponding 99.95% confidence intervals. Estimates with confidence intervals above 0 are colored green and marked with triangles; those below 0 would be colored red and marked with squares if there were any. % Black gymnasts ever competed on team is calculated using gymnasts performing for a given host university from 2015-2024. Standard errors are clustered at the meet level and Bonferroni-adjusted.