ONLINE SUPPLEMENTS

Supplement 1: Publicly posted project description

NOTE: This initial project description was publicly posted here: https://docs.google.com/document/d/1uCF5wmbcL90qvrk_J27fWAvDcDNrO9o_APkicwRkOKc/edit

Crowdsourcing Research: Many analysts, one dataset
Research Protocol
Spring 2014

Research Question: Are soccer referees more likely to give red cards to dark skin toned players than light skin toned players?

Overview

In a standard scientific analysis, one analyst or team presents a single analysis of a dataset. However, there are often a variety of defensible analytic strategies that could be used on the same data. Variation in those strategies could produce very different results.
We introduce the approach of "crowdsourcing a dataset." Multiple independent analysts are recruited to investigate the same hypothesis or hypotheses on the same dataset in whatever manner they see as best. The independent analysis strategies produce two datasets of interest: (1) the variation in analysis strategies, and (2) the variation in estimated effects. These two can be partially independent. Different analysis strategies may converge to a very similar estimated effect - indicating robustness despite variation in analysis strategies. Alternatively, the estimated effect may be highly contingent on analysis strategy. In the latter case, there are at least two methods of resolution: (1) consider the central tendency of the estimated effects to be the most accurate, or (2) critically evaluate the analysis strategies to determine whether one or more should be elevated as the preferred analysis.

This approach should be especially useful for complex datasets in which a variety of analytic approaches could be used, and when dealing with controversial issues about which researchers and others have very different priors. If everyone comes up with the same results, then scientists can speak with one voice. If not, the subjectivity and conditionality on analysis strategy is made transparent. Further, when crowdsourcing a dataset, the potential for errors and suboptimal analyses are reduced.

This first project establishes a protocol for independent simultaneous analysis of a single dataset by multiple teams, and resolution of the variation in analytic strategies and effect estimates

among them. Next, we summarize the research question, process for collaboration, and the available dataset. The Open Science Framework project page is https://osf.io/gvm2z/.

Research Questions

For this first project, we crowdsource the questions of whether soccer referees are more likely to give red cards to dark skin toned players than light skin toned players, and whether this effect is moderated by skin-tone prejudice across cultures. The available dataset provides an opportunity to identify the magnitude of the relationship among these variables. It does not offer opportunity to identify causal relations.

Research Question 1: Are soccer referees more likely to give red cards to dark skin toned players than light skin toned players?

Research Question 2: Are soccer referees from countries high in skin-tone prejudice more likely to award red cards to dark skin toned players?

Relevant background

For Question 1: Research on assimilation to stereotypes in social perception (Bodenhausen, 1988; Correll et al., 2002; Hugenberg & Bodenhausen, 2003) and cultural preferences for light skin (Maddox & Gray, 2002; Sidanius et al., 2001; Twine, 1998) predicts that darker skin tone will be associated with receiving more red cards. On the other hand, research on accountability (Lerner & Tetlock, 1999), and the debiasing effects of real world professional experience (List, 2003; Levitt & List, 2008) gives reasons to expect no such effect. Although concluding the null is always difficult, our large sample size gives us much greater leeway than usual with regard to concluding no evidence of bias.

For Question 2: Research and theory on the roots of perceptual biases in cultural socialization (Banaji, 2001; Greenwald & Banaji, 1995) suggests growing up in a society that favors light over dark skin should ingrain such prejudices in individual members of that culture. On the other hand, implicit and explicit prejudices measured at the aggregate level of societies may not related to individual-level judgments as these are different levels of analysis and relatively “distant” predictors.

Related Research

There is some relevant literature looking at other sports, specifically basketball and baseball. Price and Wolfers (2010) demonstrated a same-race bias in NBA foul calls (e.g., White referees call more fouls on Black players) and rebutted the NBA's criticisms in a follow up paper (Price & Wolfers, 2011). Parsons et al. (2011) and Kim and King (in press) demonstrate racial bias in calls by baseball umpires. Pope, Price, and Wolfers (2013) show that after the publicity around the original Price and Wolfers paper, the same-race bias shown in NBA referee calls was eliminated. This provides a strong ethical impetus for carrying out the present project. The publicity and controversy surrounding the original Price and Wolfers paper also makes it even more important than usual to get things right when looking for evidence of similar biases among soccer referees.

Project Coordination and Authorship

Raphael Silberzahn and Dan Martin are the project coordinators. Eric Uhlmann is the lead writer and Brian Nosek will supervise the project. The two project coordinators and lead writer will be the first three authors followed by alphabetical listing of all other authors, and then Brian Nosek.

Authorship is earned by completing and submitting a reproducible analysis within the stated timeframe. This includes: (1) the code for the analysis and specification of analysis package required to execute the analysis, (2) a description of the rationale for the analysis strategy, (3) a complete written description of the analysis strategy, and (4) a description of the result including specification of the effect estimate in effect size units (d, r, R^2 or odds ratio) and 95% confidence interval around the estimate.

Planned Timeline

There are seven phases for this crowdsourcing project. In order to meet the timeline, some later phases may commence while earlier phases are in process. For example, some of the report will be written while final data analyses are still in process.

1. **Registration:** Registration via Google Forms document and with the Open Science Framework: project page is https://osf.io/gvm2z/ (Complete by May 18th, 2014).
2. **1st Round Analyses**: First round of Analyses conducted until June 15, EST and analytical approaches are uploaded and shared with other research teams. Initial findings are shared with the project coordinators but not with other research teams.
3. **Round Robin Feedback Round**: Research teams comment and provide suggestions on other teams' research approaches (until June 29, 2014).

4. **2nd Round Analyses**: Research teams refine their analytical approach and upload their final analyses (until 20th of July, 2014).
5. **Working Paper**: A working paper presenting and discussing the different results will be circulated to research teams (before August 3rd, 2014) and made available for the wider public (until August 17th, 2014).

Elaboration of Project Stages

1. Registration

Research teams consisting of one or several individual researchers may register to participate in this project via the this form. After registration, participants receive an invitation on the Open Science Framework to access the project data.

2.1st Round Analyses

After registration, research teams will be given access to the data and will develop an analytical approach and engage in data analyses independently of other teams. At the end of this stage, it is expected that teams submit a short summary of their analytical approach.

In order for research teams not to converge towards a particular outcome, teams will disclose their findings from this stage to the project coordinators but not to other research teams. This procedure helps keep track of changes to analytical approaches and how initial findings and conclusions change over time, which is a potentially important insight that this crowdsourcing project may reveal.

The following will describe the dataset and available variables in greater detail.

The Dataset

From a company for sports statistics, we obtained data and profile photos from all soccer players (N = 2,053) playing in the first male divisions of England, Germany, France and Spain in the 2012-2013 season and all referees (*N* = 3,147) that these players played under in their professional career (see Fig. S1). We created a dataset of player–referee dyads including the number of matches players and referees encountered each other and our dependent variable, the number of red cards given to a player by a particular referee throughout all matches the two encountered each other.

Player's photo was available from the source for 1,586 out of 2,053 players. *Players' skin tone* was coded by two independent raters blind to the research question who, based on their profile

photo, categorized players on a 5-point scale ranging from “very light skin” to “very dark skin” with “neither dark nor light skin” as the center value.

Fig. S1: Player overview with list of referees and player-referee statistics, such as matches, goals, and cards.

