

## Many analysts, one dataset: Making transparent how variations in analytical choices affect results

### Authors

Silberzahn R.<sup>6</sup>, Uhlmann E. L.<sup>8</sup>, Martin D. P.<sup>35</sup>, Anselmi P.<sup>32</sup>, Aust F.<sup>26</sup>, Awtrey E.<sup>37</sup>, Bahník Š.<sup>39</sup>, Bai F.<sup>25</sup>, Bannard C.<sup>29</sup>, Bonnier E.<sup>16</sup>, Carlsson R.<sup>9</sup>, Cheung F.<sup>13</sup>, Christensen G.<sup>20</sup>, Clay R.<sup>4</sup>, Craig M. A.<sup>15</sup>, Dalla Rosa A.<sup>32</sup>, Dam L.<sup>28</sup>, Evans M. H.<sup>30</sup>, Flores Cervantes I.<sup>41</sup>, Fong N.<sup>18</sup>, Gamez-Djokic M.<sup>14</sup>, Glenz A.<sup>40</sup>, Gordon-McKeon S.<sup>7</sup>, Heaton T. J.<sup>33</sup>, Hederos K.<sup>17</sup>, Heene M.<sup>11</sup>, Hofelich Mohr A. J.<sup>31</sup>, Högden F.<sup>26</sup>, Hui K.<sup>12</sup>, Johannesson M.<sup>16</sup>, Kalodimos J.<sup>7</sup>, Kaszubowski E.<sup>21</sup>, Kennedy D.M.<sup>38</sup>, Lei R.<sup>15</sup>, Lindsay T. A.<sup>31</sup>, Liverani S.<sup>3</sup>, Madan C. R.<sup>22</sup>, Molden D.<sup>14</sup>, Molleman E.<sup>28</sup>, Morey R. D.<sup>28</sup>, Mulder L. B.<sup>28</sup>, Nijstad B. R.<sup>28</sup>, Pope N. G.<sup>19</sup>, Pope B.<sup>2</sup>, Prenoveau J. M.<sup>10</sup>, Rink F.<sup>28</sup>, Robusto E.<sup>32</sup>, Roderique H.<sup>34</sup>, Sandberg A.<sup>17</sup>, Schlüter E.<sup>27</sup>, Schönbrodt F. D.<sup>11</sup>, Sherman M. F.<sup>10</sup>, Sommer S.A.<sup>5</sup>, Sotak K.<sup>1</sup>, Spain S.<sup>1</sup>, Spörlein C.<sup>24</sup>, Stafford T.<sup>33</sup>, Stefanutti L.<sup>32</sup>, Tauber S.<sup>28</sup>, Ullrich J.<sup>40</sup>, Vianello M.<sup>32</sup>, Wagenmakers E.-J.<sup>23</sup>, Witkowiak M.<sup>42</sup>, Yoon S.<sup>18</sup>, & Nosek B. A.<sup>35,36</sup>

**Contact Authors:** [r.silberzahn@gmail.com](mailto:r.silberzahn@gmail.com), [eric.luis.uhlmann@gmail.com](mailto:eric.luis.uhlmann@gmail.com), [dpmartin42@gmail.com](mailto:dpmartin42@gmail.com), [nosek@virginia.edu](mailto:nosek@virginia.edu),

### Affiliations

<sup>1</sup>SUNY Oswego; <sup>2</sup>Brigham Young University; <sup>3</sup>Brunel University London, MRC Biostatistics Unit, Cambridge and Imperial College London; <sup>4</sup>City University of New York; <sup>5</sup>United States Military Academy at West Point; <sup>6</sup>University of Sussex; <sup>7</sup>Oregon State University; <sup>8</sup>INSEAD; <sup>9</sup>Linnaeus University; <sup>10</sup>Loyola University Maryland; <sup>11</sup>Ludwig-Maximilians-Universität München; <sup>12</sup>Michigan State University; <sup>13</sup>University of Hong Kong; <sup>14</sup>Northwestern University; <sup>15</sup>New York University; <sup>16</sup>Stockholm School of Economics; <sup>17</sup>Stockholm University; <sup>18</sup>Temple University; <sup>19</sup>The University of Chicago; <sup>20</sup>UC Berkeley; <sup>21</sup>Universidade Federal de Santa Catarina; <sup>22</sup>University of Alberta & University of Nottingham; <sup>23</sup>University of Amsterdam; <sup>24</sup>University of Bamberg; <sup>25</sup>Hong Kong Polytechnic University; <sup>26</sup>University of Cologne; <sup>27</sup>University of Giessen; <sup>28</sup>University of Groningen; <sup>29</sup>University of Liverpool; <sup>30</sup>University of Manchester; <sup>31</sup>University of Minnesota; <sup>32</sup>University of Padua; <sup>33</sup>University of Sheffield; <sup>34</sup>University of Toronto; <sup>35</sup>University of Virginia; <sup>36</sup>Center for Open Science; <sup>37</sup>University of Washington; <sup>38</sup>University of Washington Bothell; <sup>39</sup>University of Economics, Prague; <sup>40</sup>University of Zurich; <sup>41</sup>Westat; <sup>42</sup>Unaffiliated;

**Abstract**

Twenty-nine teams involving 61 analysts used the same dataset to address the same research question: whether soccer referees are more likely to give red cards to dark skin toned players than light skin toned players. Analytic approaches varied widely across teams, and estimated effect sizes ranged from 0.89 to 2.93 in odds ratio units, with a median of 1.31. Twenty teams (69%) found a statistically significant positive effect and nine teams (31%) observed a non-significant relationship. Overall 29 different analyses used 21 unique combinations of covariates. We found that neither analysts' prior beliefs about the effect, nor their level of expertise, nor peer-reviewed quality of analysis readily explained variation in analysis outcomes. This suggests that significant variation in the results of analyses of complex data may be difficult to avoid, even by experts with honest intentions. Crowdsourcing data analysis, a strategy by which numerous research teams are recruited to simultaneously investigate the same research question, makes transparent how defensible, yet subjective analytic choices influence research results.

