Feedback Round

Instructions Welcome to our round robin feedback system. This system works as follows: You will be shown all 28 analytical approaches in a random order. Please give detailed feedback on the first 3 reports that are shown to you. If your own is included in those then please skip to the next. You are invited to comment all other approaches and are invited to give feedback on as many as you like. After the feedback round is over we will send you an e-mail with the feedback that was submitted for your approach. Feel free share this link with collaborators from your own team, and invite them to give feedback to others. Please try to indicate your confidence for as many approaches as possible. Please aim for constructive and helpful comments and clearly point out potential weaknesses now that there is still the chance to fix things. This feedback survey will close on June 29th. The feedback is anonymous to us and to other teams! Thank you for your hard work on this project!

Dataset How confident are you that the provided dataset is suitable for answering the research questions?

	Unconfident	Rather unconfident	Somewhat unconfident	Neither confident nor unconfident	Somewhat confident	Rather confident	Confide
RQ1) Skin color and red cards	○	○	○	○	○	○	○
RQ2a) Implicit cultural preferences	○	○	○	○	○	○	○
RQ2b) Explicit cultural preferences	○	○	○	○	○	○	○

Datachanges Which different dataset or inclusion of which additional variables would increase your confidence?

QID4 Did you submit an analytical approach? (either yourself or as part of a team)

○ yes

○ no

Confid_Team3 How confident are you that the described approach below is suitable for analyzing the research questions?

	Unconfident	Rather unconfident	Somewhat unconfident	Neither confident nor unconfident	Somewhat confident	Rather confident	Confiden
RQ1) Skin color and red cards	❍	❍	❍	❍	❍	❍	❍
RQ2a) Implicit cultural preferences	❍	❍	❍	❍	❍	❍	❍
RQ2b) Explicit cultural preferences	❍	❍	❍	❍	❍	❍	❍

Team3 Please provide feedback to the analytical approach described below.

Approach3 Analytical Approach Team3   What transformations (if any) were applied to the variables. Please be specific. The rater1 and rater2 variables were averaged and rounded to the most central value. This way we could work with only one variable for player skin tone, assuming a conservative rating when the raters didn't agree. The new, averaged variable agreed with 83% of the ratings from rater1 and with 93% from rater2. / We also multiplied the meanIAT and meanExp scores by 10, because both scale have very small original values. In our model, the coefficients for those variable correspond, then, to a one-tenth increase. Were any cases excluded, and why? All cases with missing data were excluded from the analysis. Most of the missing data occurred in variables relevant to the model, like skin tone ratings and player position. To make the model more simple, we assumed that the missingness mechanism was completely at random and performed a complete-case analysis, retaining 115457 cases (79%). This assumption is strong and might be far from the truth, since a quick analysis shows that most of the missing data comes from the english and french leagues. / For the second research question, we kept only the cases which had a skin tone rating of 4 or 5, according to the averaging described above. So, for the second research question, we kept 15953 cases. What is the name of the statistical technique that you employed? Multilevel Binomial Logistic Regression using bayesian inference. Please describe the statistical technique you chose in more detail. Be specific, especially if your choice is not one you consider to be well-known. The binomial logistic regression is a generalization of the logistic regression, analyzing z successes in N trials (in traditional logistic regression, the number of trial is always equal to 1). We used a binomial-normal model to account for possible overdispersion in the data. / We adopt a multilevel approach because the dataset is structured in nested and non-nested groups which allows us to model the coefficients as varying by group. / Finally, we estimate the model coefficients using bayesian inference with Markov-Chain Monte Carlo sampling. The bayesian framework allows us to treat the parameters as random variables with prior distributions, incorporating evidence from data to update the probability distribution. The resulting posterior distribution can be used to estimate a point value for the parameter using probability to quantify our uncertainty in the estimate. What are some references for the statistical technique that you chose? Gelman, Andrew & Hill, Jeniffer (2007). Data analysis using regression and multilevel/hierarchical models. New York: Cambridge University Press. Which software did you use? If you used multiple kinds, please indicate what was accomplished with each piece of software (e.g., Data cleaning - R; Model estimation - SAS) R 3.0.2 with RStudio - Data cleaning, Exploratory Data Analysis; Stan and rstan 2.2.0 - model specification, fitting and parameter estimation. What distribution did you specify for the outcome variable of red cards? We modeled the red cards in a player-referee dyad using a binomial distribution, representing each red card as a 'success' in N trials, represented by the number of games for the dyad: / / redcard ~ Binomial(games, p) -- for each player referee dyad. / / The rationale for the binomial distribution comes from the rules of the game: a player can receive only one red card per game; a (pure) red card can be assumed to be independent from game to game and also independent from any yellow card (which is not the case for yellow-red cards). What variables were included as covariates (or control variables) when testing research question 1: The relationship between player skin tone and red cards / received? Player position; / League country. What variables were included as covariates (or control variables) when testing research question 2a: The relationship between referee country implicit skin-tone prejudice and red cards received by dark skin-toned players? Player position; / League country. What variables were included as covariates (or control variables) / when testing research question 2b: The relationship between referee country explicit skin-tone prejudice and red cards received by dark skin-toned players? Player position; / League country. What theoretical and/or statistical rationale was used for your choice of covariates included in the models? We controlled for player position because some field positions are more likely to receive a red card, since their role make them commit more fouls. We also controlled for league country because some leagues are more strict in applying the rules, so the referees are more likely to give red cards to different players. / Earlier versions of the model included more control variables, but they seemed to affect little the estimation of the main effects and had high uncertainty. What unit is your effect size in? Odds Ratio

Confid_Team16 How confident are you that the described approach below is suitable for analyzing the research questions?

	Unconfident	Rather unconfident	Somewhat unconfident	Neither confident nor unconfident	Somewhat confident	Rather confident	Confiden
RQ1) Skin color and red cards	❍	❍	❍	❍	❍	❍	❍
RQ2a) Implicit cultural preferences	❍	❍	❍	❍	❍	❍	❍
RQ2b) Explicit cultural preferences	❍	❍	❍	❍	❍	❍	❍

Team16 Please provide feedback to the analytical approach described below.

Approach16 Analytical Approach Team16 What transformations (if any) were applied to the variables. Please be specific. - Red card data was dichotomized. For the first research question, data was aggregated to the level of each individual player such that each record in the dataset represented a unique player and the dichotomized outcome variable indicated whether or not the player had ever received a red (or yellow-red) throughout the games represented in the set. For the second research question, the data was dichotomized at the level of the referee-player dyad such that the outcome variable indicated whether a red (or yellow-red) had been issued to the player within the context of the dyad. / - For the first research question, an inverse transformation was applied the 'goals' variable (aggregated at the level of the unique player) to correct positive skewness. / - For the first research question, dummy variables were created for each of the four levels of the 'leagueCountry' variable for the purpose of investigating interactions in the analysis. / - For the second research question, a log transformation was applied to the 'games' variable to correct positive skewness / - For both research questions, a skin tone index was created by taking the average of the two ratings provided. / - Player's age was calculated as the number of years from the player's birthdate to January 1, 2013. Were any cases excluded, and why? For the first research question, cases with missing skin tone ratings were excluded. For the second research question, cases with missing skin tone or prejudice ratings were excluded. In both cases, the cases were dropped so that analysis across multiple levels would be based on the same number of cases, allowing for appropriate comparison of models across steps. What is the name of the statistical technique that you employed? Hierarchical logistic regression was used in both research questions Please describe the statistical technique you chose in more detail. Be specific, especially if your choice is not one you consider to be well-known. In the first research question, hierarchical logistic regression was used to investigate whether the player's skin tone was predictive of whether they had ever received a red card (or yellow red).   / / For the second research question, hierarchical logistic regression was used to investigate whether skin tone, prejudice level of the referee's home country, and the interaction between skin tone and prejudice were predictive of whether or not the player had been ever been issued a red card (or yellow red) by the referee. Two separate models were constructed, one which used the explicit measure of prejudice as a predictor, and one which used the implicit (IAT) measure of prejudice as a predictor. / / This is the first time that I have used hierarchical logistic regression, so feedback on my analysis approach would be very helpful. What are some references for the statistical technique that you chose? Field (2012) - 'Discovering Statistics Using R' was used as the primary statistics reference for this analysis Which software did you use? If you used multiple kinds, please indicate what was accomplished with each piece of software (e.g., Data cleaning - R; Model estimation - SAS) R (version 3.01) was used for all analysis What distribution did you specify for the outcome variable of red cards? Bivariate What variables were included as covariates (or control variables) when testing research question 1: The relationship between player skin tone and red cards / received? games, age, and goals were used as covariates for research question 1. Though these variables are correlated with each other, investigation of multicolinearity indicated that this should not be problematic in the logistic regression, and all three variables predicted a significant amount of unique variance in the outcome. What variables were included as covariates (or control variables) when testing research question 2a: The relationship between referee country implicit skin-tone prejudice and red cards received by dark skin-toned players? Only 'games' was used as a control variable in research question 2a, as it was the only potential covariate that had a significant relation to the outcome. What variables were included as covariates (or control variables) / when testing research question 2b: The relationship between referee country explicit skin-tone prejudice and red cards received by dark skin-toned players? Again, only the 'games' variable was used. What theoretical and/or statistical rationale was used for your choice of covariates included in the models? For the sake of simplicity in generalizing any potential findings, only physical characteristics of the player (height, weight, age) and variables that were potentially associated with the quantity of the referee player interaction (games, goals, victories, ties, defeats) were considered for covariates. The unique association between all possible covariates and the outcome variable was assessed first. Any variables that demonstrated a significant relation to the outcome at this level were retained for a second step, which investigated multicolinearlity between the possible covariates. If multicolinearity was problematic, the decision regarding which variables to remove was subjective. This was only an issue once, as the 'victories', 'ties', 'defeats' and 'games' variables were highly related to one another in the analysis for research question 1. As the relationship was almost certainly a product of the overall number of games played, 'games' was retained as a covariate and the other three variables were dropped. What unit is your effect size in? odds ratio

Confid_Team1 How confident are you that the described approach below is suitable for analyzing the research questions?

	Unconfident	Rather unconfident	Somewhat unconfident	Neither confident nor unconfident	Somewhat confident	Rather confident	Confiden
RQ1) Skin color and red cards	❍	❍	❍	❍	❍	❍	❍
RQ2a) Implicit cultural preferences	❍	❍	❍	❍	❍	❍	❍
RQ2b) Explicit cultural preferences	❍	❍	❍	❍	❍	❍	❍

Team1 Please provide feedback to the analytical approach described below.

Approach1 Analytical Approach Team1   What transformations (if any) were applied to the variables. Please be specific. The main transformation made to these data was to change the unit of observation in the data set from a player-referee dyad observation, to a player-referee-game observation. Therefore, a player-referee dyad observation with 8 games was transformed into 8 different player-referee'game observations. This expanded the number of rows in the data set to equal the number of games played by players instead of the number of dyads that existed. In addition, certain control variables were transformed in minor ways. Namely we included squared values of age, height, and weight as addition controls and created an average rater value (a simple average of the two skin-tone ratings). We also create a standardized mean iat score and exp score with mean 0 and standard deviation 1 for each of the countries with iat and exp scores. Were any cases excluded, and why? We did not exclude any cases. For 4 dyad observations, the dyad consisted of 1 game but 2 yellow cards were issued. Typically this should be coded as a yellowred or 'soft' red card but shows up in the data as 2 yellow cards and no red cards. For these 4 observations we recoded the yellow card as 1 instead of 2 and left the redcards and yellowred variables as 0. Since this is such a small fraction of the cards awarded we see this as a trivial change. What is the name of the statistical technique that you employed? We use a variety of different regressions. First, we use ordinary least squares with robust standard errors and control for various things such as height, weight, age. We also add in fixed effects for league country, position, club, and referee. In addition, we employ a logistic regression and nonlinear regression to compare with our OLS regressions. Please describe the statistical technique you chose in more detail. Be specific, especially if your choice is not one you consider to be well-known. OLS is a canonical economic and statistical approach which calculates a linear 'best-fit' relationship between variables with the goal of minimizing the sum of square errors. Fixed effects demeans the data in such a way that in essence allows us to perform a within referee, league country, club, or position analysis. The attached tables show the specific specifications used. What are some references for the statistical technique that you chose? Aldrich, John Herbert, and Forrest D. Nelson. Linear probability, logit, and probit models. Vol. 45. Sage, 1984. Bhargava, Alok, Luisa Franzini, and Wiji Narendranathan. "Serial correlation and the fixed effects model." The Review of Economic Studies 49.4 (1982): 533-549. Which software did you use? If you used multiple kinds, please indicate what was accomplished with each piece of software (e.g., Data cleaning - R; Model estimation - SAS) STATA What distribution did you specify for the outcome variable of red cards? Bernoulli distribution. Either a player gets a red card or he does not. The probability of a player receiving a red card is p. What variables were included as covariates (or control variables) when testing research question 1: The relationship between player skin tone and red cards / received? Our regression specification is as follows: / / Yi = a + bskintonei + cXi + di + ei + fi + gi + ui (1) / / where Yi is the outcome of interest (e.g. number of yellow cards, hard red cards, or any red cards received) for a specific referee-match-player observation i. The variable skintonei is a categorical variable with values 1-5 ranging from 'very light skin' to 'very dark skin' with 'neither dark nor light skin' as the center value. Additionally, Xi is a vector of player characteristics including height, height-squared, weight, weight-squared, age, and age-squared. The variables di, ei, fi, and gi are referee, league country, club, and player position fixed effects respectively. / Our logistic and nonlinear regressions include the same controls. The controls are: height-squared, weight, weight-squared, age, age-squared, as well as referee, league country, club, and position fixed effects. / What variables were included as covariates (or control variables) when testing research question 2a: The relationship between referee country implicit skin-tone prejudice and red cards received by dark skin-toned players? To answer question 2a the following regression specification is estimated separately for light skin-tone players (ie rating equals one or two) and for dark skin tone players (ie rating equals three, four, or five): / / Yi = a + brefereenationsimplicitscorei + ci + ui   (2) / / where Yi is the outcome of interest (e.g. number of yellow cards, hard red cards, or any red cards received) for a specific referee-match-player observation i. The variable refereenationsimplicitscorei is the normalized IAT score for the referee's home country. The variable ci is a player fixed effect. Equation (2) is estimated separately for light and dark skin-tone players. The estimate to answer question 2a is the difference between these two estimates. The only covariates included are player fixed effects.   / What variables were included as covariates (or control variables) / when testing research question 2b: The relationship between referee country explicit skin-tone prejudice and red cards received by dark skin-toned players? This is done in the same way as question 2a. To answer question 2b the following regression specification is estimated separately for light skin-tone players (ie rating equals one or two) and for dark skin tone players (ie rating equals three, four, or five): / / Yi = a + brefereenationsexplicitscorei + ci + ui   (3) / / where Yi is the outcome of interest (e.g. number of yellow cards, hard red cards, or any red cards received) for a specific referee-match-player observation i. The variable refereenationsexplicitscorei is the normalized explicit score for the referee's home country. The variable ci is a player fixed effect. Equation (3) is estimated separately for light and dark skin-tone players. The estimate to answer question 2b is the difference between these two estimates. The only covariates included are player fixed effects. / What theoretical and/or statistical rationale was used for your choice of covariates included in the models? We are concerned with whether or not having darker skin causes you to receive more cautionary cards and thus we try to control for anything that reasonably could affect cautionary cards. For example, perhaps (and it appears to be the case) defenders receive many more cautionary cards than the average forward. Since this is the case we must control for it either in the form of a player-position fixed effect or a more over-arching player fixed effect. Ideally we would be able to control for all characteristics correlated with skin-tone and correlated with being given a card. Where this is not possible, we include the statistics that reasonably seem may have a relationship with cards and skin-tone. The size, position, and league of the player all seem like possible candidates. As can be seen in our tables, the addition of covariates allow for robustness checks. Due to the lack of an identification strategy for question 1, there is no way of knowing if the estimates that come from our estimates for question 1 are causal. There are likely other variables that should be controlled for but are not. I would only believe our estimates for questions 1 with skepticism. For questions 2a and 2b a good identification strategy is used. These estimates compare a player with himself when refereed by referees from countries of different levels of implicit and explicit biases. In this particular case we are able to compare this effect for light and dark skin-toned players. This allows us to have more confidence in these estimates and their causal nature. What unit is your effect size in? The unit of our effect size is the change in probability of a card.

Confid_Team2 How confident are you that the described approach below is suitable for analyzing the research questions?

	Unconfident	Rather unconfident	Somewhat unconfident	Neither confident nor unconfident	Somewhat confident	Rather confident	Confide
RQ1) Skin color and red cards	○	○	○	○	○	○	○
RQ2a) Implicit cultural preferences	○	○	○	○	○	○	○
RQ2b) Explicit cultural preferences	○	○	○	○	○	○	○

Team2 Please provide feedback to the analytical approach described below.