Schiedsrichter	Land		S	U	N	⚽			
Juan Pompei		14	9	2	3	9	2	0	0
Sergio Pezzotta		12	8	3	1	7	1	0	0
Carlos Maglio		12	4	2	6	2	1	0	0
Saul Laverni		10	4	1	5	3	0	0	0
Federico Beligoy		9	3	3	3	4	0	0	0
Pablo Lunati		9	5	0	4	2	0	0	0
Diego Abal		8	4	1	3	6	0	0	0
Héctor Baldassi		7	2	5	0	6	0	0	0
Néstor Pitana		7	2	1	4	0	0	0	0
Carlos Amarilla		6	4	0	2	2	2	0	0
Gustavo Bassi		6	3	1	2	1	0	0	0
César Ramos Palazuelos		5	2	2	1	3	0	0	0
Rafael Furchi		5	2	2	1	2	1	0	1
Carlos Chandía		5	2	2	1	1	0	0	0
Patricio Loustau		5	0	3	2	1	0	0	0
Roberto García		4	3	1	0	3	0	0	0
Alejandro Sabino		4	2	2	0	1	0	0	0
Gabriel Favale		4	1	2	1	0	0	0	0

Additionally, implicit bias scores for each referee country were calculated using a race implicit association test (IAT), with higher values corresponding to faster white | good, black | bad associations. Explicit bias scores for each referee country were calculated using a racial thermometer task, with higher values corresponding to greater feelings of warmth toward whites versus blacks. Both these measures were created by aggregating data from many online users in referee countries taking these tests on Project Implicit.

Data Structure

The dataset is available as a list with 146,028 dyads of players and referees and includes details from players, details from referees and details regarding the interactions of player-referees. A summary of the variables of interest can be seen below. A detailed description of all variables included can be seen in the README file on the project website.

Variable Name:	Variable Description:
playerShort	short player ID
player	player name
club	player club
leagueCountry	country of player club (England, Germany, France, and Spain)
height	player height (in cm)
weight	player weight (in kg)
position	player position
games	number of games in the player-referee dyad
goals	number of goals in the player-referee dyad
yellowCards	number of yellow cards player received from the referee
yellowReds	number of yellow-red cards player received from the referee
redCards	number of red cards player received from the referee
photoID	ID of player photo (if available)
rater1	skin rating of photo by rater 1
rater2	skin rating of photo by rater 1
refNum	unique referee ID number (referee name removed for anonymizing purposes)
refCountry	unique referee country ID number
meanIAT	mean implicit bias score (using the race IAT) for referee country
nIAT	sample size for race IAT in that particular country
seIAT	standard error for mean estimate of race IAT
meanExp	mean explicit bias score (using a racial thermometer task) for referee country
nExp	sample size for explicit bias in that particular country
seExp	standard error for mean estimate of explicit bias measure

3. **Round Robin Feedback Round**: After submitting their analytical approach, teams are invited to view others' approaches, take inspiration from them and comment and reflect the different strategies. Further details of this process are to be announced.

4. **2nd Round Analyses**: Based on their initial analyses, and the input received during the Round Robin Feedback round research teams refine their analytical approach and work out their final analyses and conclusion they draw from the data.

5. **Working Paper:** A single General Discussion briefly covers the results reached by each team and tries to integrate them. We also reflect on how the crowdsourcing went.

If everyone reached similar conclusions, scientist can speak with one voice on a socially important issue, which is a nice contribution. If different analysts reach very different results with multiple, defensible approaches, this is also a contribution in highlighting that there is a great deal of subjectivity in science. If errors or suboptimal analyses were uncovered when similar analyses by different analysts were compared, that's a contribution too as scientific errors were avoided through the use of many independent analysts.

There are also some potential drawbacks of crowdsourcing that may be worth discussing. The results section will likely become very long because of the need to present the results of so many

different analysts. It is also perhaps inefficient to always have many different analysts analyze the same dataset to test the same hypothesis. There is limited professional reward for many of those involved, most of whose names are lost in a long author string. In some cases crowdsourcing could lead to a "Tower of Babel" problem, where one analytic approach is actually optimal but it is lost amid less optimal (if still defensible) approaches.

Crowdsourcing is likely to be most useful in cases like this involving complicated datasets, multiple plausible hypotheses, and high levels of controversy. This is a case where all this effort will likely be worth it.

References for S1

Banaji, M. R. (2001). Implicit attitudes can be measured. In H. L. Roediger, III, J. S. Nairne, I. Neath, & A. Surprenant (Eds.), *The nature of remembering: Essays in honor of Robert G. Crowder* (pp. 117-150). Washington, DC: American Psychological Association.

Bodenhausen, G. V. (1988). Stereotypic biases in social decision making and memory: Testing process models of stereotype use. *Journal of Personality and Social Psychology*, *55*, 726-737.

Correll, J., Park, B., Judd, C.M., & Wittenbrink, B. (2002). The police officer's dilemma: Using ethnicity to disambiguate potentially threatening individuals. *Journal of Personality & Social Psychology, 83,* 1314–1329.

Greenwald, A. G., & Banaji, M. R. (1995). Implicit social cognition: Attitudes, self-esteem, and stereotypes. *Psychological Review*, *102*, 4-27.

Hugenberg, K., & Bodenhausen, G. V. (2003). Facing prejudice: Implicit prejudice and the perception of facial threat. *Psychological Science, 14*, 640-643.

Kim, J., & King, B.G. (in press). Seeing stars: Matthew effects and status bias in Major League Baseball umpiring. *Management Science*.
News: http://mobile.nytimes.com/2014/03/30/opinion/sunday/what-umpires-get-wrong.html

Lerner, J.S., & Tetlock, P.E. (1999). Accounting for the effects of accountability. *Psychological Bulletin,* 125(2), 255-275.

Levitt, S.D., & List, J.A. (2008). Homo economicus evolves. *Science, 319*, 909–910.

List, J.A. (2003). Does market experience eliminate market anomalies? *Quarterly Journal of Economics*, *118(1)*, 41–71.

Maddox, K.B. & Gray, S. (2002). Cognitive representations of African Americans: Re-exploring the role of skin tone. *Personality and Social Psychological Bulletin, 28*, 250-259.

Parsons, C., Sulaeman, J., Yates, M., & Hamermesh, D. (2011). Strike Three: Discrimination, Incentives, and Evaluation. *American Economic Review, 101,* 1410–1435.

Pope, D., Price, J., & Wolfers, J. (2013). *Awareness Reduces Racial Bias.* NBER Working Paper No. 19765.

Price, J., & Wolfers, J. (2010). Racial discrimination among NBA referees. *Quarterly Journal of Economics*.

Price, J., & Wolfers, J. (2011). Biased Referees?: Reconciling Results with the NBA's Analysis. *Contemporary Economic Policy*.

Sidanius, J., Peña, Y. & Sawyer, M. (2001). Inclusionary discrimination: Pigmentocracy and patriotism in the Dominican Republic. *Political Psychology, 22,* 827-851.

Twine, F. W. (1998). *Racism in a racial democracy*. New Brunswick, NJ: Rutgers University Press.

Supplement 2: Additional notes on the research process

1. The data included identifying information for each player such as name, club, and league played at the time the data was collected. This identifying information was helpful as soon after the initial posting of the data, one project member noted a few mismatches between players and their height, which likely had been introduced during the data cleaning process. After these issues were raised, the data was taken offline and we went back to the original data source. Two project coordinators created independent clean datasets from the original source. Both datasets were checked against each other for accuracy and spot checks with the original source revealed no differences, thus this updated dataset was provided to the analysis teams. Illustrating an important benefit of crowdsourcing science, already at this stage the multitude of researchers involved benefitted the project by helping to ensure that errors were caught at an early stage and could be addressed.

2. To aggregate the final results into a common effect size, further exchange communication occurred between the project coordinators and some team leaders after the submission of final reports. Project coordinators thereby assisted in the conversion of obtained results into the standardized effect size units reported in this paper (Cohen's d, standardized regression weight, odds ratio, or risk ratio).