Keywords: crowdsourcing science, data analysis, scientific transparency

**Many analysts, one dataset:****Making transparent how variations in analytical choices affect results**

In the scientific process, creativity is mostly associated with the generation of testable hypotheses and the development of suitable research designs. Data analysis, on the other hand, is sometimes seen as the mechanical, unimaginative process of clarifying the result. Despite methodologists' remonstrations (Bakker, van Dijk, & Wicherts, 2012; Gelman & Loken, 2014; Simmons, Nelson, & Simonsohn, 2011), it is easy to overlook the fact that results may depend on the chosen analytical strategy, which itself is imbued with theory, assumptions, and choice points. In many cases, there are many reasonable (and many unreasonable) approaches to evaluating data that bear on a research question (Carp, 2012a, 2012b; Gelman & Loken, 2014; Wagenmakers, Wetzels, Borsboom, van der Maas, & Kievit, 2012).

This may be understood conceptually, but there is little appreciation for its implications in practice. In some cases, authors use a particular analytic strategy because it is the one they know how to use, rather than there being a specific rationale. Peer reviewers may comment and suggest improvements to a chosen analysis strategy, but rarely do those comments emerge from working with the actual dataset (Sakaluk, Williams, & Biernat, 2014). Similarly, it is not uncommon for peer reviewers to take the authors' analysis strategy for granted and comment exclusively on other aspects of the manuscript. More importantly, once published, reanalysis or challenges of analytic strategies emerge slowly and occur rarely (Ebrahim et al., 2014; Krumholz & Peterson, 2014; McCullough, McGeary, & Harrison, 2006), in part due to the low frequency with which data are available for re-analysis (Wicherts, Borsboom, Kats, & Molenaar, 2006). The reported results and implications drive the impact of published articles; the analysis strategy is pushed to the background.

But what if the methodologists are correct? What if scientific results are highly contingent on subjective decisions at the analysis stage? Then, the process of certifying a particular result based on an idiosyncratic analysis strategy might be fraught with unrecognized uncertainty (Gelman & Loken, 2014) and research findings less trustworthy than they at first appear (Cumming, 2014). Had the authors made different assumptions, an entirely different result might have been observed (Babtie, Kirk, & Stumpf, 2014). The present article reports an investigation of the impact of analysis decisions on research results as 29 teams analyze the same dataset to evaluate the same research question. This investigation shows how researchers vary in their analytical approaches and makes transparent how results vary based on analytical choices. We aim to address the current lack of knowledge about just how much diversity in analytic choice exists with regard to the same data, and whether such diversity results in different conclusions.

### **Crowdsourcing data analysis: Skin-tone and red cards in soccer**

The primary research question tested in this crowdsourced project was whether soccer players with dark skin tone are more likely than those with light skin tone to receive red cards from referees. The decision to give a player a red card results in the ejection of the player from the game and has severe consequences as it obliges his team to continue with one less player for the remainder of the match. Red cards are given for aggressive behavior such as a violent tackle, a foul intended to deny an opponent a clear goal scoring opportunity, hitting or spitting on an opposing player, or threatening and abusive language. However, despite a standard set of rules and guidelines for both players and match officials, referee decisions are often fraught with ambiguity (e.g., was that an intentional foul or was the player only going for the ball?). It is

inherently a judgment call on the part of the referee as to whether a player's behavior merits a red card.

One might anticipate that players with darker skin-tone would receive more red cards because of expectancy effects in social perception, which lead ambiguous behavior to be interpreted in line with prior attitudes and beliefs (Bodenhausen, 1988; Correll, Park, Judd, & Wittenbrink, 2002; Frank & Gilovich, 1988; Hugenberg & Bodenhausen, 2003). In societies as diverse as India, China, the Dominican Republic, Brazil, Jamaica, the Philippines, the United States, Chile, Kenya, and Senegal, light skin is seen as a sign of beauty, status, and social worth (Maddox & Chase, 2004; Maddox & Gray, 2002; Sidanius, Pena, & Sawyer, 2001; Twine, 1998). Negative attitudes towards persons with dark skin may lead a referee to interpret an ambiguous foul as a severe foul and decide to give a red card (Kim & King, 2014; Parsons, Sulaeman, Yates, & Hamermesh, 2011; Price & Wolfers, 2010).

Consider for a moment how you would test this research hypothesis using a complex archival dataset with referee decisions across numerous leagues, games, years, referees, and players and a variety of potentially relevant control variables that might or might not be included. Would you treat each red-card decision as an independent observation? How would you address the possibility that some referees give more red cards than others? Would you try to control for the seniority of the referee? Would you take into account whether a referee's familiarity with a player affects their likelihood of assigning a red card? Would you look at whether players in some leagues are more likely to receive red cards, and whether might there be differences in the proportion of players with dark skin in different leagues and playing in different positions? Each of these factors requires a decision, and each decision might be defensible and simultaneously have implications for the findings observed and the conclusions drawn. You and another

researcher might each make different judgment calls of statistical method, set of covariates, or exclusion rules that are equally *prima facie* valid. This initiative to crowdsource the analysis of a complex dataset examined the extent to which such good faith, subjective choices by different researchers shape the reported results.

### **Stages of the Crowdsourcing Process**

The crowdsourced project progressed through a series of stages including collecting the unique dataset used for the project, recruiting analysts, assessing their subjective beliefs about the hypothesis being tested, repeated rounds of data analysis and peer assessments of analysis quality, online discussion and debate over email, and drafting and revising this report. Project stages are summarized in Table 1 Links to resources from this project (R1.1-R7.1) can be found in the Disclosure section.

– Place Table 1 about here –

#### **Stage 1: Building the dataset**

From a company for sports statistics, we obtained player demographics from all soccer players ( $N = 2,053$ ) playing in the first male divisions of England, Germany, France, and Spain in the 2012-2013 season. This included data about interactions of those players with all referees ( $N = 3,147$ ) that they encountered in their professional career. Thus the data entails a period of multiple years from a player's first professional match until the point in time this data was acquired (June 2014). This data included 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. The dataset was made available as a list with 146,028 dyads of players and referees ([see R1.1](#)).