Approach2 Analytical Approach Team2 What transformations (if any) were applied to the variables. Please be specific. Several versions of the skin-tone rating were used (min, max, mean, first rating, second rating) as well as a binary "dark skin tone" variable (where rating=4 or 5). / / Redcards was tranformed into a binary variable. (There were very few instances with 2 redcards to begin with.) / Were any cases excluded, and why? No. In my opinion there didn't appear to be any outliers, at least in terms of redcards, physical characteristics (height, weight), games played, etc. Even if there were outliers, I likely would have only excluded them if I thought the data was simply incorrect. What is the name of the statistical technique that you employed? Linear regression, logistic regression Please describe the statistical technique you chose in more detail. Be specific, especially if your choice is not one you consider to be well-known. I didn't limit myself to just one form of analysis, or just one specification (economists never do). I first ran a linear regression (Ordinary Least Squares) with the original, non-binary redcards, but I don't put much stock in that, since very few players have received more than one redcard. (I'd likely only put these results in an appendix, if I included them at all.) / / I put more emphasis on a linear regression with a binary redcard outcome. I again use linear regression (called the Linear Probability Model since there's a binary outcome.) I also compare this with results from a logistic regression. In my experience in economics, the standard is to run an LPM for ease of interpretability, but then add logit or probit as a robustness check. I get very similar results with the LPM and the logit, so I don't see any need to debate which model is better. / / In certain specifications I cluster the standard errors by player, since observations of the same player by different referees are almost certainly not independent. What are some references for the statistical technique that you chose? OLS/LPM and Logit regression are hopefully obvious (See Angrist & Pischke--Mostly Harmless Econometrics, or Wooldridge's Econometric Analysis of Cross Section and Panel Data). For clustering, see the previous, or Moulton, Review of Economics and Statistics, 1990. "An illustration of a pitfall in estimating the effects of aggregate variables on micro units" Which software did you use? If you used multiple kinds, please indicate what was accomplished with each piece of software (e.g., Data cleaning - R; Model estimation - SAS) Stata What distribution did you specify for the outcome variable of red cards? By using logistic regression I assumed a logit distribution. Honestly, I didn't give this any thought--I've never seen a serious difference arise from choosing logit over probit, and a quick test showed that no significant differences arise in this case. / By using the linear probability model, I assume the errors are normally distributed. I don't observe significant differences when using these different assumptions. What variables were included as covariates (or control variables) when testing research question 1: The relationship between player skin tone and red cards / received? Height, weight, player position, and player league. I've also run specifications with games played, victories, and goals scored. What variables were included as covariates (or control variables) when testing research question 2a: The relationship between referee country implicit skin-tone prejudice and red cards received by dark skin-toned players? The same: height, weight, position, and league. Of course I also included the prejudiced measure and interacted prejudice with the redcard variable. / What variables were included as covariates (or control variables) / when testing research question 2b: The relationship between referee country explicit skin-tone prejudice and red cards received by dark skin-toned players? The same: height, weight, position, and league. Of course I also included the prejudiced measure and interacted prejudice with the redcard variable. What theoretical and/or statistical rationale was used for your choice of covariates included in the models? The question, as written, is "Are soccer referees more likely to give red cards to dark skin toned players than light skin toned players?" Of course I have assumed that the actual question of interest is whether dark skin toned players receive more redcards ceteris paribus, or, solely because they have darker skin and not because of something else. If that is indeed the question we want to answer, then it seems obvious that we want to control for as many other characteristics as we can that might potentially bias the coefficient on skin tone, given that skin tone is obviously not randomly assigned. Coming up with a story of omitted variable bias for all the included controls is simple: for example, if darker skin toned players are more (or less) prevalent in Spain's league, and Spain's league gives out more (or fewer) redcards, then the coefficient on skin tone in a regression without league of play would be biased. What unit is your effect size in? Marginal Probability

Confid_Team4 How confident are you that the described approach below is suitable for analyzing the research questions?

	Unconfident	Rather unconfident	Somewhat unconfident	Neither confident nor unconfident	Somewhat confident	Rather confident	Confide
RQ1) Skin color and red cards	○	○	○	○	○	○	○
RQ2a) Implicit cultural preferences	○	○	○	○	○	○	○
RQ2b) Explicit cultural preferences	○	○	○	○	○	○	○

Team4 Please provide feedback to the analytical approach described below.

Approach4 Analytical Approach Team4 What transformations (if any) were applied to the variables. Please be specific. I made a new variable: red cards / games. Were any cases excluded, and why? Cases with no skin tone rating from either of the two reviewers. What is the name of the statistical technique that you employed? Correlations and partial correlations. Please describe the statistical technique you chose in more detail. Be specific, especially if your choice is not one you consider to be well-known. N/A What are some references for the statistical technique that you chose? N/A Which software did you use? If you used multiple kinds, please indicate what was accomplished with each piece of software (e.g., Data cleaning - R; Model estimation - SAS) MATLAB What distribution did you specify for the outcome variable of red cards? I did not specify a distribution (i.e., a point-biserial correlation). What variables were included as covariates (or control variables) when testing research question 1: The relationship between player skin tone and red cards / received? For the main analysis I correlated the coded skin tone with the [red cards/games] measure. I also tested for this relationship within each individual referee's country (with corrections for multiple comparisons). For comparison, I also conducted control correlations between skin tone and the player's height and weight. What variables were included as covariates (or control variables) when testing research question 2a: The relationship between referee country implicit skin-tone prejudice and red cards received by dark skin-toned players? I multiplied the mean IAT score with the [red cards/games] measure, effectively making it a weight score. I then correlated the coded skin tone with this measure. No covariates were included. What variables were included as covariates (or control variables) / when testing research question 2b: The relationship between referee country explicit skin-tone prejudice and red cards received by dark skin-toned players? I multiplied the mean explicit bias score with the [red cards/games] measure, effectively making it a weight score. I then correlated the coded skin tone with this measure. No covariates were included. What theoretical and/or statistical rationale was used for your choice of covariates included in the models? Height and weight correlations were conducted to test for the observed relationship strength (r) when a 'variable of no-interest' was used. Since the number of data points here is so large, this provides a control correlation to additionally evaluate the other results against, apart from the absolute effect size. What unit is your effect size in? r

Confid_Team7 How confident are you that the described approach below is suitable for analyzing the research questions?

	Unconfident	Rather unconfident	Somewhat unconfident	Neither confident nor unconfident	Somewhat confident	Rather confident	Confiden
RQ1) Skin color and red cards	○	○	○	○	○	○	○
RQ2a) Implicit cultural preferences	○	○	○	○	○	○	○
RQ2b) Explicit cultural preferences	○	○	○	○	○	○	○

Team7 Please provide feedback to the analytical approach described below.

Approach7 Analytical Approach Team7 What transformations (if any) were applied to the variables. Please be specific. None. Were any cases excluded, and why? I ran the analysis only for the pairs of player-referees where at least 1 red card had been given. This is due to the large number of pairs where no red cards were given. I also excluded all players for which the skin colour rating hadn't been recorded. What is the name of the statistical technique that you employed? Profile regression, which is a Dirichlet process Bayesian clustering and is implemented in the R package PReMiuM. Please describe the statistical technique you chose in more detail. Be specific, especially if your choice is not one you consider to be well-known. Standard regression methods encounter issues when they are used to make inferences with covariates which are highly correlated. As the covariates available for this study are highly correlated (rater1 and rater2, for example), we chose to use profile regression (see references below). This method addresses this issue by using a profile formed from a sequence of covariate values and then clusters them into groups associated via a regression model to the relevant outcome. / What are some references for the statistical technique that you chose? J. T. Molitor, M. Papathomas, M. Jerrett and S. Richardson (2010) Bayesian Profile Regression with an Application to the National Survey of Childrens Health, Biostatistics, 11, 484-498. Liverani, S., Hastie, D. I., Azizi, L., Papathomas, M. and Richardson, S. (2014) PReMiuM: An R package for Profile Regression Mixture Models using Dirichlet Processes. Forthcoming in the Journal for Statistical Software. Available at http://uk.arxiv.org/abs/1303.2836 Which software did you use? If you used multiple kinds, please indicate what was accomplished with each piece of software (e.g., Data cleaning - R; Model estimation - SAS) R What distribution did you specify for the outcome variable of red cards? Binomial distribution, with the number of games as the parameter n of the Binomial distribution. (equivalent to modelling the average number of red cards per game) / What variables were included as covariates (or control variables) when testing research question 1: The relationship between player skin tone and red cards / received? Red cards, games, rater1 and rater2. What variables were included as covariates (or control variables) when testing research question 2a: The relationship between referee country implicit skin-tone prejudice and red cards received by dark skin-toned players? Red cards, games, rater1, rater2, meanIAT. I have only included the observations for which nIAT was greater than 5,000. However, there isn't much variability between countries if the sample is large enough (hence my doubts about the validity of this measure). What variables were included as covariates (or control variables) / when testing research question 2b: The relationship between referee country explicit skin-tone prejudice and red cards received by dark skin-toned players? Red cards, games, rater1, rater2, meanExp. I have only included the observations for which nIAT was greater than 5,000. However, there isn't much variability between countries if the sample is large enough (hence my doubts about the validity of this measure). What theoretical and/or statistical rationale was used for your choice of covariates included in the models? We chose to use the minimum number of covariates to shade light on the underlying process. By using other covariate the result can be subdivided in additional clusters. We felt this was unnecessary to answer the questions asked in this project. What unit is your effect size in? Relative risk, with the cluster of people rated as light skin as the reference.

Confid_Team5 How confident are you that the described approach below is suitable for analyzing the research questions?

	Unconfident	Rather unconfident	Somewhat unconfident	Neither confident nor unconfident	Somewhat confident	Rather confident	Confide
RQ1) Skin color and red cards	❍	❍	❍	❍	❍	❍	❍
RQ2a) Implicit cultural preferences	❍	❍	❍	❍	❍	❍	❍
RQ2b) Explicit cultural preferences	❍	❍	❍	❍	❍	❍	❍

Team5 Please provide feedback to the analytical approach described below.

Approach5 Analytical Approach Team5 What transformations (if any) were applied to the variables. Please be specific. The two ratings of skin-tone were averaged and rescaled to the range between 0 (corresponding to the original value of 1) and 1 (corresponding to the original value of 5). This was done so that the difference between brightest and darkest players could later on be read off the regression coefficients directly. Were any cases excluded, and why? Cases were excluded if they had missing values on skin-tone-rating, meanIAT or meanExp (listwise deletion) because we wanted to perform all analyses (including research question 2) on the same set of cases. What is the name of the statistical technique that you employed? Mixed models (a.k.a. multilevel modeling, hierarchical linear models). Please describe the statistical technique you chose in more detail. Be specific, especially if your choice is not one you consider to be well-known. We estimated generalized linear mixed models (function glmer in R package lme4). With this technique we can estimate the desired effects while accounting for random variance of the effects across players, referees, and referees' country of origin. The crowdstorming data deviate from the standard multilevel data (e.g., where employees are members of only one team), inasmuch as they are not nested but cross-classified ' player A can have multiple games with the referee A, but player B can have multiple games with the same referee A. The package lme4 can deal with such a data structure. What are some references for the statistical technique that you chose? Bates, D.M. (2010). lme4: Mixed-effects modeling with R. Available from http://lme4.r-forge.r-project.org/lMMwR/lrgprt.pdf Which software did you use? If you used multiple kinds, please indicate what was accomplished with each piece of software (e.g., Data cleaning - R; Model estimation - SAS) R What distribution did you specify for the outcome variable of red cards? Consistent with the original research question 1 (Are soccer referees more likely to give red cards to dark skin toned players than light skin toned players?) we were interested in the likelihood of a player receiving a red card in a single game. The original response variable redCards is uninterpretable because the number of games a player has seen a given referee varies. Therefore we disaggregated the data (one game per row, redCards appear as 1s in the first n=redCards rows per player). This was possible because it does not matter in which of, for example, 3 games a player who received 1 red card in 3 games received the red card. It is sufficient that this player has three observations (three rows) associated with him, one of them indicating a red card. We used the binomial error distribution because ' after disaggregation ' our response variable specifies the occurence of an event in a single game, coded 0 and 1. What variables were included as covariates (or control variables) when testing research question 1: The relationship between player skin tone and red cards / received? none What variables were included as covariates (or control variables) when testing research question 2a: The relationship between referee country implicit skin-tone prejudice and red cards received by dark skin-toned players? none What variables were included as covariates (or control variables) / when testing research question 2b: The relationship between referee country explicit skin-tone prejudice and red cards received by dark skin-toned players? none What theoretical and/or statistical rationale was used for your choice of covariates included in the models? One obvious control variable would have been the number of games. However, we used a different approach (disaggregation, see above) to control for number of games. What unit is your effect size in? Odds-ratio = odds of getting a red card in a single game of a player with the darkest skin-tone (rating = 5) over the odds of getting a red card in a single game of a player with the brightest skin-tone (rating = 1) .

Confid_Team6 How confident are you that the described approach below is suitable for analyzing the research questions?

	Unconfident	Rather unconfident	Somewhat unconfident	Neither confident nor unconfident	Somewhat confident	Rather confident	Confide
RQ1) Skin color and red cards	○	○	○	○	○	○	○
RQ2a) Implicit cultural preferences	○	○	○	○	○	○	○
RQ2b) Explicit cultural preferences	○	○	○	○	○	○	○

Team6 Please provide feedback to the analytical approach described below.

Approach6 Analytical Approach Team6   What transformations (if any) were applied to the variables. Please be specific. None Were any cases excluded, and why? I am taking the stand that a referee views skin color as an informative signal of the likelihood of a player committing a red card worthy infraction. This may be an immutable erroneous prior of the referee in such that there is no information in this particular signal, or it could be a correct belief in that skin color acts as a visible signal of the style of play of a particular player in a game. Either way the relationship between the probability of a player being issued a red card and skin color will be weakened as the referee observes the player through multiple interactions. As such I restrict the sample to player-referee dyads that are the result of one interaction between the player and the referee. What is the name of the statistical technique that you employed? Linear probability model Please describe the statistical technique you chose in more detail. Be specific, especially if your choice is not one you consider to be well-known. My primary empirical analysis is based on a linear probability model where the incidence of a red flag is a function of the color of a player's skin, and depending on the specification, player characteristics such as height, weight, and age, the position of the player, club fixed effects, and referee fixed effects. I use an indicator variable for a particular rater's skin tone classification in order to account for any non-linearities. t-statistics are clustered at the referee country level to account for any common training in referees. / / The indicator for category 5 is the difference in the probability between the lightest skin player and the darkest skin player being issued a red card. / / In this analysis I use a linear probability model because in this context indicator variables for club, position, and referee will remove unobserved category-invariant characteristics that will bias my estimates. This is not true for non-linear models. What are some references for the statistical technique that you chose? Which software did you use? If you used multiple kinds, please indicate what was accomplished with each piece of software (e.g., Data cleaning - R; Model estimation - SAS) STATA What distribution did you specify for the outcome variable of red cards? Binary What variables were included as covariates (or control variables) when testing research question 1: The relationship between player skin tone and red cards / received? Depending on the specification, player characteristics such as height, weight, and age, the position of the player, club fixed effects, and referee fixed effects. What variables were included as covariates (or control variables) when testing research question 2a: The relationship between referee country implicit skin-tone prejudice and red cards received by dark skin-toned players? Depending on the specification, player characteristics such as height, weight, and age, the position of the player, club fixed effects, and an interaction term with an above mean indicator for high IAT. What variables were included as covariates (or control variables) / when testing research question 2b: The relationship between referee country explicit skin-tone prejudice and red cards received by dark skin-toned players? Depending on the specification, player characteristics such as height, weight, and age, the position of the player, club fixed effects, and an interaction term with an above mean indicator for high Exp. What theoretical and/or statistical rationale was used for your choice of covariates included in the models? Referee effort of observation could be conditional on any observable characteristics. What unit is your effect size in? Fraction i.e. unit probability

Confid_Team8 How confident are you that the described approach below is suitable for analyzing the research questions?

	Unconfident	Rather unconfident	Somewhat unconfident	Neither confident nor unconfident	Somewhat confident	Rather confident	Confiden
RQ1) Skin color and red cards	❍	❍	❍	❍	❍	❍	❍
RQ2a) Implicit cultural preferences	❍	❍	❍	❍	❍	❍	❍
RQ2b) Explicit cultural preferences	❍	❍	❍	❍	❍	❍	❍

Team8 Please provide feedback to the analytical approach described below.