Supplement 3: Complete surveys sent to analysis teams

1. Registration E-Mail

Dear <FirstName>,

Thank you very much for joining the Crowdsourcing Research Project. We are excited to have you in the team! I am sending you below some further information, which will help us work together. Raphael Silberzahn and Dan Martin are the project coordinators. Eric Uhlmann is the lead writer and Brian Nosek will supervise the project. Raphael (mail@raphael.rs) and Dan (dpmartin42@gmail.com) are your first points of contact for any question you may have. More information about the project itself, as well as a timeframe and further information are in our google document: https://docs.google.com/document/d/1uCF5wmbcL90qvrk_J27fWAvDcDNrO9o_APkicwRkOKc/edit We will update this document over time but will also inform you via e-mail of major changes. At this point you may likely ask what the next steps are.

(1) As a first step, I will register you as a collaborator on our project space at the Open Science Framework: https://osf.io/gvm2z/ If you are already registered at the OSF than you should be able to view this project in your dashboard. If you're not yet registered at the OSF, you will receive an e-mail.

(2) The dataset will be made available on Monday 28th of April, from which time on you may start working on your analyses. You will have time until June 15th, to upload a documentation of your analytical approach and your results. Your analytical approach but not the initial findings are then shared with other research teams and following that date, research teams will provide comments and suggestions, which should help refine your analyses thereafter. A more detailed overview of these steps is documented in our google document.

We are very excited to work on this project together with you!
All the best,
Raphael, Dan, Eric and Brian

2. Analytical Approach Collection E-Mail

Dear ${m://FirstName},

Our Crowdsourcing project is getting to the final phase! We hope you enjoyed working with the data and send you the link below to submit your analytical approach. Deadline for submission is June 15th EST. As this is a delayed submission, please submit as soon as possible and let me know by e-mail afterwards. After, we will prepare all approaches and organize the feedback round. To make sure that other teams will be able to give you high quality feedback, please try give as much information as you can regarding the analytical approach that you chose.

Best regards,
Raphael, Dan, Eric and Brian

Follow this link to submit your analytical approach:
${l://SurveyLink?d=Take the Survey}
Or copy and paste the URL below into your internet browser:
${l://SurveyURL}

${l://OptOutLink?d=To%20opt%20out%20from%20the%20crowdsourcing%20project,%20please%20click%20here.}

3. Analytical Approach Collection Questionnaire

Analytical Approach - Collection

Q1 ${m://FirstName} ${m://LastName} ${m://ExternalDataReference}
This questionnaire will be used to collect answers detailing the statistical approach that your research team has taken. Your answers will then be used to facilitate the round-robin peer review process. Please provide enough information for a naive empiricist to be able to give you valuable feedback. Remember, not all individuals involved in this project come from the same discipline, so some methods might be unfamiliar/have a different name to those in other areas. There are two sections: one that will be shared with other researchers, and one that we will use internally to get a good first idea about actual results. Only the analytic plans will be shared with the crowdsourcing groups to avoid bias.

Q20 Data Cleaning
transforms What transformations (if any) were applied to the variables. Please be specific.
exclusions Were any cases excluded, and why?

Q21 Statistical Modeling
technique: What is the name of the statistical technique that you employed?
tech_expl: 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.

tech_ref: What are some references for the statistical technique that you chose?
software: 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)
DV_dist: What distribution did you specify for the outcome variable of red cards?
cov_RQ1: What variables were included as covariates (or control variables) when testing research question 1: The relationship between player skin tone and red cards received?
cov_RQ2a: 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?
cov_RQ2b: 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?
cov_reason: What theoretical and/or statistical rationale was used for your choice of covariates included in the models?

Q24 Results
ES_unit: What unit is your effect size in?
ES_R1: What is the size of the effect for research question 1: The relationship between player skin tone and red cards received? Please specify the magnitude and direction of the effect size, along with the 95% confidence (or credible) interval in the following format: estimate [low interval, high interval]. Remember that this result will not be shared with other teams at this stage.
ES_R2a: What is the size of the effect for research question 2a: The relationship between referee country implicit skin-tone prejudice and red cards received by dark skin- tones players? Please specify the magnitude and direction of the effect size, along with the 95% confidence (or credible) interval in the following format: estimate [low interval, high interval]. Remember that this result will not be shared with other teams at this stage.
ES_R2b: What is the size of the effect for research question 2b: The relationship between referee country explicit skin-tone prejudice and red cards received by dark skin- tones players? Please specify the magnitude and direction of the effect size, along with the 95% confidence interval in the following format: estimate [low interval, high interval]. Remember that this result will not be shared with other teams at this stage.
alt_stats: What other steps/analyses did you run that are worth mentioning? Include effect sizes in a similar format as above if necessary.
script You may use the space below to paste the script you used to run the analyses. (Optional)
prior_RQ1: What is your current opinion regarding research question 1: How likely do you think it is that soccer referees tend to give more red cards to dark skinned players?
m Very Unlikely (1)
m Unlikely (2)

m Neither Likely nor Unlikely (3) m Likely (4)
m Very Likely (5)

prior_RQ2a: What is your current opinion regarding research question 2a: How likely is it that implicit cultural preferences for white over black skin tone in referees' country of origin are associated with biases in referees' decisions to give more red cards to dark skinned players?
m Very Unlikely (1)
m Unlikely (2)
m Neither Likely nor Unlikely (3) m Likely (4)
m Very Likely (5)
prior_RQ2b: What is your current opinion regarding research question 2b: How likely is it that explicit cultural preferences for white over black skin tone in referees' country of origin are associated with biases in referees' decisions to give more red cards to dark skinned players?
m Very Unlikely (1)
m Unlikely (2)
m Neither Likely nor Unlikely (3) m Likely (4)
m Very Likely (5)
comment: Please use this space for any additional comment you may have at this stage (this is for our information and will not displayed to other teams).

Q25: Please press the submit button only once you are sure that you would like to submit your respones and that no changes are needed at this stage. Deadline is midnight June 15th EST. Your name should be written here: ${m://FirstName} ${m://LastName}. If it is not, then you are in preview mode. In that case, please access the link through the personalized e-mail sent to you.

4. Feedback E-Mail

Dear <FirstName>,

We would like to thank you and your team for making this Crowdsourcing project happen! This has really been an interesting project for all of us so far. We have received your analytical approach and your feedback with thanks. Below I am sending you the feedback that your analytical approach has received from others as well as further instructions on how to proceed. We have assigned your team the identifier <Team>. This information is important for reviewing your feedback and later for submitting your results.

First, important feedback from us:
1. League vs. Referee Country. Many teams have used "League" as a control variable. We would like to emphasise that the dataset contains individuals' encounters with referees throughout their

professional careers. This means that they may have played in different leagues in different seasons. Also there have been the misconception that the dataset only covers 4 leagues. In fact, encounters from other leagues are included as the dyadic data is based on players' interactions. The fact that data originates from first league teams of major soccer leagues indicates that all players have high skill level. An alternative approach may be using the referee country of origin instead. We decided to make the referees' country of origin public. We decided to provide an updated dataset that includes the Alpha-3
country code of referees.

2. Red Cards. The question has been asked why the focus is on red cards and how red cards relate to yellow or yellow red cards. Yellow-cards are a caution, a warning vs. red cards result in the dismissal of a player as a response to a gross misconduct. We picked the indicator of a straight red card as there could have always been an alternative (a yellow card instead) and data is included on yellow cards being given to players whereas we do not know the number of fouls committed that yielded no card. If a player already has a yellow card, then a second yellow card offence results in a yellow-red card, which also means that the player is dismissed but in response to an incident that was not deemed severe. Even if a player already has a yellow card, he may be sent off with a straight red card, after a gross misconduct.