Players' photos were available from the source for 1,586 out of 2,053 players. Profiles for which no photo was available tended to be relatively new players or players who had just moved up from a team in a lower league. The variable *player skin tone* was coded by two independent raters blind to the research question who, based on the profile photo, categorized players on a 5-point scale ranging from 1 = *very light skin* to 5 = *very dark skin* with 3 = *neither dark nor light skin* as the center value ( $r = 0.92$ ;  $\rho = 0.86$ ). This variable was rescaled to be bounded by 0 (*very light skin*) and 1 (*very dark skin*) prior to the final analysis to ensure consistency among effect sizes between teams and to reflect the largest possible effect. Rescaling was done to 0, 0.25, 0.5, 0.75, and 1, making 0 to 1 the scale range.

A variety of potential independent variables were included in the dataset including information about the player, the referee, or the dyad (see Table 2). The complete codebook is available at R1.2. For players, data included their typical position, weight, and height, and for referees, their country of origin. For each dyad, data included the number of games referees and players encountered each other and the number of yellow and red cards awarded. The variables of age, club, and league— which frequently change throughout a player's career— were only available for players at the time of data collection, not at the time of receiving the particular red card sanctioning. To protect their identities given the sensitivity of the research topic, referees were anonymized and listed by a numerical identifier for each referee and for each country of origin. Importantly, our archival dataset provides the opportunity to estimate the magnitude of the relationship between variables (i.e., player skin tone and referee red card decisions), but does not offer the opportunity to identify causal relations.

– Place Table 2 about here –

## **Stage 2: Recruitment and initial survey of data analysts**

The first three authors and last author posted a description of the project online (see S1 of the Supplementary Materials). This document included an overview of the research question, a description of the dataset and the planned timeline. The project was advertised via Brian Nosek's Twitter account, blogs of prominent academics, and word of mouth.

Seventy-seven researchers expressed initial interest in participating and were given access to the Open Science Framework project page to obtain the data (see R1.1). Individual analysts were welcome to form teams. Of the initial inquiries, 33 teams submitted a report in the first round, and 29 teams submitted a final report. In total, the project involved 61 data analysts plus the four authors who organized the project. A demographic survey revealed that team leaders worked in 13 different countries and came from a variety of research backgrounds including Psychology, Statistics, Research Methods, Economics, Sociology, Linguistics, and Management. Of the 61 data analysts, at the time of conducting the research and authoring the first draft of this manuscript, 38 held a PhD (62%) and 17 a Master's degree (28%). Researchers came from various ranks and included 8 Full Professors (13%), 9 Associate Professors (15%), 13 Assistant Professors (22%), 8 Post-Docs (13%), and 17 Doctoral students (28%). In addition, 27 participants (46%) had taught at least one undergraduate statistics course, 22 (37%) had taught at least one graduate statistics course, and 24 (39%) had published at least one methodological/statistical article.



In addition to their demographic characteristics, at registration we asked team leaders for their present opinion regarding the research question, e.g. “How likely do you think it is that soccer referees tend to give more red cards to dark skinned players?” using a 5-point Likert item from 1 = *Very Unlikely* to 5 = *Very Likely*. This question was asked again at several points in the research project to track beliefs over time.

### **Stage 3: First round of data analysis**

After registration and answering the subjective beliefs survey for the first time, research teams were given access to the data. They then decided their own analytical approach to test the common research questions, and analyzed the data independently of the other teams (see S2 for further details). Then, via a standardized Qualtrics survey, teams submitted to the coordinators a structured summary of their analytical approach or approaches including information about data transformations, exclusions, covariates, the statistical technique used, the software used, the unit of effect size, and the results (see S3 for the text of the survey materials sent to team leaders, [R3.1](#) for the the Qualtrics files, and [R3.2](#) for the full list of analytical approaches).<sup>1</sup>

### **Stage 4: Round-robin peer evaluations of overall analysis quality**

For the remainder of the project, discussion and debate was encouraged between colleagues regarding their respective approaches to the dataset. First, after removing description of the results, the structured summaries were collated into a single questionnaire and distributed to all the teams for peer review. The analytic approaches were presented in a random order and researchers were instructed to provide feedback on at least the first three approaches that they

---

<sup>1</sup> This project also examined whether country-level preferences for light vs. dark skin predict the red card decisions of referees from the countries for which we had data on such preferences. In brief, little to no evidence emerged that referee decisions were moderated by explicit or implicit skin tone preferences. Data for skin tone preferences was however not available from individual referees, only from referees’ nation of origin, and the majority of analysts judged the available dataset to be inadequate to test this potential moderator. Detailed results are reported in S7.

examined. Researchers were asked for both qualitative feedback as well as the assessment: “How confident are you that the described approach below is suitable for analyzing the research questions?”, measured on a 7-point scale from 1= *Unconfident* to 7 = *Confident*. Each team received feedback from an average of about 5 other teams ( $M = 5.32$ ,  $SD = 2.87$ ).

The qualitative and quantitative feedback was aggregated into a single report and shared with all team members. As such, each team received peer review commentaries about their own and other teams’ analysis strategies. Notably, these commentaries came from reviewers that were highly familiar with the dataset, yet at this point teams were unaware of others’ results (see [R4.1](#) and [R4.2](#) for the complete survey and round-robin feedback). Each team therefore had the opportunity to learn from others’ analytic approaches, and from the qualitative and quantitative feedback provided by peer reviewers, but did not have access to each others’ estimated effect sizes. This phase offered opportunity to improve the quality of analyses and, if anything, ought to have promoted convergence in analysis strategies and outcomes.

### **Stage 5: Second round of data analysis**

Following peer review, research teams had the opportunity to change their analysis strategy and draw new conclusions (see S4). Teams submitted their formal report in a standardized format and also filled out a standardized questionnaire similar to that used in the initial round. Notably, researchers were not forced to present a single effect size without robustness checks. Rather, they were encouraged to present results in the way they would in a published article, with a formal methods and results section. Some teams did adopt a model building approach and report the results of the model that they felt was the most appropriate one. The fact that not every team did this represents yet another subjective, yet defensible analytical choice. All analysis reports can be found at R5.1. A brief summary of the methods employed by

each team and a one-sentence description of their findings are presented in S5, with a visualization of one analysis provided in S11.