Approach8 Analytical Approach Team8 What transformations (if any) were applied to the variables. Please be specific. For RQ1: No transformations were applied other than collapsing the data across playerShort, keeping RateAve for the player. We also created the variables Sum_Games that aggregated across the variable games to provide the summed the number of games played by the player across all player-referee dyads, and RedCards_sum that was the sum of redCards and therefore represented the number of red cards a player received from all referees encountered. The resulting dataset was comprised of 1585 unique cases. / / For RQ2: no transformations were applied. / Were any cases excluded, and why? For RQ1: Since there is interest in determining whether a difference in red cards exists by skin tone all cases with missing rater scores about skin tone were removed. / / For RQ2: So that every case had full values all cases with missing rater scores about skin tone, red cards, player position, games played, mean IAT and mean EXP were removed. / What is the name of the statistical technique that you employed? For RQ1: ANOVA, Linear Regression For RQ2: Linear Regression Please describe the statistical technique you chose in more detail. Be specific, especially if your choice is not one you consider to be well-known. For RQ1: We decided an ANOVA best tested RQ1 that there was a difference in the number of red cards earned between different groups (e.g., dark skin toned players and light skin toned players). Given the results we were motivated to test for a linear relationship between skin tone and red cards earned to see if a predictive relationship exists among variables. / / For RQ2: We decided that a linear regression best tested the relationship between skin tone, referee country prejudices (IAT and EXP) and number of red cards because we were looking for predictive relationships among variables. / For RQ1: We decided an ANOVA best tested RQ1 that there was a difference in the number of red cards earned between different groups (e.g., dark skin toned players and light skin toned players). Given the results we were motivated to test for a linear relationship between skin tone and red cards earned to see if a predictive relationship exists among variables. / / For RQ2: We decided that a linear regression best tested the relationship between skin tone, referee country prejudices (IAT and EXP) and number of red cards because we were looking for predictive relationships among variables. / What are some references for the statistical technique that you chose? We referenced Aiken and West (1991) as the approached used for dummy coded variables. Aiken, L. S., & West, S. G. (1991). Multiple regression: Testing and interpreting interactions. Newbury Park, CA: Sage. Which software did you use? If you used multiple kinds, please indicate what was accomplished with each piece of software (e.g., Data cleaning - R; Model estimation - SAS) For data cleaning we used Excel and for all analysis we used SPSS. What distribution did you specify for the outcome variable of red cards? We assumed that the outcome variable was normally distributed. Furthermore, the results of the assumption tests (including multicollinearity, singularity, normality, linearity, independence of errors, and non-independence of error) did not lead to any data transformations or deletion of cases because of assumption violations. What variables were included as covariates (or control variables) when testing research question 1: The relationship between player skin tone and red cards / received? While we did test for the difference in red cards by player skin tone (averaged across rater1 and rater 2), we also tested a model that controlled for games played by dividing number of red cards by number of games played. What variables were included as covariates (or control variables) when testing research question 2a: The relationship between referee country implicit skin-tone prejudice and red cards received by dark skin-toned players? We controlled for games played using the continuous variable in the dataset because players that participate in more games have more opportunity to receive red cards than players that participate in fewer games. We also controlled for the player position because certain soccer positions may be more prone to receiving red cards than other positions. What variables were included as covariates (or control variables) / when testing research question 2b: The relationship between referee country explicit skin-tone prejudice and red cards received by dark skin-toned players? We controlled for games played using the continuous variable in the dataset because players that participate in more games have more opportunity to receive red cards than players that participate in fewer games. We also controlled for the player position because certain soccer positions may be more prone to receiving red cards than other positions. What theoretical and/or statistical rationale was used for your choice of covariates included in the models?   What unit is your effect size in? R2

Confid_Team12 How confident are you that the described approach below is suitable for analyzing the research questions?

	Unconfident	Rather unconfident	Somewhat unconfident	Neither confident nor unconfident	Somewhat confident	Rather confident	Confiden
RQ1) Skin color and red cards	○	○	○	○	○	○	○
RQ2a) Implicit cultural preferences	○	○	○	○	○	○	○
RQ2b) Explicit cultural preferences	○	○	○	○	○	○	○

Team12 Please provide feedback to the analytical approach described below.

Approach12 Analytical Approach Team12   What transformations (if any) were applied to the variables. Please be specific. No transformations applied. Skin color rating ("darkSkin") was created with the average of the two raters. Were any cases excluded, and why? Cases without skin color ratings were removed from the dataset. What is the name of the statistical technique that you employed? Zero-inflated poisson (ZIP) regression Please describe the statistical technique you chose in more detail. Be specific, especially if your choice is not one you consider to be well-known. Since the

DV (red cards received in a ref-player dyad) is count data, a standard OLS cannot be used. (Violation of normality assumption.) A standard approach with count data is to employ some version of a Poisson distribution. Since these data are relatively rare events, there is a much higher number of zeros than a standard Poisson distribution can account for. There are two methods to deal with a "high" zero distribution in count data--a hurdle distribution and a zero-inflated distribution. Both of these are mixture distributions, in which one distribution is chosen for the zeroes and another for the non-zero counts. A hurdle distribution assumes there is some sort of "threshold" that must be passed, after which counts are distributed in a particular manner. A zero-inflated distribution (used in this analysis) makes no modeling assumptions about the causes for zeroes. Of important note in both of these appraches is that coefficients are produced for both parts of the distrbution, so we can answer both "what creates non-zero counts of red cards" and "what creates more red cards in one dyad than another". (UCLA has a good page describing this model: http://www.ats.ucla.edu/stat/r/dae/zipoisson.htm) What are some references for the statistical technique that you chose? Zeileis, Kleiber and Jackman 2007, Jackman 2008, Cameron and Trivedi (1998, 2005), Lambert 1992, Which software did you use? If you used multiple kinds, please indicate what was accomplished with each piece of software (e.g., Data cleaning - R; Model estimation - SAS) R was used for both data exploration and model estimation, with the pscl() package providing the ZIP regression functions What distribution did you specify for the outcome variable of red cards? Zero-inflated Poisson What variables were included as covariates (or control variables) when testing research question 1: The relationship between player skin tone and red cards / received? In all models, the following covariates were included: / weight / position / games What variables were included as covariates (or control variables) when testing research question 2a: The relationship between referee country implicit skin-tone prejudice and red cards received by dark skin-toned players? In all models, the following covariates were included: / weight / position / games What variables were included as covariates (or control variables) / when testing research question 2b: The relationship between referee country explicit skin-tone prejudice and red cards received by dark skin-toned players? In all models, the following covariates were included: / weight / position / games What theoretical and/or statistical rationale was used for your choice of covariates included in the models? weight: heavier players may be slower, and thus more prone to physical play to slow down their faster competitors (http://thomasswan.hubpages.com/hub/Top-10-Most-Red-Cards-Premier-League) / position: Some positions may have more opportuntity for physical play than others / games: More games played = more opportunities for red cards What unit is your effect size in? There are two effects reported: (1) the effect of the predictor causing the DV to be non-zero (these effects are logit coefficients, or log-odds) , and (2) the effect of the predictor causing a non-zero DV to be higher (these effects are log coefficients)

Confid_Team11 How confident are you that the described approach below is suitable for analyzing the research questions?

	Unconfident	Rather unconfident	Somewhat unconfident	Neither confident nor unconfident	Somewhat confident	Rather confident	Confiden
RQ1) Skin color and red cards	○	○	○	○	○	○	○
RQ2a) Implicit cultural preferences	○	○	○	○	○	○	○
RQ2b) Explicit cultural preferences	○	○	○	○	○	○	○

Team11 Please provide feedback to the analytical approach described below.

Approach11 Analytical Approach Team11   What transformations (if any) were applied to the variables. Please be specific. This might not be the type of transformations that you are asking about, but for all analyses, a new player skin rating variable was derived which was simply the arithmetic average of the 'rater 1' and 'rater 2' variables in the original dataset (this was justified because of the high interrater reliability of .924, p < .001). This new average skin tone variable was the one used in the analyses below, and will be referred to as average skin tone. / The birthdate variable was used to create a new variable representing player age. This was done by choosing an arbitrary date within the soccer season under consideration (01/01/2013) and then calculating player age based on their birthdate and this arbitrary date. This variable will be referred to as player age below. / The position variable was used to create 12 new variables to represent each of the 12 positions listed in the position variable. It should be noted that we would not include all 12 position variables (contrasts) in a single regression model because only 11 would be needed (the 12th would be redundant and not provide additional information). However, all 12 were created because the zero-order relations among these 12 and red cards and skin tone were conducted as part of the process for determining which should be considered as potential covariates (see answer to covariate question below for more details). For each of these new variables, players were assigned a 0 if the position variable under consideration was not their position or a 1 if it was their position (if they were missing data on the original position variable, then their data was coded as missing on all of the new 12 position variables). When referred to below in the covariates section, these new variables will be called: position-Attacking Midfielder, position-Center Back, position-Center Forward, position-Center Midfielder, position-Defensive Midfielder, position-Goalkeeper, position-Left Fullback, position-Left

Midfielder, position-Left Winger, position-Right Fullback, position-Right Midfielder, and position-Right Winger. / The leagueCountry variable was used to create 4 new variables to represent each of the 4 countries listed in this variable (similar to above, no more than 3 of these would be included as covariates because the 4th would not provide information beyond three). As with above, players were assigned a 0 if the leagueCountry variable under consideration was not their club's country or a 1 if it was. When referred to below in the covariates section, these new variables will be called: country-England, country-France, country-Germany, and country-Spain. / When testing research question 1, the dataset was restructured such that each case was a single player as opposed to the original dataset where each case was a player-referee dyad. Thus, the number of cases was reduced from 146,028 player-referee dyads to 2,053 players. As part of this data restructuring, the original variables of games, victories, ties, defeats, goals, yellowCards, yellowReds, and redCards were 'transformed' in the sense that their values were summed across referee-player dyads for each player. Below, these variables will be referred to as total games, total victories, total ties, total defeats, total goals, total yellowCards, total yellowReds, and total redCards / When testing research question 2a and 2b, the meanIAT, meanExp, and average skin tone variables were all transformed by mean centering them. This was done by calculating a mean for each variable and then subtracting that mean from each score. An interaction term was then created by multiplying the centered meanIAT variable and the centered average skin tone variable. Another interaction term was created by multiplying the centered meanExp variable and the centered average skin tone variable. / When testing research questions 2a and 2b, the redCards variable was transformed for each player-referee dyad to create a dichotomous red card variable. Thus, if no red card was given, a value of zero was assigned as in the original redCards variable, and if one or more red cards were given, a value of one was assigned to indicate that the specific referee had given at least 1 red card to the specific player. This was done as only 0.017120% (25 of 146,028) of the player-referee dyads with red card data awarded more than one red card. Were any cases excluded, and why? Any cases with missing data on the variables under consideration in a given analysis were excluded from that analysis. Thus, the final analysis for question 1 consisted of 1,433 of 2,053 (69.8%) players. Those excluded from analysis were missing position data (n = 152), skin tone ratings (n = 253), or both position data and skin tone ratings (n = 215). / The final analysis for questions 2a and 2b consisted of 116,014 of 146,028 (79.5%) player-referee dyads. Those excluded from analysis were missing position data (n = 8,454), skin tone ratings (n = 12,134), meanIAT / meanExp (n = 146), both position data and skin tone ratings (n = 9,263), both skin tone ratings and meanIAT / meanExp (n = 8), both position data and meanIAT / meanExp (n = 7), and missing position data, skin tone ratings, and meanIAT / meanExp (n = 2). What is the name of the statistical technique that you employed? Multiple linear regression with a single continuous outcome variable (total red cards) and multiple predictor variables were used to answer question 1. Multiple binary logistic regression with a single dichotomous outcome variable (dichotomized red cards) and multiple predictor variables were used to answer questions 2a and 2b. Please describe the statistical technique you chose in more detail. Be specific, especially if your choice is not one you consider to be well-known. Multiple linear regression is an extension of simple linear regression that enables the prediction of an outcome, or criterion, variable from two or more predictor variables. Multiple regression analyses enable one to determine the percentage of variance in the outcome variable explained by the model as well as the relative contribution of each of the predictors to the variance explained. / Please see the answer to the following question below for more details on how the statistical technique was applied: 'What theoretical and/or statistical rationale was used for your choice of covariates included in the models?' / Multiple binary logistic regression is an extension of multiple linear regression in which the outcome, or criterion, variable is dichotomous. Multiple binary logistic regression enables one to measure the relation between multiple predictor variables (that can be continuous or categorical) and a dichotomous outcome variable. What are some references for the statistical technique that you chose? Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). Applied multiple regression / correlation analysis for the behavioral sciences (3rd ed.). Mahwah, NJ: Lawrence Erlbaum Associates. Which software did you use? If you used multiple kinds, please indicate what was accomplished with each piece of software (e.g., Data cleaning - R; Model estimation - SAS) IBM SPSS Statistics, Version 22.0.0.0 What distribution did you specify for the outcome variable of red cards? For question 1 we treated the red cards variable as a continuous variable whereas for questions 2a and 2b we treated the variable as a binary variable. What variables were included as covariates (or control variables) when testing research question 1: The relationship between player skin tone and red cards / received? Age, total games, total goals, position-Attacking Midfielder, position-Center Back, position-Center Forward, position-Left Winger, position-Right Midfielder, country-Spain, and country-France. What variables were included as covariates (or control variables) when testing research question 2a: The relationship between referee country implicit skin-tone prejudice and red cards received by dark skin-toned players? Age, total games, total goals, position-Attacking Midfielder, position-Center Back, position-Center Forward, position-Left Winger, position-Right Midfielder, country-Spain, and country-France. What variables were included as covariates (or control variables) / when testing research question 2b: The relationship between referee country explicit skin-tone prejudice and red cards received by dark skin-toned players? Age, total games, total goals, position-Attacking Midfielder, position-Center Back, position-Center Forward, position-Left Winger, position-Right Midfielder, country-Spain, and country-France. What theoretical and/or statistical rationale was used for your choice of covariates included in the models? Because we do not have a strong knowledge of the theoretical background in this area, we examined all variables that were included, and seemed reasonable (e.g., club did not seem reasonable because there was such a large number of different clubs), as potential covariates. In the data restructured to have players as the unit of analysis (rather than player-referee dyads), the zero-order correlations were examined between all potential covariates and both total red cards and average skin tone. All variables that demonstrated a significant zero-order correlation with either total red cards, average skin tone, or both were included in the preliminary, full regression analysis for research question 1. For the preliminary, full regression analysis, there were several variables whose coefficients were not significant.

The variable with the largest p value, position-Center Midfielder, was removed and it was verified that removing this variable did not result in a significant decrement in model fit. The resulting model was also examined to see whether or not the parameter estimates for average skin tone were impacted by removal of the position-Center Midfielder variable. This was done because although the position-Center Midfielder variable did not provide significant, unique prediction of red cards, it was still possible that its inclusion influenced the relation between skin tone and red cards. There was little change in the parameter estimates for the relation between skin tone and red cards upon removal of the position-Center Midfielder variable. Thus, the position-Center Midfielder variable was not included in the final model as a covariate because it seemed more parsimonious to exclude this variable because it did not contribute significantly to the model and its removal did not have a meaningful impact on the relation of interest (the relation between skin tone and red cards). This process was carried out until all of the remaining variables had coefficients that were statistically significant and whose removal resulted in a significant decrement in model fit. These same variables were included as covariates for research questions 2a and 2b. What unit is your effect size in? For research question #1, a standardized regression coefficient is provided: Beta. For research questions 2a and 2b, the exponentiation of the unstandardized regression coefficient, or the odds ratio, is provided: Exp(B).

Confid_Team10 How confident are you that the described approach below is suitable for analyzing the research questions?

	Unconfident	Rather unconfident	Somewhat unconfident	Neither confident nor unconfident	Somewhat confident	Rather confident	Confider
RQ1) Skin color and red cards	○	○	○	○	○	○	○
RQ2a) Implicit cultural preferences	○	○	○	○	○	○	○
RQ2b) Explicit cultural preferences	○	○	○	○	○	○	○

Team10 Please provide feedback to the analytical approach described below.