3. Skin-Tone. This is a technical note. We changed the scale of the
skin tone rating from a 1,2,3,4,5 scale to a 0,0.25,0.5,0.75,1 scale. This improves the ability to which we can compare results from different approaches.
The new dataset includes this update.

4. Dataset. Apart from the two changes mentioned (Referee Country) and Skin tone metric change, no other dataset changes have occurred.

If you have already a cleaned version of the data we recommend importing only the updated variables! Please tell us if you have trouble with this. The updated dataset is available in our project folder at the OSF website: https://osf.io/gvm2z/ Second, important feedback on your analytical approach. We have attached the document with a summary of all approaches and all feedback received. Please locate your team under the identifier <Team>. We would like to point out that you are by no means restricted to stick to your current analytical technique. Feel free to learn from others and modify your approach as you see fit. You will have until July 20th to refine your final analyses and submit your final results. We will be in touch towards the end of this week outlining the detailed procedure for submitting your final results and for registering your collaborators. Please do not hesitate to contact us should you have questions meanwhile.
Best regards,
Raphael, Dan, Eric and Brian

Supplement 4: Teams that changed their analytic approaches based on peer feedback

During the project, a number of teams changed their analytic approach as a result of peer feedback they received during the round-robin feedback round or thereafter. Table S4 provides details on the initial and revised approaches.

Team	Initial Approach	Final Approach
1	Ordinary least squares, logistic regression and nonlinear regression	Ordinary least squares with robust standard errors, logistic regression
2	Linear regression, logistic regression	Linear probability model, logistic regression
3	Multilevel Binomial Logistic Regression using bayesian inference	Multilevel Binomial Logistic Regression using Bayesian inference
4	Correlations and partial correlations	Spearman correlation
5	Mixed models (aka multilevel modeling, hierarchical linear models)	Generalized linear mixed models
6	Linear probability model	Linear Probability Model
7	Profile regression, a Dirichlet process Bayesian clustering	Dirichlet process Bayesian clustering
8	ANOVA, Linear Regression	Negative binomial regression with a log link analysis
9	Generalized linear mixed effects models (GLMM), with a logit link function	Generalized linear mixed effects models with a logit link function
10	Multilevel regression analyses	Multilevel regression and logistic regression
11	Multiple linear regression with total red cards as outcome variable	Multiple linear regression
12	Zero-inflated poisson (ZIP) regression	Zero-inflated Poisson regression
13	Glm with poisson distribution	Poisson Multi-level modeling
14	WLS (weighted least squares) estimation	Weighted least squares regression with referee fixed-effects and clustered SE
15	Hierarchical log-linear modeling	Hierarchical log-linear modeling
16	Hierarchical logistic regression	Hierarchical Poisson Regression
17	Bayesian probit regression	Bayesian logistic regression
18	Hierarchical Bayes model	Hierarchical Bayes model
19	Linear Regression	Cross-classified multilevel negative binomial model
20	A four-level multilevel negative-binomial model	Tobit regression
21	Tobit regression analysis	Mixed model logistic regression
22	OLS with dummy variables for each referee and player	Multilevel logistic regression
23	Mixed model logistic regression - both frequentist and Bayesian	Multilevel logistic binomial regression
24	Multilevel linear modelling	Three-level hierarchical generalized linear modeling with Poisson sampling
25	Hierarchical generalized linear model, with a log link function	Poisson regression
26	Three-level random effects model with Poisson estimation	Mixed effects logistic regression
27	Poisson Regression	Clustered robust binomial logistic regression
28	Generalized linear mixed effects modeling	Logistic regression
29	Bayesian hierarchical modeling	Generalized linear models for binary data

Table S4. Overview of teams' initial and final analytical approaches

Supplement 5: Final results

All final submissions from analysis teams can be found here: https://osf.io/qix4g/. A summary of methods used by each team and a one-sentence summary of the findings are presented below.

Summary of Methods

Team	Method
1	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 to compare with our OLS regressions.
2	Linear probability model, logistic regression
3	Multilevel Binomial Logistic Regression using bayesian inference.
4	Spearman correlation
5	Generalized linear mixed models
6	Linear Probability Model
7	Dirichlet process Bayesian clustering
8	Analysis of covariance (ANCOVA) for RQ1, negative binomial regression with a log link analysis for RQ2
9	Generalized linear mixed effects models (GLMM), with a logit link function (binary outcome)
10	Multilevel regression (and multilevel logistic regression)
11	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.
12	Zero-inflated Poisson regression
13	Poisson Multi-level modeling
14	In our main analysis, we use WLS (weighted least squares) estimation, including fixed effects for referee, player club and player position, and clustering the standard errors on the player level. Observations are weighted by the number of games per player/referee dyad. As robustness checks, we also use a logit estimation and alternative outcome measures (yellow-red cards (getting a red card after two yellow cards in the same game) and yellow cards).
15	Hierarchical log-linear modeling
16	Hierarchical Poisson Regression
17	Bayesian logistic regression
18	Hierarchical Bayes model
20	Cross-classified multilevel negative binomial model
21	Tobit regression
23	We used mixed model logistic regression, both frequentist and Bayesian

24	Multilevel logistic regression
25	We used a multilevel logistic binomial regression with the tuple (red cards, games) as the outcome.
26	Three-level hierarchical generalized linear modeling with Poisson sampling
27	Poisson regression
28	Mixed effects logistic regression
30	Clustered robust binomial logistic regression
31	Logistic regression
32	Generalized linear models for binary data (logistic regression) with multiple measurements reflecting correlated data

Summary of Results

Team	One Sentence Summary
1	Small amounts of referee bias due to skin tone is found in red cards and no bias is found in yellow cards, however, these results have a poor identification strategy with no exogenous variation and therefore are likely confounded by unobservables such as playing style. With good identification we show that there is no relationship between referee country implicit or explicit skin-tone prejudice and red cards received by dark skin-toned players?
2	Players with darker skin receive slightly more redcards than players of lighter skin, but this correlation should be viewed with skepticism and likely not given a causal interpretation.
3	Soccer referees are more likely to give red cards to dark skin toned players.
4	Results from the simple correlational approach suggest no meaningful effect of skin tone on the issuance of red cards.
5	Soccer players with darker skin are more likely to get a red card.
6	Using a linear probability model I do not find a statistically significant conditional correlation between skin tone and the issuance of red cards.
7	Darker skin players appear to have a higher relative risk of incurring in red cards, but we also found this for other subgroups of the players, in particular those who have been rated as 'neither dark nor light skin'.
8	A multi-method analysis indicates that soccer player skin tone matters for the number of red cards awarded by a referee, but this link is not augmented by the country biases of the soccer referee.
9	Dark skin toned players received 1.5 times more red cards than light skin toned players, an effect that could not be explained by the average racial biases of the referee's countries.
10	Professional soccer referees give more red cards (and fewer yellow cards) to darker-skinned players, but this behavior is not associated with prejudice levels in the referees' country-of-origin
11	There was statistical support for a unique bivariate relation between the skin tone color of a player and the player’s receiving red cards, but there was no support for either implicit or explicit biases of the referee’s country acting as a moderator

variable of the above mentioned relation.

12 There is a relationship ($p < .10$) between player skin color, implicit racial biases of a referees' home, and red card issuance in European football.

13 Our analysis supports the hypothesis that referees are more likely to give red cards to players with darker, versus lighter, skin, but this effect was not influenced by implicit or explicit measures of racial bias collected from the referees' home country.

14 Whether the club of the player is controlled for is important for the results of the first research question; with a control for club the skin color variable is not significantly related to the likelihood of receiving a red card, whereas without a control for club the skin color variable is significant in our "baseline model".

15 Although some group of players with the same skin tone do show lower or higher than expected proportions of red cards, we found no clearly interpretable evidence of bias.