### **Stage 6: Open discussion and debate, and further analyses**

After the formal analysis reports were compiled and uploaded to the OSF project, a summary e-mail was sent to all teams inviting their review and discussion as a group about the analysis strategies and what to conclude for the primary research question. Team members engaged in a substantive e-mail discussion regarding the variation in findings and analysis strategies (the full text of this discussion can be found at R6.1). For example, one team found a strong influence of five outliers on their analysis. Other teams performed additional analyses to investigate whether their results were similarly driven by a few outliers (interestingly, they were not). Limitations of the dataset were also discussed (S9). The first three authors and last author then wrote a first draft of this paper and all authors were invited to jointly edit and extend the draft using Google Docs for collaborative editing.

When researchers scrutinized others' results, it became apparent that differences in results may have not only be due to variations in statistical models, but also due to variations in the choice of certain covariates. Doing a preliminary reanalysis, the leader of team 10 discovered that the covariates league and club may be responsible for making some results appear non-significant. A debate emerged regarding whether the inclusion of these covariates was quantitatively defensible given that league and club were only available at the time of data collection and likely changed over the course of many players' careers (see [R6.2](#)). The project coordinators thus asked the 10 teams who had included these variables in their final models to re-run their models without said covariates (S10). Additionally, we asked these teams to decide

whether to keep their prior version or use the results from the updated analysis.<sup>2</sup> The results reported in the manuscript reflect teams' choices of their final model.

### **Stage 7: More granular peer assessments of analysis quality**

We were further interested in obtaining more fine-grained expert assessments of each of the final analyses to identify potential flaws that might account for any variability in reported results. We therefore prepared an additional internal peer review assessment that occurred after the methods and results of all teams were known. The analysts participating in the crowdsourced project indicated, for each approach used by each team, their familiarity with that technique (on a five point scale ranging from 1= Very unfamiliar to 5 = Very familiar) (see S12). For some techniques such as “Multiple Regression,” most analysts (34) indicated that they were familiar or very familiar with that technique. For other techniques, such as “Dirichlet process Bayesian clustering” relatively few analysts (3) indicated that they were familiar or very familiar with that technique. Based on their expertise we then assigned researchers between one and three analytical techniques to assess in greater depth. No researcher was assigned to review the approach of their own team. Researchers were only assigned to assess techniques about which they indicated they were familiar (4) or highly familiar (5).

From initial rounds of analysis (Stages 3 to 6), we derived seven issues that presented major obstacles and potentially problematic analytic decisions (see S13). For example, the analysis may have unnecessarily excluded a large number of cases, or may have not adequately accounted for the number of games played. We then developed a questionnaire based around

---

<sup>2</sup> One of the co-authors of the present paper, D. Molden, strongly disagreed with the project coordinators' decision to allow teams to choose to retain these covariates in any final analyses. He argued that the high rate of movement of players between clubs and leagues that occurs each year (~150-200 players per league per year) invalidated the use of static club and league values from a single year in any dataset that spanned multiple years, as the present one did. He further argued that these conditions rendered the decision to use these variables a major analytic mistake, not a defensible analytic choice. For more details see [R6.2](#)

these seven issues, asking whether a given analysis did *not* take each issue into account (on a five point scale ranging from 1 = Strongly Disagree to 5 = Strongly Agree). Additionally, we asked the expert peer reviewers to indicate in an open ended question whether there was an additional issue that might bias results from that particular analytical approach and to indicate quantitatively on the same scale as before, the extent to which that point presents an issue. The lower the score, the more obstacles were avoided, and a higher score indicates that more issues were left unaddressed. In a final question, we asked the expert raters to state how convinced they were that the presented approach successfully addressed most potential concerns regarding the analysis (1= Very unconvinced, 5 = Very convinced).

### **Main Findings from the Crowdsourced Project**

#### **How variable were results from different teams using the same data to test the same hypothesis?**

Twenty-nine independent teams of researchers submitted analytical approaches and refined these throughout the crowdsourcing project. Table 2 shows each team's final analytic technique, model specifications and reported effect size.<sup>3</sup> Analytic techniques ranged from simple linear regression to complex multilevel regression and Bayesian approaches. Teams also varied highly in their decisions regarding which covariates to include (see [R7.1](#)). Table 4 shows that the 29 teams used 21 unique combinations of covariates. Apart from the variable 'games', which was used by all teams, just one covariate (player position, 62%) was used in more than half of the analytic strategies and three were used in just one analysis. Two sets of covariates

---

<sup>3</sup> Because the majority of teams used analyses that favored the reporting of odds ratios, we chose this effect size as the common effect size. For those who performed standard linear regression techniques, we used traditional conversion formulas for both Cohen's *d* and standardized regression weights (assumed to be a correlation coefficient) found in Borenstein, Hedges, Higgins, and Rothstein (2009). Additionally, because the prevalence of red cards is so low, we make the "rare disease" assumption by assuming that the risk ratios yielded in analyses adopting a Poisson regression framework yield a fair approximation to the odds ratio (Viera, 2008).

were used by three teams each, and four sets of covariates were used by two teams each. The remaining 15 teams used a unique combination of covariates.

– Place Tables 3 and 4 about here –

What were the consequences of this variability in analytic approaches? Researchers' conclusions varied regarding whether or not soccer referees were more likely to give red cards to dark skin toned players than light skin toned players. Figures 1 and 2 show the effect sizes and 95% confidence intervals alongside the description of the analytic approach provided by each team. Statistical results ranged from 0.89 (slightly negative) to 2.93 (moderately positive) in odds ratio units, with a median of 1.31. The confidence intervals for many of the estimates overlap, which is expected as they are based on the same data. From a null hypothesis significance testing standpoint, twenty teams (69%) found a significant positive relationship and nine teams (31%) observed a non-significant relationship. No team reported a significant negative relationship.