Approach10 Analytical Approach Team10   What transformations (if any) were applied to the variables. Please be specific. We chose NOT to transform the redCards measure into a proportion, because this would equate a player who received 0 cards in one game with a referee and a player who received 0 cards in 20 games. Instead, we controlled for games as a covariate in all analyses. Height and Weight were used to calculate BMI, which we thought provided a better measure of physical stature than either measure alone. Birthday was used to calculate player age (in days) as of 1/1/2013 to evaluate whether this was confounded with skin tone. An aggregate goals/game variable was calculated at the player level to examine how "star status" might influence the results. Another set of aggregate variables were calculated at the club level to examine whether mean skin tone of one's club, mean goals/game, and mean losing percentage had any additional influence. Supplementary analyses were performed with a transformed skin tone variable in which players with mean ratings < 3 were categorized as Non-African-appearing and players with mean ratings >3 were categorized as African-appearing. This was based on a visual inspection of the players sorted into the different skin tone categories and was intended to provide a more focused test of the IAT and explicit prejudice hypotheses which were built specifically on attitudes toward blacks. Were any cases excluded, and why? In the supplementary analyses, players with mean skin tone ratings of 3 were excluded, because that group included both players who were African-appearing and Non-African-appearing, which would make focused tests of prejudice toward blacks vs. whites more difficult. All other exclusions were only the result of missing data on some of the required measures. What is the name of the statistical technique that you employed? Multilevel regression analyses. Please describe the statistical technique you chose in more detail. Be specific, especially if your choice is not one you consider to be well-known. Because of the extensive nesting and multiple sources of dependency within the data, we used techniques that could account for this dependency and examine the interactions between variables at different levels of nesting. Before conducting the primary analyses, we tested which types of nesting were explaining significant variance in the outcome of interest before adopting that nesting structure for all subsequent analyses. Results showed that referee country and referee were factors that contributed significant random variance to the model. Therefore, a three-level structure was adopted in which referee was nested within referee country and intercepts were allowed to vary randomly across these levels. Further analyses showed that the skin tone slope effect did not contribute significant random variance, and thus these effects were always treated as fixed in the model. What are some references for the statistical technique that you chose? Raudenbush, S. W., & Bryk, A. S. (2002). Hierarchical linear models: Applications and data analysis methods (Vol. 1). Sage. Which software did you use? If you used multiple kinds, please indicate what was accomplished with each piece of software (e.g., Data cleaning - R; Model estimation - SAS) Data organization - Excel; Model estimation - R What distribution did you specify for the outcome variable of red cards? Because we were partialing out games, which created a wide range of adjusted values, we retained a standard (gaussian) distribution. We discussed transforming the data into proportions and then transforming the results with an arcsin function, but we decided against this because it equated people who received 0 red cards in 1 game with a referee and people who received 0 red cards in 20. We also discussed recategorizing the red cards variable into a binomial and performing a logistic regression, but we did not choose this approach because of the same problem and it potentially eliminated meaningful variance among those few cases in which players did receive multiple red cards across a varying number of games. Finally, we decided against a poisson regression because the maximum count for red cards was two.   However, I am persuadable on other approaches. What variables were included as covariates (or control variables) when testing research question 1: The relationship between player skin tone and red cards / received? Games, position and club in the player-level skin tone analysis; games, position, and player-level skin tone in the club-level aggregate skin tone analysis What variables were included as covariates (or control variables) when testing research question 2a: The relationship between referee country implicit skin-tone prejudice and red cards received by dark skin-toned players? Games, position and club in the player-level skin tone analysis; games, position, and player-level skin tone in the club-level aggregate skin tone analysis What variables were included as covariates (or control variables) / when testing research question 2b: The relationship between referee country explicit skin-tone prejudice and red cards received by dark skin-toned players? Games, position and club in the player-level skin tone analysis; games, position, and player-level skin tone in the club-level aggregate skin tone analysis What theoretical and/or statistical rationale was used for your choice of covariates included in the models? Preliminary tests showed that player skin tone was unrelated to bmi, age, aggregate goals scored/game, and aggregate losing percentage. Player skin tone was unequally distributed across player position and across club. Furthermore, position and club alone predicted frequency of red cards. Therefore, it was necessary to control for both of these potential confounds in all analyses of skin tone effects. What unit is your effect size in? Beta

Confid_Team9 How confident are you that the described approach below is suitable for analyzing the research questions?

	Unconfident	Rather unconfident	Somewhat unconfident	Neither confident nor unconfident	Somewhat confident	Rather confident	Confiden
RQ1) Skin color and red cards	❍	❍	❍	❍	❍	❍	❍
RQ2a) Implicit cultural preferences	❍	❍	❍	❍	❍	❍	❍
RQ2b) Explicit cultural preferences	❍	❍	❍	❍	❍	❍	❍

Team9 Please provide feedback to the analytical approach described below.

Approach9 Analytical Approach Team9 What transformations (if any) were applied to the variables. Please be specific. We disaggregated the data set in a way that each row is one game, which has either a red card in a specific player-referee-dyad (RC=1) or no red card (RC=0). So our DV now is binary. / / We recoded skintone that the lowest category is 0 (now ranges from 0 to 4), to make the 0 interpretable. Were any cases excluded, and why? We excluded all cases with no skin tone rating. We assume that these cases are MCAR (there was no further information provided why these players have no rating). We acknowledge that this is some loss of information (e.g., the estimation of our random effects could be improved when these cases are included), but in our approach using the lme4 package in R, we are not aware of a way to include these cases. What is the name of the statistical technique that you employed? Generalized linear mixed effects models (GLMM), with a logit link function (binary outcome) Please describe the statistical technique you chose in more detail. Be specific, especially if your choice is not one you consider to be well-known. We modeled the probability that a player gets a red card in a game. / Players and referees were entered as random effects. / Position and leagueCountry were entered as fixed effects. / Skintone was entered as fixed effect. / / For RQ2, we entered the z-transformed meanIAT as main effect and its interaction with skintone. The same for meanExp. What are some references for the statistical technique that you chose? Gelman & Hill, 2007 Which software did you use? If you used multiple kinds, please indicate what was accomplished with each piece of software (e.g., Data cleaning - R; Model estimation - SAS) R, in particular package lme4. For data handling and plotting: dplyr, ggplot2. For regression diagnostics: arm; for exploration: party package What distribution did you specify for the outcome variable of red cards? Binomial (logit link) What variables were included as covariates (or control variables) when testing research question 1: The relationship between player skin tone and red cards / received? Position + leagueCountry / We furthermore explored an interaction between leagueCountry and skintone. What variables were included as covariates (or control variables) when testing research question 2a: The relationship between referee country implicit skin-tone prejudice and red cards received by dark skin-toned players? Position + leagueCountry What variables were included as covariates (or control variables) / when testing research question 2b: The relationship between referee country explicit skin-tone prejudice and red cards received by dark skin-toned players? Position + leagueCountry What theoretical and/or statistical rationale was used for your choice of covariates included in the models? We included position and leagueCountry as covariates for a baseline model, as these variables both theoretically and empirically are related to the probability of receiving a red card. / / Furthermore, we split the data set into a training and a validation set. The split was conducted in a way that both sets had the same structure of positions, leagueCountry, and skin tone. Each player only appeared in one of both sets. There were no significant differences in these variables (p > .22) / / In the training set, we explored the impact of several other potential covariates using binary classification trees (the party package in R), and checked whether the findings replicated in the test set. Inspired by this explorative analysis, we explored whether the skintone effect differes between countries. / / What unit is your effect size in? Odds ratio

Confid_Team28 How confident are you that the described approach below is suitable for analyzing the research questions?

	Unconfident	Rather unconfident	Somewhat unconfident	Neither confident nor unconfident	Somewhat confident	Rather confident	Confiden
RQ1) Skin color and red cards	❍	❍	❍	❍	❍	❍	❍
RQ2a) Implicit cultural preferences	❍	❍	❍	❍	❍	❍	❍
RQ2b) Explicit cultural preferences	❍	❍	❍	❍	❍	❍	❍

Team28 Please provide feedback to the analytical approach described below.

Approach28 Analytical Approach Team28   What transformations (if any) were applied to the variables. Please be specific. We reduced the number of levels of the "position" variable: We collapsed the positions "Left Fullback", "Right Fullback", and "Center Back" to "Back", "Left Midfielder", "Right Midfielder", "Center Midfielder", "Attacking Midfielder", and "Defensive Midfielder" to "Middle" and "Left Winger", "Right Winger", and "Center Forward" to "Front". Furthermore, we used the mean of the two raters' skin tone ratings as predictor. Were any cases excluded, and why? Skin tone ratings and racial bias information were the predictor variables of interest and we wanted to include players' positions as covariate in our model. We, thus, excluded all cases missing skin tone ratings, information on racial bias in referees' home country ("meanIAT"/"meanExp"), or information on players' positions. Finally, we excluded all dyads with referees who in total encountered only one player because otherwise fitting our mixed effects model with random intercept term for referees would not have been possible. What is the name of the statistical technique that you employed? Generalized linear mixed effects modeling Please describe the statistical technique you chose in more detail. Be specific, especially if your choice is not one you consider to be well-known. We employed a logistic mixed effects model with crossed random effects for referees and players. Specifying crossed random effects simultaneously allows generalization of the results beyond the referees and players from the sample at hand. What are some references for the statistical technique that you chose? Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. Journal of Memory and Language, 59, 390'412. doi:10.1016/j.jml.2007.12.005; Jaeger, T. F. (2008). Categorical Data Analysis: Away from ANOVAs (transformation or not) and towards

Logit Mixed Models. Journal of Memory and Language, 59, 434'446. doi:10.1016/j.jml.2007.11.007; Pinheiro, J. C., & Bates, D. M. (2000). Mixed-effects models in S and S-PLUS (p. 535). New York, NY: Springer.; West, B. T., Welch, K. B., & Galecki, A. T. (2007). Linear mixed models: A Practical Guide Using Statistical Software (p. 339). Boca Raton, FL: Chapman & Hall. Which software did you use? If you used multiple kinds, please indicate what was accomplished with each piece of software (e.g., Data cleaning - R; Model estimation - SAS) R (lme4 package) What distribution did you specify for the outcome variable of red cards? Binomial distribution with logit link What variables were included as covariates (or control variables) when testing research question 1: The relationship between player skin tone and red cards / received? Player position and league country What variables were included as covariates (or control variables) when testing research question 2a: The relationship between referee country implicit skin-tone prejudice and red cards received by dark skin-toned players? Player position and league country What variables were included as covariates (or control variables) / when testing research question 2b: The relationship between referee country explicit skin-tone prejudice and red cards received by dark skin-toned players? Player position and league country What theoretical and/or statistical rationale was used for your choice of covariates included in the models? We assumed that the likelihood of receiving a red card would vary with a player's position. Defending player should be especially prone to foul opponents severely while trying to prevent them from scoring. We, furthermore, found that skin tone and not evenly distributed across player positions and league countries. What unit is your effect size in? Odds ratio

Confid_Team13 How confident are you that the described approach below is suitable for analyzing the research questions?

	Unconfident	Rather unconfident	Somewhat unconfident	Neither confident nor unconfident	Somewhat confident	Rather confident	Confiden
RQ1) Skin color and red cards	❍	❍	❍	❍	❍	❍	❍
RQ2a) Implicit cultural preferences	❍	❍	❍	❍	❍	❍	❍
RQ2b) Explicit cultural preferences	❍	❍	❍	❍	❍	❍	❍

Team13 Please provide feedback to the analytical approach described below.

Approach13 Analytical Approach Team13 What transformations (if any) were applied to the variables. Please be specific. The two ratings scores were averaged after checking the inter-rater reliability. We dichotomized the skin rating variable to create a new variable that treated average ratings < 3 as "light skin" and average ratings > 3 as "dark skin". We deliberately left out players from this dichotomy who were rated by both raters to be "neutral". This variable was used in all analyses. For Research Question 2, we aggregated the data for number of games and red cards, by skin color and ref's country. We then calculated the observed probability of getting a red card for dark and light skin players (#red cards/ # games) for each country, and use these variables in the analysis for Q2. Were any cases excluded, and why? No cases were excluded in question 1. For question 2, one country was removed from the aggregated data (#133) for being a major outlier in the probability score for dark skin players. Missing cases were excluded pair-wise from analyses. What is the name of the statistical technique that you employed? glm with poisson distribution Please describe the statistical technique you chose in more detail. Be specific, especially if your choice is not one you consider to be well-known. We decided to use a poisson distribution to model the number of red cards, relative to the number of games played (log(games) was used as the offset variable). This was chosen because of the count nature of this data, and the fact counts occurred in small frequencies across a variable number of games. What are some references for the statistical technique that you chose? http://www.ats.ucla.edu/stat/r/dae/poissonreg.htm Which software did you use? If you used multiple kinds, please indicate what was accomplished with each piece of software (e.g., Data cleaning - R; Model estimation - SAS) R What distribution did you specify for the outcome variable of red cards? poisson What variables were included as covariates (or control variables) when testing research question 1: The relationship between player skin tone and red cards / received? We included position in Q1. We also included the log(games) as the offset variable. What variables were included as covariates (or control variables) when testing research question 2a: The relationship between referee country implicit skin-tone prejudice and red cards received by dark skin-toned players? The probability of receiving a red card among light skin players What variables were included as covariates (or control variables) / when testing research question 2b: The relationship between referee country explicit skin-tone prejudice and red cards received by dark skin-toned players? The probability of receiving a red card among light skin players What theoretical and/or statistical rationale was used for your choice of covariates included in the models? For question 1, we felt position could influence the overall number of red cards given to players, independent of race. For question 2, we wanted to look at the red cards given to dark skin players, controlling for the probability of getting a red card among light skin players. What unit is your effect size in? Log Count of red cards relative to games (Q1: beta estimate for light vs dark skin; Q2: beta estimate for meanIAT/Exp)

Confid_Team14 How confident are you that the described approach below is suitable for analyzing the research questions?

	Unconfident	Rather unconfident	Somewhat unconfident	Neither confident nor unconfident	Somewhat confident	Rather confident	Confider
RQ1) Skin color and red cards	○	○	○	○	○	○	○
RQ2a) Implicit cultural preferences	○	○	○	○	○	○	○
RQ2b) Explicit cultural preferences	○	○	○	○	○	○	○

Team14 Please provide feedback to the analytical approach described below.

Approach14 Analytical Approach Team14   What transformations (if any) were applied to the variables. Please be specific. a) We created a new variable indicating the total number of red and yellow-red cards within each player-referee dyad by summing the number of red and the number of yellow-red cards within each player-referee dyad. / / b) We created a new variable (which we will use as our outcome variable) indicating the total number of red and yellow-red cards per game within each player-referee dyad by dividing the total number of red and yellow-red cards with the number of games within the player-referee dyad. / / c) We created a binary variable, indicating one for player-referee dyads where at least one red or yellow-red card was given, and zero otherwise. / / d) We created a variable for skin tone, which is the average value of rater 1 and rater 2. We will use this as our main explanatory variable. / / e) We created a binary variable taking the value one for players whom at least one of the raters have indicated the skin tone to be equal to 4 or more, and zero if both raters indicated the player's skin tone to be below 4. Missing values are treated as missing. / / f) We created a binary variable taking the value one for players whom at least one of the raters have indicated the skin tone to be equal to 5, and zero if both raters indicated the player's skin tone to be below 5. Missing values are treated as missing. / / g) We created an interaction term by multiplying the variable for skin tone with mean exp and mean IAT, respectively. We also created an interaction term between the binary variables for skin tone with mean exp and mean IAT, respectively. Were any cases excluded, and why? a) Player-referee dyads where skin tone has not been rated are excluded from the regression analysis. This is how Stata, the statistical software we use, deals with missing values in regression analyses. / / b) Player-referee dyads where mean exp is missing are excluded from the regression analyses addressing the second research question using exp as measure of discrimination in the referee's home country. This is how Stata, the statistical software we use, deals with missing values in regression analyses. / / c) Player-referee dyads where mean IAT is missing are excluded from the regression analyses addressing the second research question using IAT as measure of discrimination in the referee's home country. This is how Stata, the statistical software we use, deals with missing values in regression analyses. / / d) We also conducted sensitivity analyses excluding player-referee dyads where the referee's mean exp has been determined based on a sample (nexp) smaller than either 100 or 1000, respectively. We also did this for mean IAT. What is the name of the statistical technique that you employed? WLS (weighted least squares) estimation Please describe the statistical technique you chose in more detail. Be specific, especially if your choice is not one you consider to be well-known. WLS (weighted least squares) estimation is a version of OLS (ordinary least squares) estimation. OLS is a method for estimating the relationship between two variables assuming that the true relationship between these two variables is linear. The estimated relationship between the two variables is obtained by minimizing the sum of squared vertical distances between the observed values in the sample, and the predicted values of the linear approximation. Since the number of games varies across player-referee dyads, we weight the observations by the number of games per player-referee dyad. This means that we use WLS. / / To answer research question 1 we regress the total number of red and yellow-red cards per game on a continuous variable indicating the player's skin tone (the average value of rater 1 and rater 2) on a 1-5 scale. The effect we report is the coefficient on this variable. As control variables we include the player's birth year, height, weight, club and position. We also include referee fixed effects, implying that we estimate how the likelihood of receiving a red or yellow-red card varies with skin tone within referees. We cluster the standard errors on the player level in order to account for that we have multiple observations per player. / / To answer research question 2 a (b) we regress the total number of red and yellow-red cards per game on the continuous variable indicating the player's skin tone, and an interaction term between this variable and mean exp (mean iat). Since we are interested in whether the bias is larger when the referee comes from a country with a higher discriminatory index, the effect we report is the coefficient on the interaction term. As control variables we include the player's birth year, height, weight, club and position. We also include referee fixed effects, implying that we estimate how the likelihood of receiving a red or yellow-red card varies with skin tone within referees. We cluster the standard errors on the player level in order to account for that we have multiple observations per player. What are some references for the statistical technique that you chose? (1) 'Introductory Econometrics. A Modern Approach', 4th edition, 2009. Jeffrey M. Wooldridge, (2) 'Econometric Analysis', 7th edition, 2012. William H. Greene. Which software did you use? If you used multiple kinds, please indicate what was accomplished with each piece of software (e.g., Data cleaning - R; Model estimation - SAS) STATA What distribution did you specify for the outcome variable of red cards? Our chosen statistical technique relies on the assumption that the outcome variable is normally distributed, with the mean equal to the sample mean and the variance equal to the sample variance. What variables were included as covariates (or control variables) when testing research question 1: The relationship between player skin tone and red cards / received? a. Height / b. Weight / c. Player birth year / d. Position / e. Club / f. Referee fixed effects What variables were included as covariates (or control variables) when testing research question 2a: The relationship between referee country implicit skin-tone prejudice and red cards received by dark skin-toned players? a. Height / b. Weight / c. Player birth year / d. Position / e. Club / f. Referee fixed effects What variables were included as covariates (or control variables) / when testing research question 2b: The relationship between referee country explicit skin-tone prejudice and red cards received by dark skin-toned players? a. Height / b. Weight / c. Player birth year / d. Position / e. Club / f. Referee fixed effects What theoretical and/or statistical rationale was used for your choice of covariates included in the models? We control for height, weight, player birth year, position and club since these variables may be correlated with skin tone. If we do not control for these variables, we therefore risk getting a biased estimate of the relationship between skin tone and the likelihood of being awarded a red or a yellow-red card. Assume for instance that players of a certain skin tone on average are taller than other players, and that taller players are on average awarded more red and yellow-red cards. Then, if we do not control for height, it may appear as if players of that particular skin tone are discriminated against although it is rather their height, and not their skin tone, that increases their likelihood of receiving a red or yellow-red card. / / We also include referee fixed effects. This implies that we measure the effect of skin tone on the likelihood of receiving a red or yellow-red card within referees. That is, for a given referee, we analyze whether a player's likelihood of receiving a red or yellow-red card varies with his skin tone. / / It is important to point out that the data set is insufficient to identify any causal effect of skin color on the probability of getting a red or yellow-red card. We can only estimate if the likelihood of getting a red or yellow-red card varies with skin color after controlling for as many variables as possible; but this variation may be correlated with unobserved factors affecting the likelihood of getting such a card. The most important unobserved variable here is 'player style'; i.e. to what extent the player style is correlated with skin color (skin color could for instance be correlated with how physical a player plays, affecting the likelihood of getting yellow and red cards). The data set is also structured in a 'non-optimal' way with the player/referee dyads. It would have been better if each game for each player was one observation. What unit is your effect size in? Number of red and yellow-red cards per game. This can be interpreted as the likelihood of receiving a red or yellow-red card per game since the number of cards per game is between 0 and 1.

