Dear Dr. Wicker,

We are grateful to you and the reviewers for the insightful feedback you provided and for the opportunity to revise our paper once again for the Journal of Sports Economics. The comments we received helped us to think more deeply about our research question and significantly improved the manuscript, which we hope will be evident in our resubmission.

Before responding directly to each comment, we preface by noting three major revisions that we believe have greatly improved every section of the paper.

First, one important theme in reviewer comments had to do with how we assigned race to each gymnast, and we were admittedly uncomfortable with having done that 'by hand'. As a result, we have integrated a computer vision model called FairFace to make those assignments for us. Using FairFace also opened up classification of gymnasts into races other than Black and not Black (as we had previously been limited). This new addition greatly improved our paper's credibility without significantly altering our eventual conclusion, and more details about using FairFace are noted in responses to specific comments below.

Second, thinking about reviewer comments regarding Division I vs. Divisions II and III gymnasts prompted further scrutiny of the parallel trends assumption we used to justify our difference-in-differences design. As a result of this further scrutiny, we discovered that parallel trends are much more likely to hold when we limit our sample to DI gymnasts performing at regular non-invitational non-playoff meets. As with the inclusion of FairFace, this new narrowing of our sample increases the believability of our empirical strategy without changing the results of our analysis.

Third, several comments asked for more discussion of the implications of our results while calling for a greater focus on the research question as opposed to the novelty of the data. We agreed that this shift in focus would create a better paper, so we have restructured and rewritten almost every section of the paper with this in mind. We hope that this understanding will justify our inability to "highlight the changes to the paper using highlighted/colored text" as requested, but we will still be thorough and detailed in our responses to comments, citing page and paragraph numbers whenever possible.

With those major themes in mind, we now address each comment directly in the table below as requested.

Thank you again,
Tommy Morgan & Seth Cannon

Comment Source	Comment	Response
AE	1) Format Please follow the requested APA7 style more closely – text and ref list are double-spaced, headings are not numbered, ref list must be in APA style, and so on. Nothing is bold per APA.	Our manuscript now uses LaTeX's APA7 document class, which has fixed the heading style, text spacing, bolding, and so on.
AE	Table and figure insertions must be added to the text.	They are now inserted in the text.
AE	All figures are hard to read. 1. Does the comment "Figures are hard to read" refer to the quality of the image files uploaded (i.e. they got compressed) or the actual figure designs? --> font size and size of what is shown, lengthy notes.	We recreated our figures in order to emphasize readability. They are in black and white, with less text that is hopefully clearer than before. In addition, our table and figure notes have been shortened significantly throughout the paper.
AE	The typical abbreviation is Diff-in-Diff.	This abbreviation now only appears in Table 4 (p. 19) and has been corrected.
AE	Keywords are missing on the title page.	We added appropriate keywords to the title page.
AE	Title: New dataset is not an appealing term in a title – the title should reflect your RQ. Same with 'new' dataset in the abstract – nobody would write we use an old dataset.	We removed mentions of a "new" dataset throughout the paper, including in the abstract and title. Our title is now focused only on the research question. We also added a new Data section (p. 7) explaining relevant details about our dataset.
AE	Intro Please use more space to justify the relevance of the topic rather than saying what you are examining. The intro should include the purpose statement and some concrete research questions (RQs).	Our introduction (p. 1) now focuses on motivating and explaining our research question. It was completely rewritten with this goal in mind.
AE	Methods should be in chronological	We renamed our Empirical Strategy

	order: I think first you collect data, then you describe your variables, and then you run models.	section to Methods (p. 7) to reflect this idea and added a Data section to it. The order of the section now goes Data > Sample Construction > Model > Estimation.
AE	The ms lacks a discussion chapter where the findings are discussed with reference to theory and previous research.	Our conclusion section is now a Discussion & Conclusion section (p. 20) that spends more time discussing our results in the context of existing theory & research. It also expands further on the implications of our results for future research.
1	I appreciated the opportunity to review this manuscript analyzing extensive gymnastics data and potential racial discrimination. The study uses a robust dataset of gymnastics scores, and it is kind of the authors to make the data available to other researchers. The authors pose some difficult and important questions regarding gymnastics scores and the possibility that some athletes are unequally judged based on where they are competing and their racial background. The paper is well-written, and the literature review is relevant and provides a strong theoretical framework. As I outline in the comments below, I have some concerns regarding the methodology and variables included in the model that may show different results, which is why I recommend that revisions are made prior to recommending the paper be accepted for JSE.	We are grateful for the warm compliments provided by Reviewer 1 and appreciate their feedback on our manuscript. We hope to thoroughly address their comments below.
1	In the second paragraph, the authors discuss the dataset set up and note that the data is not readily available at the source, can they clarify if they are referring to NCAA or the website where they got the results from. If the dataset was not readily available, how did they get access to the data, or did they	We are referring to the website where we got the data, which is also the NCAA's official scorekeeping partner for this sport. We used internet scraping tools to aggregate data from that website, then cleaned errors we found (misspelled names compared to official rosters, missing meet dates, etc.) by hand.