– Place Figures 1 and 2 about here –

### **What types of analytic approaches were used?**

Examining the consequences of specific analysis choices more directly, teams who employed logistic or Poisson models reported estimates that tended to be larger than teams using linear models. More specifically, 15 teams used logistic models (11/15 significant, median OR = 1.34, MAD = 0.07), six teams used Poisson models (4/6 significant, median OR = 1.36, MAD = 0.08), six teams used linear models (3/6 significant, median OR = 1.21, MAD = 0.05), and two teams used models classified as miscellaneous (2/2 significant).

Teams also varied in their approaches to handling the non-independence of players and referees, which resulted in variability regarding both median estimates and rates of significance. In total, 15 teams estimated a variance component for players and/or referees (12/15 significant,

Median OR = 1.32, MAD = 0.12), eight teams used clustered standard errors (4/8 significant, Median OR = 1.28, MAD = 0.13), five teams did not account for this artifact (4/5 significant, Median OR = 1.39, MAD = 0.28), and one team used fixed effects for the referee variable (0/1 significant, OR = 0.89).

### **Did researchers' beliefs regarding the hypothesis change over time?**

Analysts' subjective beliefs about the theoretical hypothesis were assessed four times during the project: initial registration (i.e., before they had received the data), after researchers accessed the data and submitted their analytical approach, at the time of submission of their final analyses, and after a group discussion with all approaches and results available for collective review. Responses were centered in all subsequent analyses to increase interpretability. Subjective beliefs exhibited variability across time (see Figure 3). When we asked researchers at their initial registration (i.e., before they had received the data), there was slight agreement on average that a positive relationship existed between number of red cards and player skin-tone, yet opinions varied greatly ( $M = 0.61$ ,  $SD = 1.20$ ). We asked the same question again after researchers accessed the data and submitted their analytical approach. At that point, the slight initial agreement had turned into slight disagreement regarding whether a relationship existed ( $M = -0.61$ ,  $SD = 0.88$ ). At the point of the submission of their final analyses, overall slight agreement existed again of the hypothesized relationship at a magnitude similar to initial beliefs, yet again with substantial variability ( $M = 0.61$ ,  $SD = 1.20$ ). Finally, after a group discussion with all approaches and results available for collective review, overall agreement increased slightly and, notably, variability in beliefs decreased ( $M = 0.75$ ,  $SD = 0.70$ ), suggesting some convergence over time.

– Place Figure 3 about here –

In the fourth and final survey we administered items assessing more nuanced beliefs about our primary research question (i.e., whether there is an association between player skin tone and referee red card decisions). These included items such as “The effect is positive and due to referee bias” and “There is little evidence for an effect.” Analysts responded to these items on scales ranging from 1 (strongly disagree) to 7 (strongly agree). The items, means, and standard deviations are reported in Table 5. By the end of the project, a majority of teams agreed that the data showed a positive relationship between number of red cards and player skin-tone but were unclear regarding the underlying mechanism. The greatest endorsement (78% agreement) was given to the statement “The effect is positive and the mechanism is unknown” ( $M = 5.32$ ,  $SD = 1.47$ ).

-- Place Table 5 about here --

### **What is the association between scientists’ subjective beliefs regarding the hypothesis and the empirical evidence?**

Of particular interest was whether subjective beliefs that the primary research hypothesis is true were related to the results a team obtained. One might anticipate a confirmation bias, such that scientists find in a dataset what they initially expect to find. Alternatively, scientists may rationally update their beliefs in response to the empirical results they obtain, even if those results contradict their initial expectations.

Self-reported beliefs regarding research question 1 at each stage were correlated with the final reported effect size using Spearman’s rho, with the following magnitudes across the four time points (and corresponding 95% CIs): 0.14 [-0.25, 0.49], -0.20 [-0.53, 0.19], 0.43 [0.07, 0.69], 0.41 [0.04, 0.68]. Because both the magnitude of the effect and the estimate precision varied by team, Spearman’s rho correlations were also calculated between the lower bound of the



final reported effect size and self-reported beliefs regarding the primary research question, with the following magnitudes across the four time points (and corresponding 95% CIs): 0.29 [-0.09, 0.60], -0.10 [-0.46, 0.28], 0.52 [0.18, 0.75], 0.58 [0.26, 0.78].

Analysts' beliefs at registration regarding whether dark skin toned players were more likely to receive red cards were not significantly related to the observed effect size of their final report ( $\rho = 0.14$  [-0.25, 0.49]). However, as noted above, beliefs changed considerably throughout the research project, and analysts' *post*-analysis belief in the hypothesis was significantly related to their effect estimate and lower bound ( $\rho = 0.41$  [0.04, 0.68] and  $\rho = 0.58$  [0.26, 0.78], respectively), suggesting some updating of beliefs based on the empirical results. Although the sample size was small ( $N = 29$ ), the overall results of the crowdsourced project are more consistent with rational updating of beliefs based on the evidence than with confirmation bias (i.e., scientists simply finding what they expected to find).

### **Does researcher expertise explain the variability in results?**

An important question is whether the variability in the analytic choices made and results found by each team (Figures 1 and 2) simply results from teams with the greatest statistical expertise making different choices than the remaining teams. Relatedly, teams whose members have more quantitative expertise may show greater convergence in their estimated effect sizes. To examine these questions further, we dichotomized teams into two groups using latent class analysis. The first group ( $N = 9$ ) was more likely to have a team member who: had a PhD (100% vs. 53%), was professor at a university (100% vs. 37%), had taught a graduate statistics course more than twice (100% vs. 0%), and had at least one methodological/statistical publication (78% vs. 47%). Seventy-eight percent of teams with high ratings of general statistical expertise reported effects that were statistically significant (median OR = 1.39, MAD = 0.13) whereas

68% of teams with less expertise reported a significant effect (median OR = 1.30, MAD = 0.13). Further analyses of the effects of quantitative expertise on choice of statistical models is provided in S6. Note however that both teams higher and lower in expertise exhibited considerable variability in whether they found a significant effect, and had a similar degree of dispersion in their effect size estimates. Thus, overall, statistical expertise may have had some influence on analytic approaches and estimated effect sizes, but this does not explain the high variability in these choices or in the conclusions they supported.