Confid_Team15 How confident are you that the described approach below is suitable for analyzing the research questions?

	Unconfident	Rather unconfident	Somewhat unconfident	Neither confident nor unconfident	Somewhat confident	Rather confident	Confiden
RQ1) Skin color and red cards	○	○	○	○	○	○	○
RQ2a) Implicit cultural preferences	○	○	○	○	○	○	○
RQ2b) Explicit cultural preferences	○	○	○	○	○	○	○

Team15 Please provide feedback to the analytical approach described below.

Approach15 Analytical Approach Team15   What transformations (if any) were applied to the variables. Please be specific. IAT: dichotomized (median split); Explicit: Dichotomized (median split). Were any cases excluded, and why? All players without a perfect agreement between rater 1 and rater 2 skin tone assessments were discarded. This was done because 1) raters rated a picture not a real player 2) the picture could have been taken in different times across the year (e.g., just before the league, in winter, or after a holiday at the beach for some players) and because the percentage of perfect agreements was less than 80%. We thought it was important to analyze the data using a highly reliable measure of our dependent variable. What is the name of the statistical technique that you employed? Hierarchical log-linear modeling. Please describe the statistical technique you chose in more detail. Be specific, especially if your choice is not one you consider to be well-known. Useful for categorical data.   / The saturated model for a three-way table is ln(Fijk)=Î»+ Î»A+ Î»B+ Î»C+ Î»AB+ Î»AC+ Î»BC+ Î»ABC / Where Fijk=expected frequency in cell ijk and Î»=relative weight of each variable / We compared nested models using the difference in the likelihood ratio chi-square statistics of the models being compared. / What are some references for the statistical technique that you chose? Agresti, A. (2013). Categorical Data Analysis, 3rd ed. New York: Wiley. Which software did you use? If you used multiple kinds, please indicate what was accomplished with each piece of software (e.g., Data cleaning - R; Model estimation - SAS) Data preparation: SPSS, Model fit: BMDP and R. What distribution did you specify for the outcome variable of red cards? Poisson What variables were included as covariates (or control variables) when testing research question 1: The relationship between player skin tone and red cards / received? League What variables were included as covariates (or control variables) when testing research question 2a:

The relationship between referee country implicit skin-tone prejudice and red cards received by dark skin-toned players? NONE What variables were included as covariates (or control variables) / when testing research question 2b: The relationship between referee country explicit skin-tone prejudice and red cards received by dark skin-toned players? NONE What theoretical and/or statistical rationale was used for your choice of covariates included in the models? We controlled for 'League' in testing question 1 because there is an association between league and skintone and between league and Redcards such that a bias may be explained by the different distribution of skintone within each league and by the different distribution of redcards within each league. E.g. very light-tone players (skintone=1) are more likely to be found in the english league, and less red cards are given in the english league. What unit is your effect size in? don't know yet which effect size can be associated to a likelihood ratio. chi sq/df is surely an indication, but we need more time for this.

Confid_Team17 How confident are you that the described approach below is suitable for analyzing the research questions?

	Unconfident	Rather unconfident	Somewhat unconfident	Neither confident nor unconfident	Somewhat confident	Rather confident	Confider
RQ1) Skin color and red cards	○	○	○	○	○	○	○
RQ2a) Implicit cultural preferences	○	○	○	○	○	○	○
RQ2b) Explicit cultural preferences	○	○	○	○	○	○	○

Team17 Please provide feedback to the analytical approach described below.

Approach17 Analytical Approach Team17   What transformations (if any) were applied to the variables. Please be specific. We used a probit regression (on the probability of getting a red card) with a number of predictors. The mean IAT score and the mean explicit bias score were z-transformed before they were used as a predictor. The skin tone scores from the two raters were first z-transformed and then averaged, so that these ratings are represented by only a single predictor. Were any cases excluded, and why? We excluded cases for which the IAT score was unknown. We did *not* exclude cases for which skin color was unknown. These data are still useful as they provide information on referee strictness, for instance. What is the name of the statistical technique that you employed? Bayesian probit regression Please describe the statistical technique you chose in more detail. Be specific, especially if your choice is not one you consider to be well-known. The analysis is a probit regression on the probability of getting a red card, featuring the following predictors: / / 1. Referee Strictness; / 2. Player Aggression; / 3. IAT score for the referee's country (Ibias); / 4. Explicit bias score for the referee's country (Ebias); / 5. Player Skin Tone (Skin). This is quantified by the average of the z-scores for the two raters. / 6. Interaction between IAT score and Player skin tone (IbiasSkin). / 7. Interaction between explicit bias score and player skin tone (EbiasSkin). / / We decided to leave out league country as a predictor, because any differences due to this predictor will be accounted for by referee strictness and player aggression. Other possible predictors were deemed superfluous or uninformative and were not added to the regression equation. For instance, player position is not all that informative after player aggression has already been included in the model. Predictors (3) and (4) are included to keep the interaction terms interpretable. / / The first research question, "Are soccer referees more likely to give red cards to dark skin toned players than light skin toned players?", can be quantified by the importance of predictor 5, player skin tone. / / The second research question, "Are soccer referees from countries high in skin-tone prejudice more likely to award red cards to dark skin toned players?", can be quantified by the importance of predictors 6 and 7 (interaction between implicit and explicit bias and player skin tone). / / To quantify the importance of predictors 5, 6, and 7, we compare the full regression model (that includes all of the predictors) to three simplified regression models: (1) model A has all predictors except 5; (2) model B has all predictors except 6; (3) model C has all predictors except 7. Whenever the full model outperforms its simplified versions, this constitutes evidence for the inclusion of the associated predictor. / / To fit the models and quantify this evidence we use Bayesian inference. We first assign each regression coefficient an independent standard Normal distribution. This is not important for the nuisance predictors 1-4, but it is important for the predictors of interest. We still need to conduct a sensitivity analysis and consider ways to motivate the choice for a particular width. / / Next we fit the full regression model to the data (using the JAGS code provided later in this form). Parameter estimation results can be gauged by plotting the posterior distributions for the relevant regression coefficients. Each posterior distribution can be summarized by its mean and a 95% credible interval. For hypothesis testing, we compute three Bayes factors based on the comparison between the full model against each of three simplified models that omit one of the three key predictors. / / Because the simple models are nested in the full model, we can compute the Bayes factor using the Savage-Dickey density ratio; that is, we compute the ratio between the prior and the posterior ordinate at the value under test (i.e., beta = 0). This also allows an immediate visual representation of the test outcomes. / What are some references for the statistical technique that you chose? Two books: Gelman & Hill (2007); Ntzoufras (2009) Which software did you use? If you used multiple kinds, please indicate what was accomplished with each piece of software (e.g., Data cleaning - R; Model estimation - SAS) JAGS to fit the model, R to do everything else. What distribution did you specify for the outcome variable of red cards? We modeled the probit of the probability of getting a red card. What variables were included as covariates (or control variables) when testing research question 1: The relationship between player skin tone and red cards / received? We used a probit regression with 7 predictors: / 1. Referee Strictness; / 2. Player Aggression; / 3. IAT score for the referee's country (Ibias); / 4. Explicit bias score for the referee's country (Ebias); / 5. Player Skin Tone (Skin). This is quantified by the average of the z-scores for the two raters. / 6. Interaction between IAT score and Player skin tone (IbiasSkin). / 7. Interaction between explicit bias score and player skin tone (EbiasSkin). / What variables were included as covariates (or control variables) when testing research question 2a: The relationship between referee country implicit skin-tone prejudice and red cards received by dark skin-toned players? We used a probit regression with 7 predictors: / 1. Referee Strictness; / 2. Player Aggression; / 3. IAT score for the referee's country (Ibias); / 4. Explicit bias score for the referee's country (Ebias); / 5. Player Skin Tone (Skin). This is quantified by the average of the z-scores for the two raters. / 6. Interaction between IAT score and Player skin tone (IbiasSkin). / 7. Interaction between explicit bias score and player skin tone (EbiasSkin). / What variables were included as covariates (or control variables) / when testing research question 2b: The relationship between referee country explicit skin-tone prejudice and red cards received by dark skin-toned players? We used a probit regression with 7 predictors: / 1. Referee Strictness; / 2. Player Aggression; / 3. IAT score for the referee's country (Ibias); / 4. Explicit bias score for the referee's country (Ebias); / 5. Player Skin Tone (Skin). This is quantified by the average of the z-scores for the two raters. / 6. Interaction between IAT score and Player skin tone (IbiasSkin). / 7. Interaction between explicit bias score and player skin tone (EbiasSkin). / What theoretical and/or statistical rationale was used for your choice of covariates included in the models? Predictors 1 and 2 are key "nuisance" factors; Predictors 3 and 4 need to be in the model to keep the interaction interpretable. We could have included other nuisance predictors in an ad-hoc fashion, but our approach was motivated by the desire for simplicity. Exploratory analyses may suggest that particular other predictors are important, but we believe that many are subsumed under referee strictness and player aggression. Our approach here is purely confirmatory: we did not tinker with the predictors. What unit is your effect size in? To obtain a measure of effect size for probit regression, we took the increase in probability of a red card, for an "average" player/ref combination at a value of 0 for the other predictors.

Confid_Team18 How confident are you that the described approach below is suitable for analyzing the research questions?

	Unconfident	Rather unconfident	Somewhat unconfident	Neither confident nor unconfident	Somewhat confident	Rather confident	Confiden
RQ1) Skin color and red cards	○	○	○	○	○	○	○
RQ2a) Implicit cultural preferences	○	○	○	○	○	○	○
RQ2b) Explicit cultural preferences	○	○	○	○	○	○	○

Team18 Please provide feedback to the analytical approach described below.

Approach18 Analytical Approach Team18   What transformations (if any) were applied to the variables. Please be specific. Skin color ratings were treated as factors. I was considering recoding of skin color levels to a 3 point scale (light, undecided, dark), but in the end I stayed with original ratings. Were any cases excluded, and why? Cases with missing IAT and Exp scores were excluded, model depends on availability of this data. / From the example photos and ratings it was clear that both raters' scale was calibrated in a different way. I wanted to use one score, so the dataset was subset to contain only cases where both raters agree (76% of all cases). What is the name of the statistical technique that you employed? hierarchical Bayes model Please describe the statistical technique you chose in more detail. Be specific, especially if your choice is not one you consider to be well-known. The number of red cards was expressed as number of successes with probability p from given number of games per player (Q1) or per player+referee (Q2). / / This probability p was estimated using logistic function dependent on sum of parameters: overall average, average for given player (his play style), average for given referee (his rigorousness), skin color effect and IAT / Exp scores. / / IAT / Exp scores were sampled from normal distribution according to N and std. error provided in the data. / / The priors for all parameters were dependent on hyperparameters distributed normally with zero mean and vauge precision. / What are some references for the statistical technique that you chose? "Doing Bayesian Data Analysis" by Kruschke; "Bayesian Data Analysis" by Gelman, Carlin, Stern, Rubin Which software did you use? If you used multiple kinds, please indicate what was accomplished with each piece of software (e.g., Data cleaning - R; Model estimation - SAS) Data cleaning, processing - R; model estimation - JAGS What distribution did you specify for the outcome variable of red cards? Binomial distribution with red card count as number of successes out of given number of games. What variables were included as covariates (or control variables) when testing research question 1: The relationship between player skin tone and red cards / received? Player id, his skin tone, number of red cards, number of games played What variables were included as covariates (or control variables) when testing research question 2a: The relationship between referee country implicit skin-tone prejudice and red cards received by dark skin-toned players? Player id, player skin tone, number of red cards, number of games played, referee id, referee country IAT/Exp score parameters What variables were included as covariates (or control variables) / when testing research question 2b: The relationship between referee country explicit skin-tone prejudice and red cards received by dark skin-toned players? Player id, player skin tone, number of red cards, number of games played, referee id, referee country IAT/Exp score parameters What theoretical and/or statistical rationale was used for your choice of covariates included in the models? I didn't use any other covariates because this model estimates effect of each players play style and each referee rigorousness separately. I assumed that any information provided by other possible covariates (age, league, ...) would be already hidden in those parameters. Here we investigate only their interaction given the skin tone. What unit is your effect size in? Odds ratio

Confid_Team19 How confident are you that the described approach below is suitable for analyzing the research questions?

	Unconfident	Rather unconfident	Somewhat unconfident	Neither confident nor unconfident	Somewhat confident	Rather confident	Confide
RQ1) Skin color and red cards	○	○	○	○	○	○	○
RQ2a) Implicit cultural preferences	○	○	○	○	○	○	○
RQ2b) Explicit cultural preferences	○	○	○	○	○	○	○

Team19 Please provide feedback to the analytical approach described below.

Approach19 Analytical Approach Team19   What transformations (if any) were applied to the variables. Please be specific. Positions were grouped as: goalkeeper, centerback, fullback, defensive midfielder, midfielder, attacking midfielder, winger, forward. Means for red card per game were examined, and then the positions are rearranged in order from least-often red carded to most, for a nominal variable. (This will then be used as a control.) / / The follow variables are computed for analysis: / bmi = weight / (height/100)^2 / [Straight Red] strRed = redCards - yellowReds / redpergame = redCards / games / yellowpergame = yellowCards / games / [Straight Red per Game] stredpergame = straightRed / games / explicit = skintone * meanExp / implicit = skintone * meanIAT / / Since I used the automatic linear modeling function in SPSS 21, other transformations may have been automatically applied. Were any cases excluded, and why? Cases were excluded listwise if there's a missing value in one of the relevant variables. What is the name of the statistical technique that you employed? Linear Regression Please describe the statistical technique you chose in more detail. Be specific, especially if your choice is not one you consider to be well-known. Automatic Linear Modeling in SPSS 21, all default settings (e.g. standard model as opposed to robust). What are some references for the statistical technique that you chose?   Which software did you use? If you used multiple kinds, please indicate what was accomplished with each piece of software (e.g., Data cleaning - R; Model estimation - SAS) Data Cleaning: Excel // Analysis: SPSS What distribution did you specify for the outcome variable of red cards? Normal. What variables were included as covariates (or control variables) when testing research question 1: The relationship between player skin tone and red cards / received? Player Position. What variables were included as covariates (or control variables) when testing research question 2a: The relationship between referee country implicit skin-tone prejudice and red cards received by dark skin-toned players? Player Position. What variables were included as covariates (or control variables) / when testing research question 2b: The relationship between referee country explicit skin-tone prejudice and red cards received by dark skin-toned players? Player Position. What theoretical and/or statistical rationale was used for your choice of covariates included in the models? Positions that require more cardable activities would get carded more independent of skin tone. What unit is your effect size in? r (I think? SPSS calls it Importance).