	combine data from the noted website that is available to everyone?	This point is still mentioned in the final paragraph of our introduction (p. 2) and in our abstract, but it is clarified in greater detail in the new Data subsection of the Methods section (p. 7).
1	Regarding the model definitions and criteria, the authors note they conducted a difference-in-differences model, this however is not the case, as they do not examine a change over time or an intervention (e.g., change in policy or event) that impacted the "treated" group. At one point they note the events following George Floyd's death as a possible triple difference in design, that would be a potential notable difference-in-difference variable implying that an event took place that may have changed the trends at a given moment for the treated group. As it currently stands, the dataset is a time-series one, though once again, time does not appear to be an important factor in the model unless the authors wish to examine if something changed overtime and that biases have changed over time (either increased or decreased) and if time plays a factor (e.g., previous results as leads or future results as lags). This needs to be clarified in the methods section and corrected. The interaction term in the model examines the potential bias but not a diff-in-diff.	The first difference we examine in the paper is the difference between Black and White gymnasts (or White and not White gymnasts in our new second specification) at a given university. The second difference we examine is the difference between Black (White) gymnasts' performance at that given university compared to their performance at other meets not hosted by that university. Because the other meets happen at different times, we view the first difference as the "treated" difference and the second difference as the "post" difference. However, in this case, "post" is not the correct word; it's more like the "post" difference is inactive, then activates for meets hosted by the focus university, then deactivates again. For that reason, we still view our interaction term as a difference-in-differences estimator. We have clarified this greatly in our Estimation Strategy section (p. 15).
1	Is there perhaps a "judge" fixed effect that needs to be included in the model, identify if differences in scoring are not where the event is taking place but the specific judge? Can you run the models for each event separately (or why did you choose not to do so?) and possibly identify a bias by some judges who judge specific events?	This comment prompted us to adjust our fixed effect in the more robust second equation we estimate (p. 16) from a meet fixed effect to an event-by-meet fixed effect, which we explain in detail in its section. We argue in that section that doing so controls for the nuances of each event and for the average bias of the pair of judges that scores each event, which is constant within a given meet. Our models with event or event-by-meet fixed effects

		are functionally equivalent to running the model separately for each event, and we chose to run them together to maintain our ability to display the results clearly and compactly.
1	I will also be interested to see the relative score athletes get as opposed to individual score as the dependent variable, this can provide insights on how they were scored relative to others. This raises one "concern" I have with the analysis, as we do not know if the athlete made a mistake and only have the subjective scoring (similar to referee bias in other sports, e.g. soccer, if a penalty was awarded or not awarded depends on a lot of factors included subjective decision making). We cannot determine if an athlete received a lower score for doing the same routine conducted in similar manner (both made the same error or not error - or from soccer, why was one awarded a penalty and the other was not, but both appear to be similar fouls).	The interaction terms we estimate in equations 1 and 2 (p. 15 & 16) quantifies this relative difference: we compare Black gymnasts' scores at a meet hosted by a given university both to themselves at other meets (the time difference) and to their White peers at the same university (the treatment difference). On the actual paper scoring sheets collected at most modern meets, there is a section marking the amount of ND (non-discretionary) deductions received, such as stepping out of bounds, going beyond a time limit, etc. However, these have not been digitized, so collecting them for our entire sample is not possible for us given our current resources. Yet even collecting them would still not be necessary, as an environmental effect on performance could potentially affect both a gymnast stepping out of bounds (non-discretionary deduction) or not holding their handstand long enough (discretionary deduction), and either deduction would be useful to our analysis.
1	In the third model the authors included a gymnast fixed effect, my concern with that is the inclusion of that variable and the Black variable which may be "double counting". More clarification is necessary and perhaps replace the Black variable with the gymnast fixed effect to identify if certain gymnasts are subjected to biases and what are the characteristics of those gymnasts (post estimation analysis).	In equation 1 (p. 15), we include an indicator for being Black and do not include gymnast fixed effects. In equation 2 (p. 16), we include gymnast fixed effects and exclude an indicator for being Black. Including both in one equation would certainly be "double counting", which we avoid by splitting the equations. In equation 2, the indicator (Black*atHost) is not collinear with the gymnast fixed effect, as there would be within-gymnast variation in that indicator (1 when at the focus university, 0 else). This is enough to prevent these two sets of variables from double counting with each other. Our research design identifies if a certain