16 Evidence from Poisson regression analysis indicates that darker skin tone soccer players receive more red cards relative to lighter skin tone players, but it does not appear that average prejudice levels in the home country of the referee play a role in this bias.

17 After removing seven outliers –0.3% of the complete dataset– a Bayesian logistic regression model no longer revealed any evidence for the assertion that soccer referees are more likely to give red cards to players with darker skin tone.

18 This study found that although it may be likely that the dark-skinned players receive more red cards than other players, the prejudices in referees' country of origin play no significant role.

20 Soccer players with darker skin-tones were more likely to receive red cards from referees, but this association was not moderated by implicit or explicit racial bias.

21 A Tobit regression method showed that skin color was weakly related to the number of red cards received, but this was not moderated by skin-tone prejudice as determined by referee country.

23 Darker skinned players are more likely to be sent off the soccer pitch, but – since this is not predicted by measures of implicit or explicit bias associated with the country of the referee - the locus of this bias remains unclear.

24 Dark skin toned players were more likely to get a red card, but the effect of skin tone did not seem to be dependent on explicit or implicit attitudes.

25 Results show that darker skinned players are more likely to receive a red card, and referees from countries with higher mean implicit association test score are more likely to give red cards; however, they do not seem to be particularly more likely to punish darker toned players than other referees, on average.

26 Soccer referees are more likely to give red cards to darker skin toned players.

27 We found an incidence rate ratio of 8.24, suggesting that players whose skin tone was rated darkest were more than 8 times more likely to receive red cards than those whose skin tone was rated lightest, however this finding was not significant and no significant impact of implicit or explicit bias in the country of origin of referee was found.

28 A mixed effects logistic regression analysis with crossed random effects for

referees and players revealed that soccer players with darker as opposed to lighter skin tones receive more red cards ($OR_{lightest,darkest}$ = 1.382 [1.120, 1.705]) regardless of explicit or implicit racial prejudice in the referees' home countries.

30 Using a clustered robust binomial regression adjusted for several potentially confounding variables, we find that dark skinned players receive more red cards, but that this is not related to the average levels of implicit or explicit skin bias in the referee's home country.

31 Our logistic regression results showed that the players' skin colors, and the explicit and implicit attitudes held by the referee's country of origin do not influence the distribution of red cards.

32 The odds of a dark skin toned player (scale=1) receiving a red card are 1.39 times higher than the odds for a light skin toned player (scale=0) receiving a red card. The 95% confidence interval of the odd ratio is (1.10, 1.75).

Supplement 6: Additional analyses of research team expertise and statistical model choice

Further analyses examined the effects of research teams' quantitative expertise on choices of statistical models. With regard to the choice modeling distribution, 7 of 9 teams who had comparatively high levels of expertise chose a logistic model, and 5 of these 7 found a statistically significant result (median OR = 1.38, MAD = .10). Of those who had comparatively lower expertise, 8 of 19 used a logistic model and 6 of these 8 found a statistically significant result (median OR = 1.33, MAD = .08). All 5 teams who chose a Poisson model were in the comparatively lower expertise group, with 4 of these 5 teams detecting a statistically significant effect (median OR = 1.40, MAD = .12). Additionally, all 6 teams who chose a linear model were in the comparatively lower expertise group, with 3 of these 6 teams detecting a statistically significant effect (median OR = 1.21, MAD = .05).

With regard to handling the non-independent nature of the dataset, 6 of 9 teams who had comparatively high levels of expertise used a variance component for players and/or referees, and 4 of these 6 found a statistically significant result (median OR = 1.35, MAD = .16). Of those who had comparatively lower expertise, 9 of 19 used a variance component for players and/or referees and 8 of these 9 found a statistically significant result (median OR = 1.32, MAD = .09). More teams with comparatively lower rankings on expertise chose to use clustered standard errors (7/19 teams, versus 1/9 teams comparatively higher in expertise). Three of 7 relatively less expert teams who used clustered standard errors detected a statistically significant result (median OR = 1.28, MAD = .10).

Supplement 7: Research Questions 2a and 2b

This project additionally examined whether national level preferences for light vs. dark skin predict the red card decisions of referees from those countries. Research question 2a examined whether national level implicit preferences for light vs. dark skin predict referee card decisions, which research question 2b did the same with explicit preferences.

For the country of each referee, we included average scores of implicit and explicit preferences for light vs. dark skin tone that had been gathered in independent research by Project Implicit (Nosek et al., 2007; Nosek, Banaji, & Greenwald, 2002). Implicit preference scores for each referee country had been calculated using a skin tone Implicit Association Test (IAT) (Greenwald, McGhee, & Schwartz, 1998), a speeded response task that assesses strength of associations. Higher scores on the IAT reflect a stronger automatic association between dark skin, relative to light skin, and negative valence. Explicit preference scores for each referee country were calculated using a feeling thermometer task, with higher values corresponding to greater self-reported feelings of positivity toward light skin tone versus dark skin tone. Both these national-level measures were created by aggregating data from many online users from referees' countries taking these tests on Project Implicit (https://implicit.harvard.edu/; see alsoMarini et al., 2013).

At the outset of the project, analysts expressed serious concerns as to the suitability of the available data to test these hypotheses. In an initial survey, 75% and 72% of respondents were unconfident to somewhat unconfident regarding how appropriate the dataset was for answering either research question 2a or 2b, respectively. In contrast, only 32% of respondents felt the same way regarding the primary research question (whether an association exists between players' skin tone and referee red card decisions). Teams commented one reason they felt this way is the lack of variability in the country-level implicit/explicit measures, as well as sampling issues regarding the measures from a particular country. For example, it is difficult to determine how well the bias from a non-random sample of drastically different sample sizes for each country might map on to how biased a given referee might be. Because of this, we chose to not include the aggregated results for these research questions in the main text.

Results for both research questions 2a and 2b from the majority of teams yielded extremely wide confidence intervals. When submitting their final report, only 3 team leaders found it likely that implicit cultural preferences for light over dark skin tone in referees' country of origin are associated with biases in referees' decisions to give more red cards to dark skinned players. In contrast, 14 team leaders found this to be unlikely and 12 neither likely nor unlikely. Similarly, only 1 team leader found it likely that explicit cultural preferences for light over dark skin tone had this same association, whereas 18 team leaders found this to be unlikely and 10 neither likely or unlikely. In total, all but one team found no significant evidence for an effect in this sample. See Fig. S7 below for team's beliefs regarding the effects for research question 2a

and 2b.