Confid_Team20 How confident are you that the described approach below is suitable for analyzing the research questions?

	Unconfident	Rather unconfident	Somewhat unconfident	Neither confident nor unconfident	Somewhat confident	Rather confident	Confiden
RQ1) Skin color and red cards	❍	❍	❍	❍	❍	❍	❍
RQ2a) Implicit cultural preferences	❍	❍	❍	❍	❍	❍	❍
RQ2b) Explicit cultural preferences	❍	❍	❍	❍	❍	❍	❍

Team20 Please provide feedback to the analytical approach described below.

Approach20 Analytical Approach Team20   What transformations (if any) were applied to the variables. Please be specific. 1. We created unique identifier for players, clubs, and leagues using the variables "playerShort," "club," and "leagueCountry." / / 2. For players with missing data ("NA") in position, we used Wikipedia to find out what position they played. We coded them using 4 categories: goalkeeper, defender, midfielder, and forward/winger. We used these 4 categories because Wikipedia did not provide specific enough information for certain players. We then transformed the original position variable using these 4 categories: goalkeeper, defender (center back, left fullback, and right fullback), midfielder (defensive midfielder, center midfielder, attacking midfielder, left midfielder, and right midfielder), and forward/winger (forward, left winger, and right winger). / / 3. We created the age variable from "birthday." Specifically, we subtracted the last 4 characters of the "birthday" variable from 2013. / / 4. We created the variable "rater" by averaging the ratings of skin tone from the two raters ("rater1" and "rater2"). We used the variable "rater" as our predictor. / / 5. We grand-mean centered all predictors and covariates ("rater," "meanIAT," "meanExp," "height," "weight," "games," "victories," "defeats," "goals," and "age"). To clarify, grand-mean centering means that we computed the mean of a variable and subtracted each value of the variable from the mean. This procedure improves the interpretability of the intercept and the interaction terms. / Were any cases excluded, and why? Players with missing data on skin tone ('rater1' and 'rater2') because skin tone is the main predictor in the current study. What is the name of the statistical technique that you employed? A four-level multilevel negative-binomial model. Please describe the statistical technique you chose in more detail. Be specific, especially if your choice is not one you consider to be well-known. We used a multilevel model with player-referee dyads as level-1, players as level-2, clubs as level-3, and leagues as level-4. This model accounts for the interdependence within players, clubs, and leagues. The likelihood of receiving a red card may differ from players to players, from clubs to clubs, and from leagues to leagues. As a hypothetical example, Arsenal as a club may tend to receive more red cards compared to Manchester United. The multilevel structure helps account for similarities in likelihood to receive red cards within Arsenal and within Manchester United. / / We used a negative-binomial model because the dependent variable (redCards) is a count variable. / What are some references for the statistical technique that you chose? Gelman, A., & Hill, J. (2007). Data analysis using regression and multilevel/hierarchical models. Cambridge University Press. / Greene, William H. "Econometric analysis, 5th." Ed.. Upper Saddle River, NJ(2003). Which software did you use? If you used multiple kinds, please indicate what was accomplished with each piece of software (e.g., Data cleaning - R; Model estimation - SAS) Recoding of the position variable ' SPSS; Everything else ' R What distribution did you specify for the outcome variable of red cards? We used a negative-binomial distribution. What variables were included as covariates (or control variables) when testing research question 1: The relationship between player skin tone and red cards / received? Games, victories, defeats, height, weight, age, positions, goals What variables were included as covariates (or control variables) when testing research question 2a: The relationship between referee country implicit skin-tone prejudice and red cards received by dark skin-toned players? Games, victories, defeats, height, weight, age, positions, goals What variables were included as covariates (or control variables) / when testing research question 2b: The relationship between referee country explicit skin-tone prejudice and red cards received by dark skin-toned players? Games, victories, defeats, height, weight, age, positions, goals What theoretical and/or statistical rationale was used for your choice of covariates included in the models? We controlled for numbers of games because encountering a referee more time increases the likelihood of receiving a red card. / / Players may be less likely to commit a foul (and thus receive a red card) if their team won the game. In contrast, they may play more aggressive defense and commit fouls if their team lost the game. Thus, we controlled for the outcomes of the games (victories and defeats) and number of goals. / / We controlled for height and weight because conceivably, bigger players may have more advantage fighting for position and thus be more likely to engage in bodily contact, which may increase the chance of committing fouls (and thus, receiving red cards). / / We controlled for positions because defensive players (goalkeepers and defenders) may commit more fouls and thus receive more red cards. / / We controlled for age because impulsivity, which may be associated with receiving red cards, tends to decrease with age (Steinberg et al., 2008). / Steinberg, L., Albert, D., Cauffman, E., Banich, M., Graham, S., & Woolard, J. (2008). Age differences in sensation seeking and impulsivity as indexed by behavior and self-report: evidence for a dual systems model. Developmental psychology, 44(6), 1764. / What unit is your effect size in? The log of expected count

Confid_Team21 How confident are you that the described approach below is suitable for analyzing the research questions?

	Unconfident	Rather unconfident	Somewhat unconfident	Neither confident nor unconfident	Somewhat confident	Rather confident	Confiden
RQ1) Skin color and red cards	❍	❍	❍	❍	❍	❍	❍
RQ2a) Implicit cultural preferences	❍	❍	❍	❍	❍	❍	❍
RQ2b) Explicit cultural preferences	❍	❍	❍	❍	❍	❍	❍

Team21 Please provide feedback to the analytical approach described below.

Approach21 Analytical Approach Team21   What transformations (if any) were applied to the variables. Please be specific. We transformed the main dependent measure, the number of red cards, to a frequency score of 'red cards per game per player'. This was done by standardizing the number of games to one, by calculating the number of red cards per game. By construction the new variable ranges from 0 to 1, and can be interpreted as the probability of receiving a red card. Note that we made no distinction in the 'type' of red card (i.e. a direct red or twice yellow where treated equally). / We also standardized (z-values) all continuous predictors before analysis. For skin color, we took the average of the two skin color ratings, given high agreement. This variable ranges from 1 to 5. / Were any cases excluded, and why? We did not exclude any cases from the data analysis. Due to variables with missing values, the N-size did decrease to 424 player-referee dyads. What is the name of the statistical technique that you employed? Tobit regression analysis Please describe the statistical technique you chose in more detail. Be specific, especially if your choice is not one you consider to be well-known. At first sight, methods using count data may seem appropriate to use (i.e. a negative binomial or Poisson regression). However, these methods assume that the number of trials (in this case 'games played') is the same for every observation. Including the number of games as a regressor mitigates this problem, but only to some extent. To solve this issue, we transformed the dependent variable and bounded it between 0 and 1. OLS is then no longer appropriate. We choose to apply a Tobit analysis instead because the different numbers of games per player put an upper bound on how many red cards each player can receive. This analytical technique adequately handles this kind of censored data. / As the number of times that a player encounters the same referee can affect the average frequency of receiving a red card, we ran standard model specifications and model specifications with frequency weights to each player-referee dyad. Frequency weighing entailed that, for example, a player-referee dyad that has 10 encounters is weighted 10 times more than a player-referee dyad that only had one encounter. Because this weighted method blows up the number of observations, we adjusted the standard errors for clusters of repeated observations at different levels. This adjustment generally does not affect the point estimate, merely the significance of the effect. We ended up running 4 different model specifications depending on the clusters (either club or country) and frequency weights (yes vs. no). This did not make much difference for the results. The results we report here are the ones in which we used frequency weights and clustered for country. / / Because the level of analysis for this question is player, we aggregated the data to the player level. Across all four model specifications, we found that the effect of Skin color was statistically significant at a 10% significant level. / We found that a player with a very dark skin color has 0.2% higher likelihood of receiving a red card compared to a player with a very light skin. This seems very small, but considering that the average number of red cards is about 0.8%, this effect may be considered as quite large. It implies that for every 4 red cards a player with skincolor 1 receives, a player with skin color 5 will receive an additional red card (to a total of 5.)   / / To test Hypothesis 2, we tried to reproduce the results for the aggregated data with the disaggregated data, either clustered on player or referee observations. Now, the main effect of skin color on frequency of cards is even larger (b = 0.0278, t = 2.59, p < .01). We subsequently included the two prejudice variables in specifications 5 and 6 and the first interaction term (IAT x skin color) in specification 7. This interaction-term did not yield a significant effect. This result suggests that the relationship between skin color and the average frequency of red cards does not depend on this implicit prejudice measure. The second interaction term (Explicit x Skin color), which was entered in specification 8, was also insignificant. This result suggests that the relationship between skin color and the average frequency of red cards does not depend on this explicit prejudice measure.   / / / What are some references for the statistical technique that you chose? William H. Greene, Econometric Analysis, 6th ed., 2008 New Jersey: Prentice or any other good Econometrics textbook. Which software did you use? If you used multiple kinds, please indicate what was accomplished with each piece of software (e.g., Data cleaning - R; Model estimation - SAS) STATA What distribution did you specify for the outcome variable of red cards? (see question 1) We standardized the number of games to one, by calculating the number of red cards per game. This can be interpreted as the probability of receiving a red card. The Tobit analysis assumes a censored normal distribution. What variables were included as covariates (or control variables) when testing research question 1: The relationship between player skin tone and red cards / received? Number of games played / Age / Height / Dummie League: France / Dummie League: Germany / Dummie League: Spain / Dummie position: defender / Dummie position: midfielder / Dummie position: attacker / What variables were included as covariates (or control variables) when testing research question 2a: The relationship between referee country implicit skin-tone prejudice and red cards received by dark skin-toned players? Number of games played / Age / Height / Dummie League: France / Dummie League: Germany / Dummie League: Spain / Dummie position: defender / Dummie position: midfielder / Dummie position: attacker / What variables were included as covariates (or control variables) / when testing research question 2b: The relationship between referee country explicit skin-tone prejudice and red cards received by dark skin-toned players? Number of game played / Age / Height / Dummie League: France / Dummie League: Germany / Dummie League: Spain / Dummie position: defender / Dummie position: midfielder / Dummie position: attacker / What theoretical and/or statistical rationale was used for your choice of covariates included in the models? We included all variables that correlated significantly with both skin color and red cards. This is because these variables may form an possible alternative explanation for the effect of skin color on red cards. This is what we wanted to rule out. Note that we choose for height over weight (that correlated with each other, obviously) because weight did not predict red cards when controlled for height. What unit is your effect size in? We base our effect size estimate on the regression weight b and the skin color rating 1-5.

Confid_Team22 How confident are you that the described approach below is suitable for analyzing the research questions?

	Unconfident	Rather unconfident	Somewhat unconfident	Neither confident nor unconfident	Somewhat confident	Rather confident	Confiden
RQ1) Skin color and red cards	○	○	○	○	○	○	○
RQ2a) Implicit cultural preferences	○	○	○	○	○	○	○
RQ2b) Explicit cultural preferences	○	○	○	○	○	○	○

Team22 Please provide feedback to the analytical approach described below.

Approach22 Analytical Approach Team22   What transformations (if any) were applied to the variables. Please be specific. Yes. In the variable birthday, we extracted the year of birth and created a variable: 2013 - year of birth. Thus, we are assuming that the month and day of birth has a negligible effect. The variables redCards, yellowCards and yellowReds were divided by the variable games. So they should be interpreted as average number of red cards per referree-player pair and per game. The variables rater1 and rater2 were transformed into one variable: (rater1 + rater2)/2. This new variable denotes the average rate. Were any cases excluded, and why? No. What is the name of the statistical technique that you employed? The classic OLS, but with dummy variables for each referee and player. Clustered standard errors. Please describe the statistical technique you chose in more detail. Be specific, especially if your choice is not one you consider to be well-known. We are in the context of OLS. The model can be written as, / y_it = X_it beta + epsilon_it, / where i denotes the individual player and t the referee. / / What are some references for the statistical technique that you chose? Cameron and Trivedi (2005). Microeconometrics: Methods and Applications ; Wolfers and Price (2010). Which software did you use? If you used multiple kinds, please indicate what was accomplished with each piece of software (e.g., Data cleaning - R; Model estimation - SAS) R version 3.1.0 . What distribution did you specify for the outcome variable of red cards? This is a linear model, which implies normal distribution for the dependent variable. What variables were included as covariates (or control variables) when testing research question 1: The relationship between player skin tone and red cards / received? Before doing variable selection: averRate (average rater1 and rater 2) + weight + height + position + refCountry + yellowCards_game + leagueCountry + club + refNum (dummy variables) + player (dummy variables) + age + victories + defeats What variables were included as covariates (or control variables) when testing research question 2a: The relationship between referee country implicit skin-tone prejudice and red cards received by dark skin-toned players? Similar (See 1). Just replaced averRate by meanIAT What variables were included as covariates (or control variables) / when testing research question 2b: The relationship between referee country explicit skin-tone prejudice and red cards received by dark skin-toned players? Similar (See 1). Just replaced averRate by meanExp / What theoretical and/or statistical rationale was used for your choice of covariates included in the models? The selection criteria AIC (akaike information criteria) and BIC (bayesian information criteria) are used. We eliminate variables one by one and stop if the varible is statistically significant or the AIC/BIC criteria do not improve. What unit is your effect size in? Do not have an estimate yet.

Confid_Team23 How confident are you that the described approach below is suitable for analyzing the research questions?

	Unconfident	Rather unconfident	Somewhat unconfident	Neither confident nor unconfident	Somewhat confident	Rather confident	Confiden
RQ1) Skin color and red cards	○	○	○	○	○	○	○
RQ2a) Implicit cultural preferences	○	○	○	○	○	○	○
RQ2b) Explicit cultural preferences	○	○	○	○	○	○	○

Team23 Please provide feedback to the analytical approach described below.

Approach23 Analytical Approach Team23   What transformations (if any) were applied to the variables. Please be specific. a variable 'skintone' was constructed by averaging rater1 and rater2. In the final analysis this was considered as a continuous variable and was centred around the mean. We also centred the implicit and explicit bias scores. / / a variable 'allreds' was constructed by adding yellowReds and redCards / / the data was recoded from player-ref dyads per row into single games per row, with the allreds becoming a binary variable (and, obviously, the sum of allreds for each player-ref dyad remaining the same as before recoding) Were any cases excluded, and why? We dropped all rows where there were missing values in any of: / 1) the position of the player / 2) the skin tone of the player / 3) the bias scores of the country from which the referee came / 4) the league in which the player was active /   / This was done on the presumption that such data were missing at random What is the name of the statistical technique that you employed? Mixed model logistic regression - both frequentist and Bayesian. Our analysis was performed in R using the lme4 package (frequentist) developed by Bates et al.; and the MCMCglmm package (Bayesian) developed by Jarrod Hadfield (references below) Please describe the statistical technique you chose in more detail. Be specific, especially if your choice is not one you consider to be well-known. logistic regression predicts the outcome on a binary variable - in this case whether a game resulted in a red card for that player. We suppose that the probability p_{ij} of any individual i obtaining a red card in a game j depends upon a number of predictor variables. These predictor variables can either be 'fixed effects' x_{ij} (which are typically of most interest and would stay the same if the experiment were to be hypothetically repeated e.g. skintone of the player) and 'random effects' (which are usually not of specific interest and may well differ in any new experiment e.g. referee identifier). In this data study we might expect there to be some referees who have a greater propensity to award cards in any game and similarly some who award fewer. Which specific referees these are is not our primary question of interest and may be different in a new set of games. We would therefore include a 'random effect' to incorporate this random variation in referee strictness. Random effects are required in order to be able to generalise the results of the analysis outside of the population of study. / / Specifically we model the probability of a red card to player i in game j to be / / log p_{ij}/(1-p_{ij}) = sum beta^T_{ij} x_{ij} + w_{ij} / / where x_{ij} are the fixed effects and w_{ij} the random effect. / What are some references for the statistical technique that you chose? McCullagh and Nelder (1989). Generalised Linear Models. Chapman and Hall Hadfield JD (2010). MCMC methods for Multi-response Generalised Linear Mixed Models: The MCMCglmm R Package. Journal of Statistical Software, 33(2), 1-22 for general background we consulted Baayen, R. H. (2008). Analyzing linguistic data. A practical introduction to statistics using R. Which software did you use? If you used multiple kinds, please indicate what was accomplished with each piece of software (e.g., Data cleaning - R; Model estimation - SAS) Data recoding was done using Python, using the pandas library The analysis was run using R, using the lme4 (for frequentist analysis) and MCMCglmm (for Bayesian) packages What distribution did you specify for the outcome variable of red cards? We considered each game separately with the outcome for any player being binary - either a red (coded as 1) or not (coded as 0) What variables were included as covariates (or control variables) when testing research question 1: The relationship between player skin tone and red cards / received? Model selection was performed via Akaike's Information Criterion (AIC). Multiple models were considered before selecting that with the lowest AIC as our final model. / / The player and referee were included as random effects. The player random effect aims to account for variation amongst players in propensity to be booked (independent of skin tone) and the referee random effect aims to account for variation in strictness amongst referees (again independent of player skin tone) / / As fixed effects we included the covariates (as factors) of position played and league. This aimed to take account of the potentially unbalanced nature of the dataset i.e. that the four different leagues may have different numbers of cards awarded and that the skin tone of players in different leagues could be unbalanced. Similarly we felt that different positions are more/less likely to be red carded and certain positions may be more associated with particular skin tone. / / One could potentially consider league as a random (instead of fixed) effect but we felt there may be specific interest in the four major European leagues and not more general. We did also investigate if there was an interaction between league and skin tone (i.e. some leagues were more likely to give red cards to certain skin tones) but found no evidence for this. / / Finally as specific interest was in whether the bias score of the referee's home country was significant we also included two interactions: the first between the meanIAT (Implicit Bias) score and skin tone in our model; and the second between the mean (Explicit Bias) score and skin tone. Inclusion of both bias scores simultaneously means that one needs to be careful when interpreting results - comparisons are made conditional on all other variables being equal. / / It is not therefore possible to consider the Implicit and Explicit bias effects separately. One can instead compare two countries with the same Implicit bias scores but different Explicit bias scores; or vice versa. We are therefore concerned that results could be misinterpreted due to high correlation between implicit and explicit bias. Generalised statements about a general effect of explicit bias to compare two countries (as if one can ignore also the implicit bias) must be treated with caution. We are hence presuming that in any specific comparison/prediction between countries all the fixed variables are known or at least considered to be equal. All questions were answered using this model. / / We would have preferred to have had a single measure of bias but this did not seem possible according to the project brief. In our investigation we fitted separate models including explicit and implicit bias separately (and both together). Explicit bias was seen to be less significant than implicit bias - although there was v. little evidence to support either in the model. As a result, if we were to choose our 'best' model we would have dropped explicit bias and calculated the odds ratio simply for implicit bias. What variables were included as covariates (or control variables) when testing research question 2a: The relationship between referee country implicit skin-tone prejudice and red cards received by dark skin-toned players? As above - we used the same model What variables were included as covariates (or control variables) / when testing research question 2b: The relationship between referee country explicit skin-tone prejudice and red cards received by dark skin-toned players? As above - we used the same model What theoretical and/or statistical rationale was used for your choice of covariates included in the models? From background knowledge of soccer we reasoned that player identity, league, player position and referee identify would effect likelihood of a red card being awarded. / / We tested a number of different models and selected that with the lowest AIC as our final model (conditional on including the terms required to answer the questions posed). / / Since the data set was very large we felt it was appropriate to model the parameter estimate as normally distributed and so used +/- 1.96 * standard error to estimate the confidence intervals. What unit is your effect size in? odds ratio