		set of gymnasts is subject to a certain kind of bias: are Black (White) gymnasts experiencing an environmental effect on their performance at any DI universities? We believe we do not need to do any post-estimation analysis to answer our research question given the empirical strategy we use.
1	Can the authors explain the substantial decrease in Black athletes seen in Table 2 in 2021 and the subsequent increase in 2022?	Yes! These statistics are displayed in greater detail in our new Table 1 (p. 8). Due to the COVID-19 pandemic, several schools across the country canceled their 2021 gymnastics seasons, leading to a drop of gymnasts who recorded a score in 2021 across all racial categories. No schools canceled their 2022 seasons, so we see a resulting increase across all categories again. Notably, the number of enrolled gymnasts recorded in Panel B of table 1 did not drop concurrently; this suggests that universities that cancelled their seasons still kept their student-athletes enrolled and on their teams even though they were not participating. Thankfully, because the year in which a score is received is constant within our event-by-meet fixed effects, these unique yearly circumstances are controlled for when we estimate equation 2.
2	Thank you for the opportunity to review this manuscript. This was a very interesting study and the researchers are commended for gathering such a large dataset spanning 10 years of NCAA women's gymnastics competitions. Below, I will share feedback and questions about the manuscript, which I hope will help the authors to strengthen this work.	We are especially grateful to Reviewer 2 for the excellent insight and careful attention to detail they provided to us in their comments. We hope to thoroughly address their comments below.
2	--First, please make sure you follow the journal's specified style, which is APA 7th edition. The journal specifically asks authors to submit	We have adapted the manuscript to follow APA 7th edition convention as explained previously.

	using this style, 12 point Times New Roman font, and double-spacing.	
2	--Along with that point, several of your in-text citations are formatted incorrectly. When you have multiple sources cited within parentheses, you should list these in alphabetic order, not chronological order. So, for example, on page 2 the Joustra et al. citation should be listed after Damisch et al. and before Morgan & Rotthoff. Additionally, APA Style specifies that parenthetic citations use the & sign instead of the word "and". Please make sure you proofread the article thoroughly, as these errors are persistent throughout.	We have corrected all citations and our reference list to APA 7 conventions.
2	--I've provided some more minor points regarding grammar, formatting, and style at the bottom of the review. MINOR POINTS: --p. 3, "typically complete" should be "typically compete". --p. 3, when you list "BYU" please spell out the full name of the university rather than the acronym. --There are a few instances where you spell out all author names for a three-person authored publication (i.e., Caselli, Falco, and Mattera), but when there are more than two authors you can use et al. in the first instance. Therefore, these should all be Caselli et al. Please double check the document for other cases of this and adjust accordingly. --p. 11 – there should be no parentheses around the Armstrong reference the way you have it written. Please get rid of these.	We have corrected all of these minor points as well, either through direct correction or by rewriting them out of the manuscript.
2	--On page 2 you state that the dataset contained all NCAA women's gymnastics scores from 2015-2024. Does this include all	This comment was the impetus for our deeper dive into the differences between gymnasts in each division; the revision describes why we chose to narrow our