### **Do peer ratings of overall analysis quality explain the variability in results?**

We further examined whether peer-evaluations of the overall quality of each analytic approach were associated with the reported results. During the round robin feedback phase when the methods (but not results) from each team were known, each analytical plan received ratings of peers' confidence regarding the suitability of the approach. The final effect sizes from teams whose analytic approach received high (4/5 or 5/5) confidence ratings (median OR = 1.31, MAD = 0.15) did not differ from effect sizes of those of teams who received lower confidence ratings (Median OR = 1.28, MAD = 0.12). Thus little evidence emerged that the variability in estimated effect sizes observed across teams was attributable to a subset of analyses that were lower in quality overall.

### **Do peer assessments of specific problematic issues with each analysis explain variability in results?**

Toward the end of the crowdsourcing process, we matched researchers based on their statistical expertise to final analytical approaches conducted by other teams. The qualitative feedback to the first round of analytical approaches had indicated that analytical techniques would need to address seven different analytical issues. Expert researchers assessed the extent to

which the final analytical approach taken by another team addressed each of the seven statistical issues. Additionally, each assessor provided a rating of overall confidence in the approach. For the 29 approaches there was an average of 2.55 assessors, with 16 approaches reviewed by 3 expert assessors and 13 approaches reviewed by 2 expert assessors. The average number of statistical issues that remained across teams was ( $M = 2.18$ ,  $SD = .55$ ) on a scale of 1 to 5, with lower numbers indicating that the approach included fewer analytical issues.

Researchers tended to be more convinced by approaches in which fewer problematic issues remained, as indicated by a correlation between the average rating of the seven statistical issues and assessors' rating of confidence in an approach ( $r = -0.75$  [-0.60, -0.86]). Interestingly, however, analytical issues were unrelated to the OR for the relationship between darker skin tone and red cards received ( $r = 0.06$  [-0.35, 0.31]). Likewise, overall peer confidence in each analytic approach was unrelated to the OR for skin tone and red cards ( $r = -0.03$ , [-0.39, 0.60]). Overall, relatively little evidence emerged that analytic approaches with identifiable statistical problems accounted for the variability in results across teams, for example by producing abnormally large or small effect sizes. S14 reports exploratory analyses attempting to identify subsets of analyses that exhibited more convergence across teams.

### **Implications for the Scientific Endeavor**

It is easy to understand that effects can vary across independent tests of the same research question using different sources of data. Variation in measures, samples, and random error in assessment naturally produce variation in results. Here, we demonstrate that variation in estimated effect sizes emerges for analyses using the same data, contingent on researchers' choices and assumptions during the analysis. Independent teams estimated effects for the primary research question ranging from 0.89 to 2.93 in odds ratio units (1.0 indicates a null

effect), with zero teams finding a negative effect, nine teams finding no significant relationship, and twenty teams finding a positive effect. If, as in virtually all other research projects, a single team had conducted the study, selecting randomly from the present teams, there would have been a 69% probability of reporting a positive result and a 31% probability of reporting a null effect from an identical data set and when testing the same hypothesis.

This variability in results could not be readily accounted for by differences in expertise. Analysts with high and comparatively lower levels of quantitative expertise both exhibited high levels of variability in their estimated effect sizes. Further, analytic approaches that received highly favorable evaluations from peers showed the same variability in final effect sizes as analytic approaches that were less favorably rated. The latter was true both in terms of 1) peer ratings of overall quality and 2) a lack of specific issues or problems with the analysis as assessed by scientists selected for their familiarity with that type of analysis.

### **Analysis-contingent results are distinct from p-hacking, the garden of forking paths, and re-analyses of published data**

The main contribution of our paper is in directly demonstrating the extent to which good faith, yet subjective, analytic choices can have an impact on research results. This is related to but distinct from p-hacking (Simonsohn, Nelson, & Simmons, 2013), the garden of forking paths (Gelman & Loken, 2014), or questioning published findings based on re-analyses of the original data.

*P-hacking.* As originally defined by Simonsohn, Nelson, and Simmons (2013), p-hacking is either consciously or unconsciously exploiting researcher degrees of freedom in order to achieve statistical significance. For instance, Simonsohn et al. (2013, p. 534) write that “researchers may file merely the subsets of analyses that produce nonsignificant results. We refer

to such behavior as *p-hacking*.” Thus *p-hacking* is driven by the implicit or explicit goal to obtain statistically significant support for a particular conclusion. Although the specific decisions made in the process of *p-hacked* analyses may be independently justifiable, it is not justifiable to choose an analytic strategy based on whether it provides a desired result. Few editors would accept a paper, even based on a series of *prima facie* defensible analytic choices, if the researchers admitted they made their analytic choices to reach  $p < .05$ .

In the context of crowdsourcing data analysis, all teams knew that their analyses would be observed and public, and the perceived need to achieve a significant result for publishability was lessened by the nature of the project. Distinct from *p-hacking*, highly defensible analytic decisions made without direct incentive to achieve statistical significance can still produce wide variability in effect size estimates. In the case of the hypothesized relationship between player skin tone and referee red card decisions, the findings collectively suggest a positive correlation, but we glimpse this through the fog of variable subjective analytic decisions.

*The garden of forking paths.* Gelman and Loken’s (2014) concept of a garden of forking paths does not require any selection between different analytic options to achieve significant results (as in *p-hacking*), but is contingent on first observing patterns in the data and only then testing for significance. Such data-contingent analyses do capitalize heavily (perhaps unintentionally) on chance, since patterns that emerged randomly are subjected to significance tests whose validity requires *a priori* predictions. This leads to “researcher degrees of freedom without fishing, which consists of computing a single test based on the data, but in an environment where a different test would have been performed given different data” (p. 460).