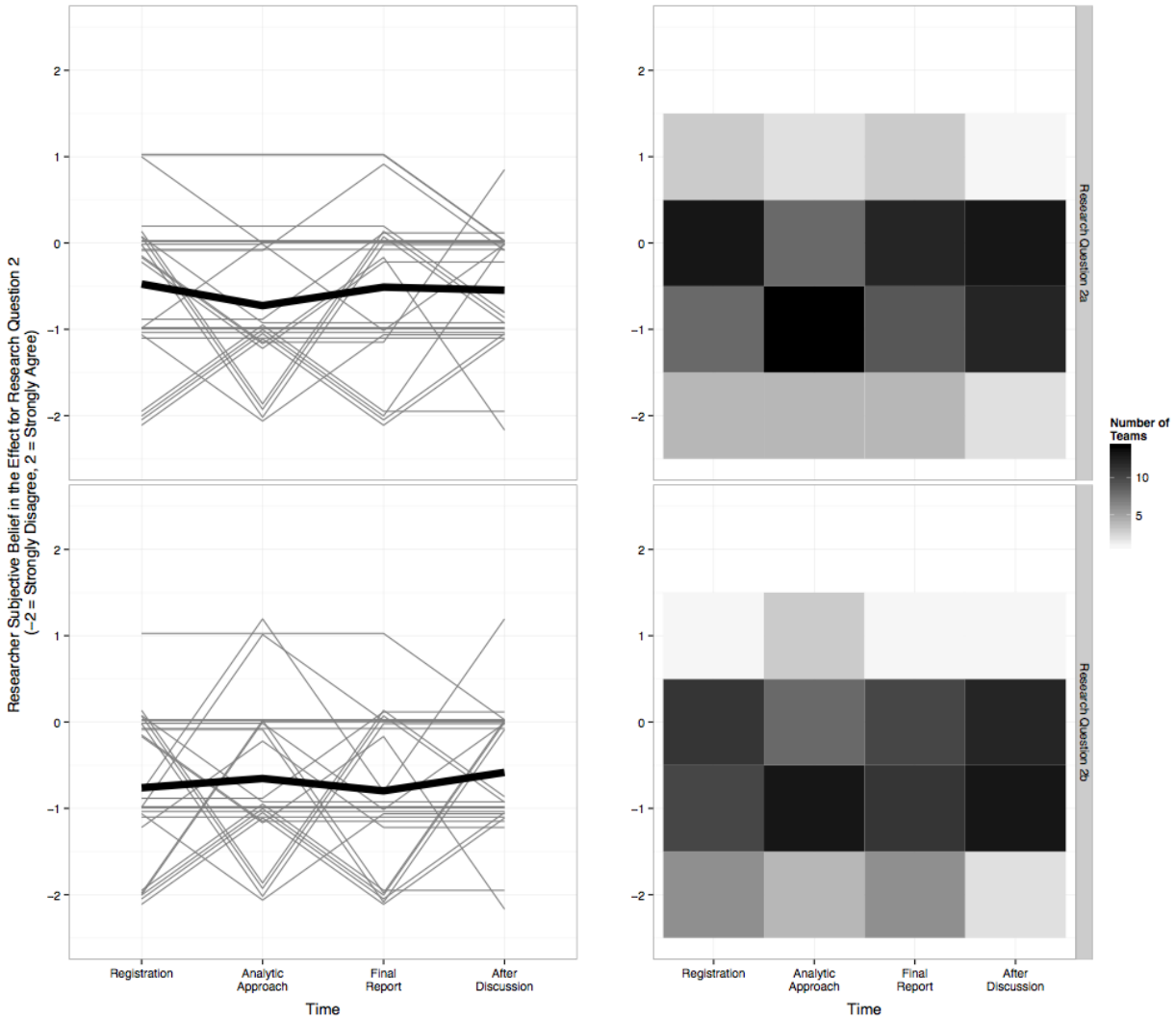

Fig. S7. The top panels reflects team leaders beliefs regarding research question 2a (whether national level implicit preferences for light vs. dark skin predict referee red card decisions). The bottom two panels reflect team leader beliefs for research question 2b (whether national level explicit skin tone preferences predict red card decisions). The plots on the left show belief trajectories, where each light gray line represents a single team leader's belief trajectory throughout the project and the black trajectory represents the mean value at each time point. The plots on the right represent the consensus (or lack thereof) by plotting the number of team leaders endorsing a particular response at each time.

References for S7

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. (1998). Measuring individual differences in implicit cognition: the implicit association test. *Journal of Personality and Social Psychology*, *74*(6), 1464–1480.

Marini, M., Sriram, N., Schnabel, K., Maliszewski, N., Devos, T., Ekehammar, B., … Nosek, B. A. (2013). Overweight people have low levels of implicit weight bias, but overweight nations have high levels of implicit weight bias. *PloS One*, *8*(12), e83543.

Nosek, B. A., Banaji, M., & Greenwald, A. G. (2002). Harvesting implicit group attitudes and beliefs from a demonstration web site. *Group Dynamics: Theory, Research, and Practice: The Official Journal of Division 49, Group Psychology and Group Psychotherapy of the American Psychological Association*, *6*(1), 101–115.

Nosek, B. A., Smyth, F. L., Hansen, J. J., Devos, T., Lindner, N. M., Ranganath, K. A., … Banaji, M. R. (2007). Pervasiveness and correlates of implicit attitudes and stereotypes. *European Review of Social Psychology*, *18*(1), 36–88.

Supplement 8: Author Contribution Forms from Analysis Teams

Team	Name	Contribution
1	Nolan G. Pope	Analysis and writing
1	Bryson Pope	Analysis and writing
2	Garret Christensen	Details unavailable
3	Erikson Kaszubowski	Details unavailable
4	Christopher R. Madan	Analysis, interpretation, writing
5	Johannes Ullrich	Coordinated project, planned analyses, conducted analyses, wrote report
5	Elmar Schlüter	Coordinated project, planned analyses, discussed report
5	Christoph Spörlein	Planned analyses, conducted analyses, discussed report
5	Andreas Glenz	Planned analyses, disaggregated data, conducted analyses, checked R script
6	Jonathan Kalodimos	All
7	Silvia Liverani	Details unavailable
8	S. Amy Sommer	Data Analysis, Writing
8	Deanna M. Kennedy	Data Analysis, Writing
9	Felix D. Schönbrodt	Data Analysis, Writing
9	Moritz Heene	Data Analysis, Writing
10	Daniel Molden	Helped design analyses; conducted analyses; wrote the report
10	Maureen Craig	Helped design analyses; conducted analyses
10	Ryan Lei	Helped design analyses
10	Monica Gamez-Djokic	Helped design analyses
11	Jason M. Prenoveau	Analyses and write-up
11	Martin F. Sherman	Analyses and write-up
12	Eli Awtrey	All
13	Alicia J. Mohr	Analysis plan, analysis, writing report
13	Thomas A. Lindsay	Analysis plan, writing report
14	Anna Sandberg	Analysis plan, analysis, writing report
14	Evelina Bonnier	Analysis plan, analysis, writing report
14	Karin Hederos	Analysis plan, analysis, writing report
14	Magnus Johannesson	Analysis plan, writing report
15	Michelangelo Vianello	Analyzed data; Wrote report
15	Egidio Robusto	Analyzed data
15	Pasquale Anselmi	Analyzed data
15	Luca Stefanutti	Analyzed data
15	Anna Dalla Rosa	Analyzed data
16	Russ Clay	All
17	Eric-Jan Wagenmakers	Conceptualizing the analyses, writing
17	Richard D. Morey	Conceptualizing the analyses, conducting the analyses, writing
18	Maciej Witkowiak	All
20	Felix Cheung	Collection of data on players' position; Data analysis; Interpretation of the results; Draft the final results
20	Kent Hui	Collection of data on players' position; Interpretation of the results; Provide feedback on written drafts
21	Laetitia B. Mulder	Coordination, feedback on other teams, and final writing
21	Lammertjan	Performing (and informing on) Tobit regressions, feedback on the other teams
21	Eric Molleman	Initial analyses
21	Bernard A. Nijstad	Initial analyses and deciding on final analyses
21	Floor Rink	Advise, input on analyses and writing, feedback on other teams, and fellow-coordination
21	Susanne Tauber	Advise, feedback on other teams
23	Tom Stafford	Analysis coordination, analysis; writing up
23	Mathew H. Evans	Visualisation; writing up
23	Tim J. Heaton	Analysis, frequentist models; writing up
23	Colin Bannard	Aanalysis, Bayesian models
24	Štěpán Bahník	Details unavailable
25	Seth Spain	Analysis plan, data preparation, analysis, report writing and editing
25	Kristin Sotak	Analysis planning, analysis, report write-up
26	Feng Bai	Details unavailable
26	Hadiya Roderique	Details unavailable
27	Shauna Gordon-McKeon	Design and execution of analysis plan; write up.
28	Frederik Aust	Data analysis, reporting of results
28	Fabia Högden	Data analysis, reporting of results
30	Rickard Carlsson	All
31	Sangsuk Yoon	Data analysis, write up
31	Nathan Fong	Data analysis
32	Ismael Flores Cervantes	All

Supplement 9: Limitations of the dataset

A number of significant limitations of the dataset were discussed during the project, and are worth further elaborating on. Given the correlational nature of the available field data, the present research cannot identify causal relationships between variables. Most teams observed a significant relationship between player skin tone and referee red card decisions, but this correlation could be driven by referee biases, player behavior (e.g., due to national differences in playing styles), or unmeasured third variables.