	NCAA divisions (i.e., Division I, II, and III)? Or only Division I? Please just specify that in the text.	sample to only DI gymnasts in our Sample Construction section (p. 9).
2	--I'm wondering why you only chose to focus on Black athletes. There are numerous gymnasts from other races who compete in NCAA gymnastics – Asian, Hispanic/Latina, etc. Why not code every gymnast by race instead of only focusing on Black gymnasts? This would be quite interesting and would potentially uncover biases related to other non-white athletes. You have the data, so why not go back and examine these groups as well?	This comment was one of several that led us to integrate FairFace. Originally, we were focusing only on Black athletes because manually assigning race across several categories would have been much more labor-intensive for us. Now that we were able to use FairFace to do the classification, we no longer have to restrict ourselves so much. However, as we describe in our Sample Construction section (p. 9), the NCAA's Two or More Races and Other categories make it difficult for us to "verify" FairFace's predictions for categories other than Black and White. This shows up in the Latina category in Table 1 (p. 8). We are able to do a Black-White comparison and a White-everyone else comparison now, as described throughout the paper; the decision to do so was motivated by this comment.
2	--There is a lot of gymnastics-specific information included in your paper that makes sense to someone who knows and follows NCAA gymnastics, but will be confusing or possibly raise additional questions for those readers who do not follow it. I will point some of these out throughout the review. The first one I noticed was on page 3 when you talk about gymnasts typically only competing in one or two events. Please briefly explain WHY so many gymnasts only compete in a few events rather than all four.	This has been implemented throughout the paper, with the specific question regarding athletes only competing in a few events being answered on p. 3, third paragraph.
2	--You also state on page 3 that each performance is scored out of 10, but then later you explain this is not always the case. I would recommend saying that each performance is scored out of a	This note was implemented exactly as suggested (p. 3, second paragraph).

	maximum of 10 points, which leaves the door open for some routines being scored out of a lower number.	
2	--The first paragraph under section 2.1 contains a great deal of factual information, yet no sources are cited. Please make sure you cite sources for all of this information, otherwise it is unclear where this came from or whether it is actually accurate.	Our revised Background & Related Literature section (p. 2) now includes two citations; one is a general guide to understanding NCAA gymnastics scoring written by gymnastics journalists, and the other is a paper that explains some of the sensitivities of gymnastics scoring in its early sections.
2	--Under the 2nd paragraph of section 2.1 on page 3, you state the start value “is the score a gymnast would receive by performing their prepared routine perfectly”. For clarity’s sake, I recommend adding a brief explanation that greater difficulty in a routine leads to a higher start value.	The discussed paragraph is now the fourth paragraph of our Background & Related Literature section (beginning on p. 3). We added a lot of clarification on this point.
2	--I question the statement that “collegiate level gymnasts are typically sufficiently skilled to begin at the maximum 10 point start value”. This is really only true of the very top teams in the NCAA, and even for those teams it is usually not the case on vault, where most teams include multiple gymnasts whose maximum start value is only a 9.95. Perhaps tone this down or indicate that only the top teams in the nation tend to have all 10.0 SV routines.	In our revised Background & Related Literature section, this statement was removed in favor of the more detailed explanation of scoring variance contained in its fourth paragraph (p. 3).
2	--You do a nice job of setting up the purpose and rationale for your study on page 4.	Thank you! The sections referred to have now moved around and been emphasized throughout the paper as our motivation for the study.
2	--On page 5 you state that your research could “uncover causality”, but it is unclear how your results would do this. I don’t think you can claim to uncover a cause based on the research method you’re using.	If the parallel trends assumption holds, our diff-in-diff estimate can be interpreted as the causal effect of all race-host factors on scoring, including the potential environmental microaggression effect we focus on and any other unobserved race-host interactions. We aren’t sure what those could be, but they would play

		into any estimate as well. We clarify this point in the final two paragraphs of our Estimation Strategy section, beginning with the second paragraph on page 17.
2	--Also on page 5 you talk about similar previous research that focused on sport teams as opposed to individuals, but you should further explain the significance of studying this at the individual level. Beyond "it hasn't been done before", why should we care? What will these findings contribute to the literature and our understanding of the issue?	In the final paragraph of our Model section (p. 15, paragraph on p. 16), we suggest that coaches and gymnasts who are score maximizing may not be accounting for this potential environmental effect. In our Discussion & Conclusion section (p. 20-21), we explain that because we see no effect, they aren't missing anything (i.e. rationally accounting for [non]existing effects such that they go unobserved. We also now emphasize how this does not necessarily align with expectations in the same section.
2	--On page 6 you state that it is possible to fall off the uneven bars, but not the floor exercise. I believe later you also restate this point and include that a gymnast can fall from the beam but not the vault. I find these statements to be problematic, as gymnasts can certainly fall on the floor exercise as well as on the vault and these falls will incur the same deduction (0.5 in NCAA competition) as the falls from the bars or beam. I don't think you should be differentiating falls between events, as falls are possible on EVERY event. Does this impact your analysis at all?	The score averages in our new Table 2 (p. 11) show that gymnasts score higher on average on vault and floor than they do on beam and bars. We attributed this to the fact that completely falling off of the apparatus is only common in bars and beam routines (she fell *off* the beam) as opposed to falling on the other two events (she fell *on* her final floor pass). Regardless, because we control for event fixed effects at some level in both estimating equations, this distinction does not impact our analysis at all; we therefore removed this statement in lieu of our more detailed descriptions of general gymnastics scoring previously mentioned.
2	--In terms of the sample construction, did you only include dual meets or tri-meets? What about conference championships – were these considered "playoff" meets? You state later that you only included "non-invitational meets" but it's unclear what constitutes an invitational versus a non-invitational. Please just be more specific in explaining these criteria.playoffs,	The Sample Construction section (p. 9) is now much clearer. We motivate dropping everything but an "ordinary" meet, but we do not exclude scores from tri- or quad-meets if they meet our criteria for inclusion: 1) the meet is hosted by one university; and 2) the meet is not an invitational or playoff meet (as measured by their meet titles recorded on Road to Nationals). We believe our explanation is now much clearer and more specific.