The analysis contingent results we examined is broader than forking paths, in that variability in effect sizes can occur even when the researcher has not looked for patterns in the

data first and only tested for significance after the fact. For example, we asked analysts to test a specific relationship between player skin tone and referee red card decisions, arguably limiting opportunities for a “garden of forking paths” process, which might take the form of examining relationships between a players’ various group-based characteristics (skin tone, ethnicity, pGDP of country of origin) on the one hand and various referee decisions (red cards, yellow cards, stoppage time, offside calls, disallowed goals), and running formal significance test only for the relationships that seem to emerge as potentially meaningful.

Moreover, imagine if we had required the 29 teams to preregister their analysis plans before observing the data (Nosek, Ebersole, DeHaven, & Mellor, 2017). Preregistration solves the forking paths and p-hacking challenges by removing flexibility of data contingent analyses and reducing the opportunity to present post-hoc tests as a priori (Wagenmakers et al., 2012). However, preregistration would not have prevented the observed variability in effect estimates across teams in our study. Variation in outcomes can be observed based on different, defensible analytic decisions whether they are made *post hoc* or *a priori*.

*Re-analyses of published data.* Making data from published papers more accessible to facilitate re-analyses and post-publication peer review (Hunter, 2012; Simonsohn, 2013; Wicherts et al., 2006) is important for science, but also does not make fully transparent the contingency of analytic decisions on observed findings. For example, few scientists would bother to write (and even fewer editors would publish) a commentary presenting new analyses and results unless they suggest a different conclusion from the original publication. This creates perverse incentives for both original authors and commenters. Original authors have strong incentives to find positive results to achieve publication, and commenters have strong incentives to find different (usually negative) results to achieve publication. Thus, published commentaries

will almost inevitably differ from original articles in their analytic approaches and conclusions, introducing a strong selection bias.

In contrast, when data analysis is crowdsourced prior to publication, any individual analysis will not play a major role in the final publication decision, and the approach is collaborative rather than conflict-oriented. The most obvious incentive may be to avoid making a public error analyzing an open dataset. Thus, crowdsourcing data analysis may reduce dysfunctional incentives for both original and commenter positions, build connections between colleagues, and make transparent all approaches used and all results obtained. Crowdsourcing analysis can provide a much more accurate picture of the robustness of results, and the dependency of the findings on subjective analytic choices.

In sum, our crowd of analysts had no incentive to try different specification and choose one that supported the hypothesis (p-hacking), to first examine the data and only test for significant patterns after-the-fact (the garden of forking paths), or to confirm or disconfirm a finding to achieve publication. Even so, the variability in analytic choices led to variability in observed results. This illustrates the breadth of the challenge posed by analytic choices influencing observed outcomes.

### **How much variability in results is too much?**

As scientists, we can have comparatively more faith in a finding when there is less variability across different approaches to investigating the same phenomenon. In a follow-up to this project, Crowdsourcing Data Analysis 2 (Schweinsberg et al., 2017), a group of over 40 analysts have independently analyzed the same complex dataset to test hypotheses regarding the effects of gender and status on intellectual debates. This new crowd of analysts are reporting radically dispersed effect sizes, and in some cases significant effects in opposite directions for

the same hypothesis tested with the same data. In such extreme cases of little to no convergence in results, the crowdsourcing process suggests the scientific community should have no faith the theoretical hypothesis is true, even if one two teams did find significant support with a defensible analysis that might have been publishable on its own. In the present project on referee decisions, the degree of convergence in results is relatively high by comparison, with over two thirds of teams supporting the hypothesis and the vast majority of teams returning effect size estimates in the predicted direction.

There will almost always be variability in a measured effect depending on analysis choices. As transparency about this increases with data posting rules and further crowdsourced projects, scientists and policymakers will need to make ultimately subjective decisions about how much consistency is enough (and not enough) to conclude an effect is worth believing. Similar subjective and continually debated decisions have had to be made about the cut-off for statistical significance (Benjamin et al., 2017; Johnson, 2013). Setting cut-offs may be particularly challenging for policymakers because a decision must be made, and the ideal information includes both whether an effect exists and its magnitude. For example, some economic interventions might have both societally positive and societally negative effects, and policymakers will want to have precise estimates of each to evaluate the tradeoffs. Policy makers and practitioners may require greater convergence in effect size estimates than scientists, for whom establishing a directional effect is often sufficient for building theory. We believe that crowdsourcing data analysis initiatives will help improve estimation of confidence and uncertainty for policymakers. Crowdsourced analysis, combined with preregistered investigations and replications, will provide more informed benchmarks for the contingency of observed findings on sample, setting, procedures, *and* analysis decisions.



**Generalizability to other datasets**

The results of the present crowdsourced initiative are striking because the present research question—the relationship between player skin tone and referee red card decisions—was clear and, ostensibly, straightforward to investigate. Compared to many research questions in neuroscience, economics, biology, and psychology, this is a research problem of relatively modest complexity. And yet the process of translating this question from natural language to statistical models gave rise to many different assumptions and choices that influenced the conclusions. This raises the possibility of hidden uncertainty due to the wide range of analytic choices available to the researchers across a wide variety of research applications.

Of course, more than one such investigation is needed to determine how contingent research results are on analytic decisions more generally. The conclusions from this demonstration are thus limited to being a case example with plausible, but untested, generalizability. For example, the project coordinators framed a specific research question for the analysts (does player skin tone correlate with referee red card decisions?), which may have artificially reduced the variability in estimated effect sizes. The research question could have been posed more broadly (“is there evidence of bias against minority groups in referee decisions?”), or the key outcome measure (e.g., yellow cards, red cards, stoppage time) left up to each research team. This is being examined in the second crowdsourcing data analysis project, on the roles of gender and status in intellectual debates (Schweinsberg et al., 2017). In this follow-up project, analysts are also choosing how to operationalize each construct (e.g., is academic status best measured by citation counts, job rank, school rank, or some combination?). As noted earlier, this second project finds an even greater variability in the effect size estimates reported by different analysts for the same hypothesis tested using the same data than in the

present initiative. Systematic investigation via crowdsourcing will facilitate more general conclusions about how contingent research results are on analytic choices, and what characteristics of the research question, dataset, and analyses serve as moderating variables.