Another major limitation is that data on explicit and implicit skin tone preferences (the focus of research questions 2a and 2b) were only available for referees' country of origin, not for the individual referees themselves. Referees may or may not have skin tone preferences similar to those of the average person in their home country. This could be one reason why our analysis teams converged on the conclusion that skin tone preferences did not predict referee decisions, and that the dataset was not adequate to answer the question effectively (see S7). Another explanation, of course, is that neither explicit nor implicit attitudes exhibit significant predictive validity in this particular field context. To address these issues, it will be productive to directly measure the social attitudes of sports officials and examine whether these predict their judgments of players.

More generally, to investigate the research questions more effectively, access to more detailed and fine-grained data would be ideal. The amount of time a player was on the pitch during the game, details of all other players playing that same match, whether the game was an international game or league game and if the latter in which league the game was played, as well as the importance of the particular game were all mentioned by analysts as information they would have liked to have included but that was not available.

Supplement 10: Club and league as covariates

During the round-robin feedback stage, it became clear that some variables were not interpreted by researchers in the same way. Players' club and leagues was a static variable in the dataset, gathered from players' profile page at the time of data collection. Whereas weight and height for players are relatively static, club and league information is not actually static across time. Players may switch clubs and leagues between seasons. Consequently while the project coordinators saw those two variables as identifying variables, the lack of labeling as such meant that some researchers worked with club and league information in their first analyses. As the information for each player-referee-dyad referred to all games played in individuals' professional career, single club and league information for each player did not necessarily reflect the state of the world at the time of each particular game. This information was clarified in an e-mail to project members. However, teams were not obliged to change their analytical approach based on the round-robin feedback.

To examine whether using league and club as covariates affected final effect size estimates, we asked those ten teams who had used the league and club variables in their analyses to reconduct their analyses without these variables. The removal of the two covariates corresponded to a slight increase in effect size (Median OR = 1.25, MAD = 0.12 to Median OR = 1.32, MAD = 0.07). We offered teams the choice of whether to include or exclude these covariates in their final models. The overviews in the tables and figures in the main text reflect teams' final model choice.

Supplement 11: IPython notebook visualisation of the dataset

Team 23 (Tom Stafford, Mathew H. Evans, Tim Heaton, Colin Bannard) created a walkthrough of some exploration and visualisation of the data steps taken in support of their analysis. This illustrates some of the process Team 23 went through as part of this project. This is in an IPython notebook which can be viewed statically here:

http://nbviewer.ipython.org/github/mathewzilla/redcard/blob/master/Crowdstorming_visualisation.ipynb

The notebook can also be downloaded for interactive use on a local machine.

Supplement 12: Survey of familiarity with each analytic approach

The subsequent pages feature the complete survey assessing each researchers' level of familiarity with each analytic approach used. The sample of scientists for the survey consisted of the researchers participating in the crowdsourced project.

Please indicate how familiar you are with each of the following analytical techniques.	Very unfamiliar (1)	Rather unfamiliar (2)	Somewhat familiar (3)	Familiar (4)	Very familiar (5)
Ordinary least squares with robust standard errors, logistic regression (1)	○	○	○	○	○
Linear probability model, logistic regression (2)	○	○	○	○	○
Multilevel Binomial Logistic Regression using Bayesian inference (3)	○	○	○	○	○
Spearman correlation (4)	○	○	○	○	○
Generalized linear mixed models (5)	○	○	○	○	○
Linear Probability Model (6)	○	○	○	○	○
Dirichlet process Bayesian clustering (7)	○	○	○	○	○
Negative binomial regression with a log link analysis (8)	○	○	○	○	○
Generalized linear mixed effects models with a logit link function(9)	○	○	○	○	○
Multilevel regression and logistic regression (10)	○	○	○	○	○
Multiple linear regression (11)	○	○	○	○	○
Zero-inflated Poisson regression (12)	○	○	○	○	○
Poisson Multi-level modeling (13)	○	○	○	○	○
Weighted least squares regression with referee fixed-effects and clustered SE (14)	○	○	○	○	○
Hierarchical log-linear modeling (15)	○	○	○	○	○
Hierarchical Poisson Regression (16)	○	○	○	○	○
Bayesian logistic regression (17)	○	○	○	○	○

Hierarchical Bayes model (18)	○	○	○	○	○
Cross-classified multilevel negative binomial model(19)	○	○	○	○	○
Tobit regression (20)	○	○	○	○	○
Mixed model logistic regression (21)	○	○	○	○	○
Multilevel logistic regression (22)	○	○	○	○	○
Multilevel logistic binomial regression (23)	○	○	○	○	○
Three-level hierarchical generalized linear modeling with Poisson sampling (24)	○	○	○	○	○
Poisson regression (25)	○	○	○	○	○
Mixed effects logistic regression (26)	○	○	○	○	○
Clustered robust binomial logistic regression (27)	○	○	○	○	○
Logistic regression (28)	○	○	○	○	○
Generalized linear models for binary data (29)	○	○	○	○	○

Covariates: I'm willing to review additional aspects of the dataset (i.e. the validity of particular covariates). This is a great way to help, particularly if you are not familiar with analytical techniques.

- ○ Yes (1)
- ○ No (2)
- ○ If at all needed (3)

Supplement 13: Peer review survey of final analytical choices for potential issues

Thank you very much for reviewing the final report assigned to you. Below you will find a series of guiding points to help your assessment. These points are based on the feedback given to the initial analytical approaches. Please carefully examine the final report. We would like to know to what extent each point is (still) an issue in the described approach, or whether it has been (fully) addressed. If you need verifying information from the authors, please get in touch. Please note that the validity of the inclusion of covariates will be assessed separately. You can re-open the questionnaire. This review is for Team ${e://Field/Team}.Click here to locate this team's report in a new window: https://osf.io/j5v8f/

Q1 Dependent Variable Point 1. In the dataset the dependent variable (red cards given) needs to take into account the number of games played. Examples of how this issue could be resolved: It has been suggested that a remedy is to transform the data (for instance so that each line represents a single referee player interaction). Alternatively, it has been suggested that 'Games' should be used as an offset in a regression (rather than a predictor) so that observations are weighted depending on the number of games in each player/referee dyad.The approach from Team ${e://Field/Team} DOES NOT adequately account for the number of games played.

- ○ Strongly Disagree (1)
- ○ Disagree (2)
- ○ Neither Agree nor Disagree (3)
- ○ Agree (4)
- ○ Strongly Agree (5)

Q2 Point 2. The value of red cards in the dataset is either 0, 1 or 2 and there are many cases in which no red card was given and two red cards was very few. Example: The dependent variable cannot be assumed to be linear (assuming an interval-scale). The approach from Team X assumes an interval-scale (linear model).

- ○ Strongly Disagree (1)
- ○ Disagree (2)
- ○ Neither Agree nor Disagree (3)
- ○ Agree (4)
- ○ Strongly Agree (5)

Q3 Point 3. Red cards are dependent on the number of games played. If red cards per game was specified as a proportion, this represents a ratio and a linear model would also not be appropriate. Further, transforming red cards into a proportion has limitations in that it equates getting 0 red cards in only 1 or 2 games with a referee and getting 0 red cards in 20 games with the

referee. The approach from Team ${e://Field/Team} specifies 'red cards per game' as a proportion.