Confid_Team24 How confident are you that the described approach below is suitable for analyzing the research questions?

	Unconfident	Rather unconfident	Somewhat unconfident	Neither confident nor unconfident	Somewhat confident	Rather confident	Confiden
RQ1) Skin color and red cards	○	○	○	○	○	○	○
RQ2a) Implicit cultural preferences	○	○	○	○	○	○	○
RQ2b) Explicit cultural preferences	○	○	○	○	○	○	○

Team24 Please provide feedback to the analytical approach described below.

Approach24 Analytical Approach Team24   What transformations (if any) were applied to the variables. Please be specific. Data were transformed to a format where a single match was taken as one observation. The format was obtained by multiplicating each row from the original data by the number of games for a given player-referee pair. For x of these rows a newly computed variable 'red' was set to 1, where x was sum of the yellowRed and redCards variables from the original data. The variable red took value of 0 for the remaining observations for a given player-referee pair. This variable then served as the dependent variable in all models. / For each player, a variable indicating the average number of goals in a game was computed from all matches in the dataset and the variable was then standardized. The variable was computed from all matches instead of just matches for a given player-referee pair because the number of red cards might itself influence number of goals or be influenced by it. Taking the average number of goals as a characteristic of a player instead of characteristic of a player-referee pair should mitigate influence of this confound. / birthday was transformed to age in days and standardized. / Six positions were recoded in three that omitted information about the side of the player which was deemed irrelevant with regard to the number of red cards (e.g. right winger and left winger were recoded to the position 'winger'). / height, weight, meanIAT, meanExp variables were standardized. Standardization for meanIAT and meanExp was done only from values of countries that were included in the subsequent analysis, i.e. those that had standard errors of the measure lower than â…• of standard deviation of mean values of the measure (described in data exclusion). / rating variable was computed as the standardized average of rater1 and rater2 variables. The standardization was done from values on player level and not match level to help interpretation. Were any cases excluded, and why? All cases that had missing values for any of the predictors used in a given model. Additionally, for models using meanIAT and meanExp, observations for referees from countries that had standard errors of the measure higher than 1/5 of standard deviation of mean values of the measure for all countries were excluded. The 1/5 of standard deviation threshold was chosen arbitrarily and meant that referees from countries that had a measure of attitudes computed from approximately 25 people or less were removed from the analysis. This exclusion was done so that the results were not influenced by unreliable values of attitude measures. What is the name of the statistical technique that you employed? Multilevel linear modelling Please describe the statistical technique you chose in more detail. Be specific, especially if your choice is not one you consider to be well-known. Multilevel modelling takes into account hierarchical structure of data. In the case of the current dataset, it means that it can model effects of individual players and referees. What are some references for the statistical technique that you chose? Baguley, T. (2012). Serious stats: A guide to advanced statistics for the behavioral sciences. Palgrave Macmillan.   Bates, D., Maechler, M., & Bolker, B. (2012). lme4: Linear mixed-effects models using S4 classes.   Gelman, A., & Hill, J. (2007). Data analysis using regression and multilevel/hierarchical models. Cambridge University Press. Which software did you use? If you used multiple kinds, please indicate what was accomplished with each piece of software (e.g., Data cleaning - R; Model estimation - SAS) R 3.0.2; package lme4 1.1.6 What distribution did you specify for the outcome variable of red cards? The outcome variable was binary. While multilevel logistic regression would be probably more appropriate given the binary nature of the outcome variable, multilevel linear regression was used because multilevel logistic regression turned out to be too computationally intensive. What variables were included as covariates (or control variables) when testing research question 1: The relationship between player skin tone and red cards / received? position, height, weight as predictors; player and referee as random effects / age and average number of goals were dropped from predictors because they increased AIC What variables were included as covariates (or control variables) when testing research question 2a: The relationship between referee country implicit skin-tone prejudice and red cards received by dark skin-toned players? position, height, weight, meanIAT as predictors; player and referee as random effects What variables were included as covariates (or control variables) / when testing research question 2b: The relationship between referee country explicit skin-tone prejudice and red cards received by dark skin-toned players? position, height, weight, meanExp as predictors; player and referee as random effects What theoretical and/or statistical rationale was used for your choice of covariates included in the models? The full model with all covariates was specified and subsequently covariates that increased AIC were discarded from the model. The models for questions 2a and 2b were built from the final model for question 1 by adding meanIAT or meanExp and its interaction with the average rating of skin color. What unit is your effect size in? Percentage increase of probability of getting a red card in a match when the predictor increases by one. For question 1, that corresponds to increase in predictor by one standard deviation. For questions 2a and 2b, the predictor is an interaction between standardized average skin color rating and standardized meanIAT or meanExp of country of origin of a referee.

Confid_Team27 How confident are you that the described approach below is suitable for analyzing the research questions?

	Unconfident	Rather unconfident	Somewhat unconfident	Neither confident nor unconfident	Somewhat confident	Rather confident	Confide
RQ1) Skin color and red cards	○	○	○	○	○	○	○
RQ2a) Implicit cultural preferences	○	○	○	○	○	○	○
RQ2b) Explicit cultural preferences	○	○	○	○	○	○	○

Team27 Please provide feedback to the analytical approach described below.

Approach27 Analytical Approach Team27   What transformations (if any) were applied to the variables. Please be specific. For question 1, 'rater1' and 'rater2' were averaged into a single variable, 'rating'. Spearman's rank correlation coefficient showed that the raters were 85.8% in agreement, so this was considered sufficient. (Spearman's was chosen over Pearson's due to the non-normal distribution of ratings.) / For question 2, meanIAT and meanExp ratings were multiplied by 100 to ease interpretation of effect sizes. Were any cases excluded, and why? Of the 146028 dyads in the original dataset, 21407 were missing ratings from both raters and were excluded from all analyses. For Q2a + b only, meanIAT and meanExp ratings were missing for for 153 dyads, which were excluded. What is the name of the statistical technique that you employed? Poisson Regression Please describe the statistical technique you chose in more detail. Be specific, especially if your choice is not one you consider to be well-known. I chose a regression model, since regression is suited to a single dependent variable and multiple independent variables/covariates. I chose a Poisson Regression specifically as the dependent variable, redCards, had a Poisson distribution (see below). What are some references for the statistical technique that you chose? http://courses.education.illinois.edu/EdPsy589/lectures/4glm3-ha-online.pdf https://www.casact.com/pubs/forum/07wforum/07w109.pdf http://www.csm.ornl.gov/~frome/BE/FP/FromeBiometrics83.pdf Which software did you use? If you used multiple kinds, please indicate what was accomplished with each piece of software (e.g., Data cleaning - R; Model estimation - SAS) I used the python standard library, scipy, and matplotlib to test the IRR of the skin color ratings. I used pandas and statsmodels to clean the data, test my assumptions re: the Poisson distribution, and run the regression analysis. What distribution did you specify for the outcome variable of red cards? I specified a Poisson distribution, which is commonly used for variables which are counts and for rare events. A Poisson distribution should be right-skewed - the outcome variable of redCards contains 98.7% datapoints with zero red cards, making the data strongly right-skewed. In a Poisson distribution, the variance should be roughly equal to the mean. In this data, variance is 0.01297 while the mean is 0.01275. I could find no references suggesting how close is 'close enough', but this seemed fine to me. What variables were included as covariates (or control variables) when testing research question 1: The relationship between player skin tone and red cards / received? rating'= The skin color rating, i.e. the independent variable of interest. / / 'games' = The number of games played by the player. This was included as a larger number of games provide more opportunities for a red card to be issued. Alternatively, familiarity with a player might enhance or attenuate bias. Consequently, no direction was predicted. / / 'goals' = The number of goals scored by a player. This was included as success might enhance or attenuate bias. No direction was predicted. / / 'yellowCards' = The number of yellow cards given by referee to player. As yellow cards and red cards are both awarded for aggressive play, one might expect them to be positively correlated with each other. However, the role of the yellow card as a "caution" or "warning" indicates the potential for a more complicated role (for instance, players with lighter skin might disproportionately receive yellow cards and those with darker skin red cards for the same behavior). / / 'meanIAT' = The mean implicit bias of the referee's country. / 'meanExp' = The mean explicit bias of the referee's country. / As implicit and explicit bias are conceptually separate things (which may well be correlated), I chose to consider them separately. / What variables were included as covariates (or control variables) when testing research question 2a: The relationship between referee country implicit skin-tone prejudice and red cards received by dark skin-toned players? meanIAT' = The mean implicit bias of the referee's country, i.e. the independent variable of interest. / / 'rating'= Skin color rating. / / 'games' = The number of games played by the player. This was included as a larger number of games provide more opportunities for a red card to be issued. Alternatively, familiarity with a player might enhance or attenuate bias. Consequently, no direction was predicted. / / 'goals' = The number of goals scored by a player. This was included as success might enhance or attenuate bias. No direction was predicted. / / 'yellowCards' = The number of yellow cards given by referee to player. As yellow cards and red cards are both awarded for aggressive play, one might expect them to be positively correlated with each other. However, the role of the yellow card as a "caution" or "warning" indicates the potential for a more complicated role (for instance, players with lighter skin might disproportionately receive yellow cards and those with darker skin red cards for the same behavior). / / 'meanExp' = The mean explicit bias of the referee's country. / What variables were included as covariates (or control variables) / when testing research question 2b: The relationship between referee country explicit skin-tone prejudice and red cards received by dark skin-toned players? meanExp' = The mean explicit bias of the referee's country, i.e. the independent variable of interest. / / 'rating'= Skin color rating. / / 'games' = The number of games played by the player. This was included as a larger number of games provide more opportunities for a red card to be issued. Alternatively, familiarity with a player might enhance or attenuate bias. Consequently, no direction was predicted. / / 'goals' = The number of goals scored by a player. This was included as success might enhance or attenuate bias. No direction was predicted. / / 'yellowCards' = The number of yellow cards given by referee to player. As yellow cards and red cards are both awarded for aggressive play, one might expect them to be positively correlated with each other. However, the role of the yellow card as a "caution" or "warning" indicates the potential for a more complicated role (for instance, players with lighter skin might disproportionately receive yellow cards and those with darker skin red cards for the same behavior). / / 'meanIAT' = The mean implicit bias of the referee's country. / What theoretical and/or statistical rationale was used for your choice of covariates included in the models? I chose to include all potential variables which might moderate the effect, regardless of subjective likelihood, as the number of potential variables was less than < 10. A robust effect should not be lost when correcting for that number of comparisons. What unit is your effect size in? For Qs 1, 2a and 2b, the effect sizes are Poisson regression coefficients (r). In Q1, 1 unit is equal to 1 on the skin color rating scale. In Q2a, 1 unit is equal to a change of 1 on the mean implicit bias score (which contains values from 0-1). In Q2b, 1 unit is equal to a change of 1 on the mean explicit bias score (which contains values from 0-1). Z scores are also provided.

Confid_Team25 How confident are you that the described approach below is suitable for analyzing the research questions?

	Unconfident	Rather unconfident	Somewhat unconfident	Neither confident nor unconfident	Somewhat confident	Rather confident	Confide
RQ1) Skin color and red cards	○	○	○	○	○	○	○
RQ2a) Implicit cultural preferences	○	○	○	○	○	○	○
RQ2b) Explicit cultural preferences	○	○	○	○	○	○	○

Team25 Please provide feedback to the analytical approach described below.

Approach25 Analytical Approach Team25   What transformations (if any) were applied to the variables. Please be specific. We used a Poisson model, so there's an implicit log transform on the redCard variable. / / We z-transformed all metric predictors, and used standard R factor scoring for our categorical predictors. / / Additionally, as we discuss later, we averaged the two raters' skin tone ratings for use in the primary research questions. Were any cases excluded, and why? We used listwise deletion for missing values, which effectively eliminated entries without ratings for skin tone. / / Additionally, at this stage, we randomly selected 250 referees as the focal analysis, leading to a total sample size of 10282. We intend to use the full data set for the final analysis, barring missing data deletions (imputing skin color based on the other variables seems problematic, at best). What is the name of the statistical technique that you employed? We used a hierarchical generalized linear model, with a log link function. Please describe the statistical technique you chose in more detail. Be specific, especially if your choice is not one you consider to be well-known. The above descirbes a multilevel Poisson regression (e.g., Gelman & Hill, 2007). We classified the data according to referees (i.e., the "random" intercepts vary by referee). / / That is, we believe that the data are probably best described as a conscious-or-unconscious choice process by the individual referees. We could also, potentially classify the data according to the players, but that could lead to confounding with the primary variable of interest, the player's skin color. We believe that there is enough information available in the form of covariates providing information about players to make the model exchangeable in players (e.g., Gelman, Carlin, Stern, Dunson, Vehtari, & Rubin, 2013, p. 5). / / Citation: / ----------- / Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B., Vehtari, A., & Rubin, D.B. (2013). Bayesian / data analysis (3rd Ed.). New York, NY: CRC Press. / / What are some references for the statistical technique that you chose? Gelman & Hill (2007). Data analysis using regression and hierarchical/multilevel models. New York, NY: Cambridge University Press. is a good pedagogical reference for this model. Which software did you use? If you used multiple kinds, please indicate what was accomplished with each piece of software (e.g., Data cleaning - R; Model estimation - SAS) R What distribution did you specify for the outcome variable of red cards? Poisson, though we suspect zero inflation and extra-Poisson variation that may not be adequately captured by our model. The hierarchical structure of the model should help account for some of the extra-Poisson variation, but may not be sufficient to model the excess zeroes with sufficient accuracy. / / We may wish to explore hierarchical extensions of a zero-inflated model for the final analysis. This may need to be implemented using a fully Bayesian model, but so far, we have had little luck in getting that model to run. What variables were included as covariates (or control variables) when testing research question 1: The relationship between player skin tone and red cards / received? height, weight, goals, position What variables were included as covariates (or control variables) when testing research question 2a: The relationship between referee country implicit skin-tone prejudice and red cards received by dark skin-toned players? height, weight, goals, position What variables were included as covariates (or control variables) / when testing research question 2b: The relationship between referee country explicit skin-tone prejudice and red cards received by dark skin-toned players? height, weight, goals, position What theoretical and/or statistical rationale was used for your choice of covariates included in the models? We wanted to keep some consistency with the Price & Wolfers paper, we believed that these variables would help make the model more exchangeable in the players, and they made sense to my research assistant (a former college soccer player). What unit is your effect size in? We report exponentiated coefficients from the multilevel Poisson regression below. We believe this is the most appropriate approach to reporting the effect, given the discrete nature of the outcome variable. We are open to a more standardized approach, but would like a strong rationale for it. A variance explained measure may be appropriate, but would, of course be an approximation based on the reduction in deviance. Also of note, at this point, the confidence intervals we are reporting are relatively naive, being based on the point estimate plus and minus two standard errors.