	invitationals, etc.	
2	--While reading about the sample I had the same question that I raised earlier – was your sample limited only to Division I programs, or also Divisions II and III?	We chose to limit our sample to scores from DI gymnasts as a result of our increased scrutiny, which was motivated by this comment among others. This is clear throughout the text now.
2	--I have several questions relating to your use of photos from web pages to identify which gymnasts were coded as “Black”. What was your definition of “Black”? This is extremely important, as different researchers might code these athletes in different ways. For example, if someone’s skin tone was darker than an average white person, was this person coded as “Black”? An example of this might be a gymnast like Paityn Walker from the University of Alabama, whose hair appears to be more typically Black, but her skin tone appears white. Or what about Luisa Blanco from the same university? Her skin tone is darker than her white teammates, but her ancestry is Colombian. Does she count as “Black” or was she considered “white” for your study? And what about a gymnast like Ruthuja Nataraj from the University of Illinois? Her skin is dark, but she is of Indian ancestry. Is she considered “Black”? I think you need to be much more detailed in your explanation of how you coded gymnasts’ race, and this also begs the question asked earlier about why you did not include other races beyond Black and white.	This comment motivated our integration of FairFace, and we are particularly grateful to Reviewer 2 for its detail. Referring to the specific examples mentioned by Reviewer 2: Paityn Walker is not in our dataset, as she only began competing in 2025. However, FairFace assigned her to the Latina category; judging by her parents’ names (Tara Nunez-Walker and Mark Walker) as listed on her Alabama bio, she appears to be at least half Latina, so this prediction feels solid. Luisa Blanco is correctly predicted as Latina by FairFace. However, whether she also identifies as Black is outside the scope of our capabilities to determine; cases like this one motivate not using the Latina category on its own in our analysis. Ruthuja Nataraj is correctly predicted as Indian by FairFace. We feel that this new integration of a computer vision model is sufficient to assuage concerns about subjective predictions of race, and we relate this throughout the text.
2	--In your Table 2, it’s not clear if you are presenting the NCAA demographics of ALL NCAA athletes, or if these are the figures only of NCAA gymnasts under the “NCAA Demo.” heading. Please specify.	Table 2 (p. 11) is now very clear about this distinction, and it helps motivate our decision to narrow our sample to only scores from DI gymnasts.

2	--The results are presented in a way that is easy to understand for the reader, so thank you for that.	Thank you!
2	--I was surprised to see the paper jump from a “Results” section directly into a “Conclusion” section. There should be a “Discussion” section included, and I strongly encourage the authors to add this. Without it, the paper leaves many unanswered questions which are vital to answer if this paper wants to make a contribution in the literature. First, you presented your results in a straightforward manner so we know what you found. But what does all of this tell us? How do your findings compare to those of previous research on team sports (and individuals, although it sounds like not much has been done in that regard until now)? What do the findings mean from a practical perspective? The “so what” question is not answered in the paper’s current format. Implications need to be clearly stated and explained.	Our paper now concludes with a Discussion & Conclusion section (p. 20-21) that explains what our results imply practically and expands on specific opportunities for future research to build on our work. We believe this expanded closing section sufficiently describes our contribution to this literature and makes the implications of our results clear.
2	Overall, this is a well written paper, but there are a lot of details that should be included, a stronger justification for how you chose a gymnast’s race needs to be explained, and a Discussion section is imperative. Best wishes to the authors as you work to strengthen your manuscript.	We believe that we have strengthened our manuscript in all three categories mentioned in this comment as described in our specific replies above, and we appreciate the well wishes.