- ○ Strongly Disagree (1)
- ○ Disagree (2)
- ○ Neither Agree nor Disagree (3)
- ○ Agree (4)
- ○ Strongly Agree (5)

Q4 Point 4. Many players received 0 red cards from a referee. Therefore the dependent variable often takes the value of 0. Was a model chosen that addresses this issue? For example: It has been suggested that a negative binominal regression is more appropriate than a Poisson regression, because of the high number of zeros in the distribution (and the associated low mean and high variance in this variable). The approach from Team ${e://Field/Team} DOES NOT adequately take into account that the dependent variable often takes the value of 0.

- ○ Strongly Disagree (1)
- ○ Disagree (2)
- ○ Neither Agree nor Disagree (3)
- ○ Agree (4)
- ○ Strongly Agree (5)

Q5 Point 5. Both YellowRed and Red result in the send-off of players. Yet YellowRed and RedCards are qualitatively different: the YellowRed is a second yellow card offense, given after a previous yellow card had been shown. Yellow cards are typically given for less serious fouls than pure red cards. There is no consensus whether pooling YellowRed and Red cards is appropriate or not. Nevertheless we want to record this distinction. The approach from Team ${e://Field/Team} predicts NOT ONLY red cards but also yellow-red and/or yellow cards.

- ○ Strongly Disagree (1)
- ○ Disagree (2)
- ○ Neither Agree nor Disagree (3)
- ○ Agree (4)
- ○ Strongly Agree (5)

Q6 Model: Point 1: The dataset is based on repeated observations of referees and players. Many regression analyses such as OLS – classical linear regression models and also standard logistic regression requires each observation to be independent. It is an issue if the analytical technique treats the data as independent, instead of nested, multi-level, and thus accounting for repeated observations of referees and players. The approach from Team ${e://Field/Team} DOES NOT adequately take into account that observations are non-independent.

- ○ Strongly Disagree (1)
- ○ Disagree (2)
- ○ Neither Agree nor Disagree (3)
- ○ Agree (4)
- ○ Strongly Agree (5)

Q7 Exclusions & Missing Data: Point 1. Have cases been unnecessarily been excluded, potentially leading to a loss in information? For instance, dichotomizing skintone (and excluding "neutrals"); excluding cases where the raters disagree; excluding dyads or players for whom no red card was given. The approach from Team ${e://Field/Team} unnecessarily excludes a substantial number of cases.

- ○ Strongly Disagree (1)
- ○ Disagree (2)
- ○ Neither Agree nor Disagree (3)
- ○ Agree (4)
- ○ Strongly Agree (5)

Q8 You can use this space to describe whether the particular approach includes any issue that has not been mentioned in the list above. [Free response text box].

Q9 This additional issue seriously affects the validity of this approach

- ○ Strongly Disagree (1)
- ○ Disagree (2)
- ○ Neither Agree nor Disagree (3)
- ○ Agree (4)
- ○ Strongly Agree (5)

Q10 Overall, how convinced are you that the presented approach successfully addressed most concerns regarding the analysis?

- Very unconvinced (1)
- Rather unconvinced (2)
- Neither convinced nor unconvinced (3)
- Rather convinced (4)
- Very convinced (5)

Supplement 14: Exploratory analyses in search of converging results

We also carried out further coding and exploratory analyses to see if any subcategory of analytic approaches could be identified for which there was greater convergence in results across teams.

Of particular interest was whether results might cluster by differences in use of covariates. From the pool of researchers participating in this crowdsourced project we recruited a sub-team of those interested in discussing the advantages and disadvantages of including each covariate. This was done via e-mail (see https://osf.io/g3k8h/). The purpose of this discussion was to see whether we could arrive at a conclusion about which covariates warrant inclusion into the models and which ones should not be used. From the arguments it was concluded that teams pursued different motivations regarding the treatment of covariates and that there were three distinguishable approaches. A first group of teams attempted to use as few covariates as possible so that any obtained effect would relate to observable outcomes (across leagues, player sizes, positions or other covariates). A second group of teams tried to include as much information as available into the models, albeit at the cost of increasing noise. A third group of teams tried a balanced approach between including many and few covariates.

There thus appeared to be different philosophies between teams regarding the most appropriate strategy to best model the effect and answer the research question. Importantly, the research question did not specify clearly whether any effect was to be modeled with or largely without covariates. We therefore aimed to differentiate results based on teams' strategies, as observed by the number of covariates included by teams, and take into account peer ratings of confidence in each approach.

Supplementary Table S14 shows the results grouped into three categories, 0-1 covariates, 2-3 covariates, and more than 3 covariates. Results are ordered so that in each category, the approach with the highest confidence ratings from peers is ranked on top. This overview shows that the top-ranked approaches in each category are quite similar in terms of their OR (an average odds ratio of OR 1.40 [95% CI: 1.15, 1.71], as evidenced by a low standard deviation (SD = 0.02). Thus, within the sets of analyses that included relatively few, a moderate number, or a high number of covariates, higher quality analyses tended to find an OR of around 1.40. Future research should examine whether this exploratory evidence of convergence among high quality-analyses within each category of covariate use can be replicated in a confirmatory analyses with a larger sample size.

Covariate Group	Team	Covariates	Analytical Issues	Confidence	OR	Min	Max
0-1	**20**	**1**	**1.06**	**5.00**	**1.40**	**1.15**	**1.71**
0-1	13	1	1.75	4.33	1.41	1.13	1.75
0-1	7	0	1.94	3.50	1.71	1.70	1.72
0-1	5	0	2.06	4.00	1.38	1.10	1.75
0-1	27	1	2.44	2.00	2.93	0.11	78.66
0-1	8	0	2.58	3.00	1.39	1.17	1.65
0-1	32	1	2.67	2.00	1.39	1.10	1.75
0-1	15	1	3.00	2.33	1.02	1.00	1.03
2-3	**28**	**2**	**1.54**	**5.00**	**1.38**	**1.12**	**1.71**
2-3	3	2	1.54	4.67	1.31	1.09	1.57
2-3	17	2	1.63	4.00	0.96	0.77	1.18
2-3	23	2	1.63	4.33	1.31	1.10	1.56
2-3	18	2	1.69	3.00	1.10	0.98	1.27
2-3	16	2	1.75	4.33	1.32	1.06	1.63
2-3	24	3	1.94	5.00	1.38	1.11	1.72
2-3	9	2	2.00	4.00	1.48	1.20	1.84
2-3	30	3	2.19	3.00	1.28	1.04	1.57
2-3	10	3	2.31	3.00	1.03	1.01	1.05
2-3	12	2	2.44	1.50	0.89	0.49	1.60
2-3	4	3	3.08	1.67	1.21	1.20	1.21
>3	**25**	**4**	**1.83**	**4.67**	**1.42**	**1.19**	**1.71**
>3	11	4	1.94	3.00	1.25	1.05	1.49
>3	31	6	2.00	3.00	1.12	0.88	1.43
>3	26	6	2.21	4.00	1.30	1.08	1.56
>3	2	6	2.44	3.50	1.34	1.10	1.63
>3	21	4	2.58	3.33	2.88	1.03	11.47
>3	14	6	2.75	3.67	1.21	0.97	1.46
>3	1	7	2.75	3.00	1.18	0.95	1.41
>3	6	6	3.54	2.33	1.28	0.77	2.13

Table S14 – Teams split by number of covariates used in the final model, assessment of analytical issues and peer ratings of confidence in each analysis