Confid_Team26 How confident are you that the described approach below is suitable for analyzing the research questions?

	Unconfident	Rather unconfident	Somewhat unconfident	Neither confident nor unconfident	Somewhat confident	Rather confident	Confide
RQ1) Skin color and red cards	○	○	○	○	○	○	○
RQ2a) Implicit cultural preferences	○	○	○	○	○	○	○
RQ2b) Explicit cultural preferences	○	○	○	○	○	○	○

Team26 Please provide feedback to the analytical approach described below.

Approach26 Analytical Approach Team26 What transformations (if any) were applied to the variables. Please be specific. Missing values in player position is re-coded as a new category. These players may be able to play multiple positions and hence have special features. Player position is controlled in the analyses as a dummy variable for each position. Were any cases excluded, and why? First, cases that do not include skin tone measures are exclued. Second, as required for multilevel analyses, referees with fewer than 3 players played under are excluded, and referees from countries with fewer than 3 referees are exclued. What is the name of the statistical technique that you employed? Three-level random effects model with Poisson estimation Please describe the statistical technique you chose in more detail. Be specific, especially if your choice is not one you consider to be well-known. We choose the three-level random effects model with Poisson estimation for two reasons. First, multilevel random effects models are appropriate for analyzing nested or non-independence dataset like this one. In this dataset, each player-referee data point is nested within a referee who is further nested within a country. Second, the DV - red cards given to a player is a count variable with a distribution similar to that of a poisson distribution. For this count variable, generalized linear models with Poisson estimation is most appropriate. What are some references for the statistical technique that you chose? Notes on Intermediate Data Analysis by Blair Wheaton. Which software did you use? If you used multiple kinds, please indicate what was accomplished with each piece of software (e.g., Data cleaning - R; Model estimation - SAS) SAS 9.2 What distribution did you specify for the outcome variable of red cards? Poisson distribution. What variables were included as covariates (or control variables) when testing research question 1: The relationship between player skin tone and red cards / received? Games; yellowcards; age, weight, height, victories, ties, meanExp, meanIat, position, league, and a generated variable Refexp indicating how many players have played under the referee. What variables were included as covariates (or control variables) when testing research question 2a: The relationship between referee country implicit skin-tone prejudice and red cards received by dark skin-toned players? Games; yellowcards; age, weight, height, victories, ties, meanExp, position, league, and a generated variable Refexp indicating how many players have played under the referee. What variables were included as covariates (or control variables) / when testing research question 2b: The relationship between referee country explicit skin-tone prejudice and red cards received by dark skin-toned players? Games; yellowcards; age, weight, height, victories, ties, meanIat, meanIat, position, league, and a generated variable Refexp indicating how many players have played under the referee. What theoretical and/or statistical rationale was used for your choice of covariates included in the models? Games: the number of red cards one received is positively affected by how many games he played; / Yellowcards: since the purpose of the study is to examine the unique effect of skin tone on red cards received, yellow cards; indicating less servere punishment is necessary to be controlled for; / Age, as a common demographic variable is normally controlled for; / Physical conditions (i.e., weight and height) may confound with skin tone as darker skinned players may be stronger or weaker; / Victories and ties: players often commit a servere foul out of frustration from being defeated; / meanIat and meanExp: it is necessary to control for implict or explicit bias score when studying the other's interaction with skin tone to rule out potential confounding effect; / Position, players playing defensive positions may be more likely to commit fouls in general; / League, different leagues may have different styles. Some may be more physical and competitive than others. / Refexp, it may be easier for an experienced referee to give red cards to players. What unit is your effect size in? Unstandardized parameter estimates (b)

Confid_T29 How confident are you that the described approach below is suitable for analyzing the research questions?

	Unconfident	Rather unconfident	Somewhat unconfident	Neither confident nor unconfident	Somewhat confident	Rather confident	Confider
RQ1) Skin color and red cards	❍	❍	❍	❍	❍	❍	❍
RQ2a) Implicit cultural preferences	❍	❍	❍	❍	❍	❍	❍
RQ2b) Explicit cultural preferences	❍	❍	❍	❍	❍	❍	❍

Team29 Please provide feedback to the analytical approach described below.

ApprT29 Analytical Approach Team29   What transformations (if any) were applied to the variables. Please be specific. The current transformations are: / log(player matches) / log(referee matches) / dummy code (position) / dummy code (league) / Were any cases excluded, and why? For simplicity, cases with missing data were excluded. That said, if I have more time I would have liked to explore various missing data imputation procedures. What is the name of the statistical technique that you employed? In broad terms, the technique might be called Bayesian hierarchical modeling. Please describe the statistical technique you chose in more detail. Be specific, especially if your choice is not one you consider to be well-known. At present, this is a work in progress. Given that these analyses are somewhat incomplete, I considered not posting my analyses, but Raphael Silberzahn suggested that it still might be useful to share my approach. In particular, progress has been slowed by the fact that estimating the model on the full dataset on a standard desktop takes around a day to run with reasonable MCMC settings. This has substantially slowed the process of exploring the best way to specify the model, how to code covariates, and deciding on which covariates to retain. I had also planned on extending the model to explicitly incorporate country level effects for assessing research question 2. Nonetheless, here is the model at present: / / Given that it is not possible to enter math into this online form and mathematics is the essence of the model, I have uploaded an image of the description of the model here: http://i.imgur.com/im59WrH.png / / What are some references for the statistical technique that you chose? Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). Bayesian data analysis. CRC press. Gelman, A., & Hill, J. (2006). Data analysis using regression and multilevel/hierarchical models. Cambridge University Press. Which software did you use? If you used multiple kinds, please indicate what was accomplished with each piece of software (e.g., Data cleaning - R; Model estimation - SAS) R was used for data cleaning and preparation. JAGS was used for Bayesian model estimation. JAGS was called from R using the rjags package. R was used to process the output from JAGS. What distribution did you specify for the outcome variable of red cards? A binomial distribution was used. What variables were included as covariates (or control variables) when testing research question 1: The relationship between player skin tone and red cards / received? What variables were included as covariates (or control variables) when testing research question 2a: The relationship between referee country implicit skin-tone prejudice and red cards received by dark skin-toned players? What variables were included as covariates (or control variables) / when testing research question 2b: The relationship between referee country explicit skin-tone prejudice and red cards received by dark skin-toned players? What theoretical and/or statistical rationale was used for your choice of covariates included in the models? Given that this is a Bayesian model, one of the benefits is to have a model that integrates the uncertainty in parameter estimation over all parameters. Thus, the aim is to have a single model that would be used to provide parameters for research questions 1, 2a, and 2b. As such the set of control variables would be the same for all three research questions. / While I have not finalized this aspect of the analysis, I had flagged the following as likely control variables: dummy coded(league_country); log(total player matches); dummy coded (player position); log(referee matches) / What unit is your effect size in? I think an odds ratio is a good choice.

Confid_T30 How confident are you that the described approach below is suitable for analyzing the research questions?

	Unconfident	Rather unconfident	Somewhat unconfident	Neither confident nor unconfident	Somewhat confident	Rather confident	Confide
RQ1) Skin color and red cards	❍	❍	❍	❍	❍	❍	❍
RQ2a) Implicit cultural preferences	❍	❍	❍	❍	❍	❍	❍
RQ2b) Explicit cultural preferences	❍	❍	❍	❍	❍	❍	❍

Team30 Please provide feedback to the analytical approach described below.

ApprT30 Analytical Approach Team30 What transformations (if any) were applied to the variables. Please be specific. The two ratings of skin color had good internal consistency (alpha = 0.9602) so we averaged them into a single index. / / 123,056 of these dyads resulted in zero red cards, 1541 in one red card, and 24 in two red cards. Of course, having received a red card in 47 games is far different from having received a red card based on a single played game. The obvious thing to do here is to disaggregate the dyads into games. However, this was not straightforward, since the correlation with other variables also aggregated (e.g., number of yellow-cards, number of goals) is unknown on the level of games. Instead, we divided each game-specific outcome (red cards, yellow cards, victories, defeats, goals, ties) by the number of games played, thus rendering the variables the probability of the event occurring during a game. The exception was the goals per game variable where it is possible to score more than one goal per game. Although this approach resulted in fewer observations than if the data had been disaggregated, the multiple observations should add reliability to the estimate. / / All predictors were standardized. Were any cases excluded, and why? 21407 observations had missing values on the skin color variable and were thus omitted from subsequent analysis What is the name of the statistical technique that you employed? Logistic regression with multi-way clustering for Q1, and logistic regression with clustering for Q2 Please describe the statistical technique you chose in more detail. Be specific, especially if your choice is not one you consider to be well-known. Because there was very little variation beyond 0 or 1 red cards, we decided on treating the dependent variable as a dummy and choose to analyze the results by means of logistic regression, treating any non-zero probability as having received (at least) one red card. Another reason for choosing logistic regression is that the dependent variable does not have very much meaning in an absolute sense and thus absolute changes in probabilities (e.g., from an OLS, probit or tobit) would be hard to interpret. Odds ratios, in contrast, are commonly used as effect sizes for relative effects. In this case, the relative change in the odds of receiving a red card as a function of your skin color.Because the 123,056 observations was based on observations from 1572 players and 3147 referees, and one could expect correlation both within players and within referees in the number of red cards, it was necessary to take into account this hierarchical structure of the data. One approach to do so is to adjust the standard errors through clustering. The main advantage of this over other approaches (e.g., multi-level modeling) is that conventional regression analysis, such as logit in this case, can be used with a simple adjustment. However, the program we used (STATA 12) does not include a command for multi-way clustering. We thus used an add-on (logit2) for this. Thus for Q1, we used a logistic regression with two-way clustering on referee and player, with the dependent variable being whether an encounter between a player and a referee resulted in a red-card or not. In the first model, we included only skin tone (standardized) as predictor. In the second model we also included controls: weight and height (standardized) and position (effect coded in order to base the constant on the grand mean and simply comparison to unadjusted model). For Q2, both referees and players are nested within the country of which the IAT and explicit measure is taken from, and we thus clustered only on this highest level (using the logit command in STATA 12). As in the previous analysis, we started without controls and thus only included the IAT and the explicit variable (standardized) as well as their interaction with skin tone rating. These interaction effects served as our answer to Q2. If prejudice matters, then there should be a positive interaction effect in that the skin tone effect is stronger for those with higher prejudice on the IAT and/or the explicit measures. Finally, we re-ran the analysis with the same-control mentioned above. What are some references for the statistical technique that you chose? A. Colin Cameron, Jonah B. Gelbach & Douglas L. Miller (2011) Robust Inference With Multiway Clustering, Journal of Business & Economic Statistics, 29:2, 238-249, DOI: 10.1198/jbes.2010.07136 for the article and http://www.kellogg.northwestern.edu/faculty/petersen/htm/papers/se/se_programming.htm for the STATA scripts Which software did you use? If you used multiple kinds, please indicate what was accomplished with each piece of software (e.g., Data cleaning - R; Model estimation - SAS) STAT/transfer 11 to convert the data to STATA-format. STATA 12 with logit2 add-on for everything else. What distribution did you specify for the outcome variable of red cards? Logit What variables were included as covariates (or control variables) when testing research question 1: The relationship between player skin tone and red cards / received? First model: rating / Second model: rating, position (effect coded), weight, height. What variables were included as covariates (or control variables) when testing research question 2a: The relationship between referee country implicit skin-tone prejudice and red cards received by dark skin-toned players? First model: rating, IAT, explicit, rating*iat, rating*explicit / Second model: rating, IAT, explicit, rating*iat, rating*explicit, position (effect coded), weight, height. What variables were included as covariates (or control variables) / when testing research question 2b: The relationship between referee country explicit skin-tone prejudice and red cards received by dark skin-toned players? First model: rating, IAT, explicit, rating*iat, rating*explicit / Second model: rating, IAT, explicit, rating*iat, rating*explicit, position (effect coded), weight, height. What theoretical and/or statistical rationale was used for your choice of covariates included in the models? The focus was on variables that may theoretically co-vary with both skin color and with number of red cards. Because I have very poor understanding of soccer, the choice was not obvious to me. However, the position played seemed relevant (e.g., defenders receiving more red cards then attackers) and so did weight and height. These three variables can also be expected to correlate with skin tone. This was confirmed in bivariate analysis before entering them as controls in the regression. What unit is your effect size in? Odds ratio

Confid_T31 How confident are you that the described approach below is suitable for analyzing the research questions?

	Unconfident	Rather unconfident	Somewhat unconfident	Neither confident nor unconfident	Somewhat confident	Rather confident	Confiden
RQ1) Skin color and red cards	❍	❍	❍	❍	❍	❍	❍
RQ2a) Implicit cultural preferences	❍	❍	❍	❍	❍	❍	❍
RQ2b) Explicit cultural preferences	❍	❍	❍	❍	❍	❍	❍

Team31 Please provide feedback to the analytical approach described below.

ApprT31 Analytical Approach Team31   What transformations (if any) were applied to the variables. Please be specific. First, we encoded several nominal string variables into numeric variables. For example, club, leaguecontry, and position were encoded into numeric values to put them as fixed effects in the regression model. Next, we created a new binary dependent variable. Since we are interested in whether or not referees gave red cards, we created independent variable including 0 and 1. 0 indicates no red card, while 1 indicates red cards either from red card or from two yellows. Indeed, for research question 2, we created interaction terms by multiplying rating and two cultural preference scores on lighter skin color (e.g., mean IAT and explicit bias score). Were any cases excluded, and why? First of all, we excluded responses which have at least one missing in any of the variables. Next, to rule out rater effect for the skin color rating, we excluded the responses that rater 1 and rater 2 rated differently. What is the name of the statistical technique that you employed? We used logistic regression for the analysis. Please describe the statistical technique you chose in more detail. Be specific, especially if your choice is not one you consider to be well-known. For both research questions, we used logistic regression. Since the key research questions are whether skin color or cultural preference on skin color influence the referees' tendency to give more red cards to darker skin players than lighter skin players, we created the dependent variable as a categorical variable (e.g., 0: no red card, 1: red card either from two yellows or red card). For binary dependent variable, it is proper to use logistic regression, so our analysis was mainly based on logistic regression. What are some references for the statistical technique that you chose? http://www.udel.edu/FREC/ilvento/BUAD820/MOD504.pdf, http://www.jstor.org/discover/10.2307/23045634?uid=3738392&uid=2&uid=4&sid=21104202913557, Which software did you use? If you used multiple kinds, please indicate what was accomplished with each piece of software (e.g., Data cleaning - R; Model estimation - SAS) We used STATA Version 12 for data cleaning and model estimation. What distribution did you specify for the outcome variable of red cards? Since we created a new variable for the red cards as a binary variable (0: no red card, 1: red cards either from red card or from two yellows), we could say we specified the outcome variable as a binominal distribution. What variables were included as covariates (or control variables) when testing research question 1: The relationship between player skin tone and red cards / received? We put height, weight, goals, and the number of games as control variables in the model. Also, we put club, leaguecontry, and position as control variables by putting them as dummy variables, since they are nominal or categorical variables. What variables were included as covariates (or control variables) when testing research question 2a: The relationship between referee country implicit skin-tone prejudice and red cards received by dark skin-toned players? The control variables and fixed effects that we used for research question 2a were the same as those for the research question 1, except that we excluded refcountry from the control variables since here we are interested in the effect of cultural preference on lighter skin across different countries. What variables were included as covariates (or control variables) / when testing research question 2b: The relationship between referee country explicit skin-tone prejudice and red cards received by dark skin-toned players? The control variables that we used for research question 2b were the same as those for the research question 2a. What theoretical and/or statistical rationale was used for your choice of covariates included in the models? First, we thought physical features could influence red card, since being hit by taller and heavier players may give greater effect. Next, we controlled the number of games because it is possible that the reason for more red cards may be caused by the fact that they simply played more games. Also, the number of goals can be an indicator of the character of the game, so we included it as a control variable. We are interested in the effect of skin color or cultural preference in referees' country on red cards in general across positions, leagues, clubs, and countries, not specific effect, we added those variables as control variables by adding them as dummy variables in the regression model. What unit is your effect size in? We looked at marginal effects of rating (research 1), the interaction between rating and mean IAT, and the interaction between rating and mean explicit bias score.

Q89 This was the last approach. You may go back to review your comments. Once you press submit, you will no longer be able to make changes. Thank you for your feedback!

Q90 Any comment you would like to make to us us can be written in the form below. For a quicker response, please write an e-mail.