There are also constraints on the useful application of crowdsourcing strategies. For example, the flexibility in analytical choices and thus their impact on estimated effect sizes is likely to increase with the complexity of the dataset (e.g., longitudinal datasets with missing data, many potential covariates, levels of nesting, statistical models to be chosen). It remains an empirical question how great a role analysis contingent results play in comparatively simple experimental studies with two to four conditions and relatively fewer measured variables. There may still be enough choice points (outlier exclusions, transformations), even when analyzing a relatively simple dataset, to introduce considerable variability in results based on those choices (Steege, Tuerlinckx, Gelman, & Vanpaemel, 2016).

Notably, assessments of the robustness of experimental laboratory effects can also be pursued via replication initiatives repeating the same experimental design with new research participants (Ebersole et al., 2016; Klein et al., 2014). Crowdsourcing analysis approaches are particularly relevant for datasets with many choice points in analysis and that cannot easily be independently replicated with new observations. Crowdsourcing may also add a great deal of value when addressing controversial research questions, or areas where there are many competing theoretical predictions to be adjudicated empirically.

**Recommendations for individual researchers and teams**

Due to practical constraints, most future scientific investigations will not involve crowds of researchers. For a lone analyst working without the benefit of a crowd we would recommend use of a specification curve (Simonsohn, Simmons, & Nelson, 2016) or multiverse analysis (Steege et al., 2016). The analyst in effect tries to come up with every different defensible analysis she can, runs them all, and then computes the likelihood that the number of observed significant results would be seen if there really is no effect (Simonsohn et al., 2016).

Crowdsourcing the analysis of data greatly reduces efficiency relative to a single analyst attempting many specifications. However, when feasible a crowdsourced approach adds value in a number of ways. A globally distributed crowdsourced project will leverage skills, perspectives, and approaches to data analysis that no single analyst or research team can realistically muster on their own. In addition, a crowd of analysts has no perverse incentive to conduct a primary analysis or robustness check that produces statistically significant support for the research hypothesis. In contrast, a traditional research team seeking to publish in a top academic journal has a strong perverse incentive to select both a primary analysis and robustness checks that return publishable results, something that is relatively easy to do given the numerous possible specifications typically available to choose from. Further, the crowdsourcing data analysis allows for debate and discussion between different research teams of a richness and depth not typically seen in the academic review process— in which reviewers and editors rarely have access to the data themselves, and often choose to focus on other aspects of the paper besides the analytical approach chosen.

**Conclusion**

The observed results from a complex dataset can be highly contingent on justifiable, but subjective, analytic decisions. Uncertainty in interpreting research results is therefore not just a function of statistical power or the presence of questionable research practices, it is also a function of the many reasonable decisions that researchers must make in order to conduct the research. This does not mean that data analysis and drawing research conclusions is a subjective enterprise with no connection to reality. It does mean that many subjective decisions are part of the research process and can affect the outcomes. The best defense against subjectivity in science is to expose it. Transparency in data, methods, and process gives the rest of the community opportunity to see the decisions, question them, offer alternatives, and test these alternatives in further research.

## **Disclosures**

### **Author Contribution Statement**

The first and second author contributed equally to the project. EU proposed the idea of crowdsourcing data analysis and wrote the initial project outline. RS, EU, DPM, and BAN developed the research protocol. RS and EU developed the specific research question regarding skin tone influencing referee decisions. RS and DPM collected the referee decisions data and prepared the dataset for analysis. RS and DPM coordinated the different stages of the crowdsourcing process. All other authors worked in teams to analyze data, give feedback and produce individual reports. A detailed list of contributions for each team is provided in S8 of the Supplementary Materials. RS and DPM combined and analyzed the results of the different teams. EU outlined the paper and wrote the first draft of the abstract, introduction, and discussion. RS wrote the first draft of the methods and online supplement. RS and DPM wrote the first draft of the results section. DPM and RS created the figures and tables. BAN heavily revised the manuscript, gave critical comments, and provided overall project supervision. All authors reviewed the paper and many authors provided crucial comments and edits that were then incorporated into this manuscript.

### **Acknowledgments**

Silvia Liverani acknowledges support from a Leverhulme Trust Early Career Fellowship (ECF-2011-576). Tom Stafford was supported by a Leverhulme Trust Research Project Grant. Richard Morey and Eric-Jan Wagenmakers' contribution was supported by an ERC grant from the European Research Council. Daniel P. Martin was supported by the Institute of Education Sciences, U.S. Department of Education (Grant No. R305B090002). Christopher Madan was

supported by a Canadian Graduate Scholarship, Doctoral-level (CGS-D) from the National Science and Engineering Research Council (NSERC) of Canada. Magnus Johannesson received funding from the Jan Wallander and Tom Hedelius Foundation (P2015-0001:1), as well as from the Swedish Foundation for Humanities and Social Sciences (NHS14-1719:1).

### Online Resources

<b>Resource</b>	<b>Project Stage and Web Reference</b>
	<i>Stage 1</i>
R1.1	OSF Project Page: <a href="https://osf.io/47tnc/">https://osf.io/47tnc/</a>
R1.2	Codebook: <a href="https://osf.io/9yh4x/">https://osf.io/9yh4x/</a>
	<i>Stage 3</i>
R3.1	Collection form for analytical approaches: <a href="https://osf.io/yug9r/">https://osf.io/yug9r/</a>
R3.2	List of analytical approaches: <a href="https://osf.io/3ifm2/">https://osf.io/3ifm2/</a>
	<i>Stage 4</i>
R4.1	Survey of analytical strategies: <a href="https://osf.io/evfts/">https://osf.io/evfts/</a>
R4.2	Round robin feedback: <a href="https://osf.io/ic634/">https://osf.io/ic634/</a>
	<i>Stage 5</i>
R5.1	Report of all analyses: <a href="https://osf.io/qix4g">https://osf.io/qix4g</a>
	<i>Stage 6</i>
R6.1	E-mail discussion of analytical approaches: <a href="https://osf.io/8eg94/">https://osf.io/8eg94/</a>
R6.2	Discussion regarding covariates: <a href="https://osf.io/2prib/">https://osf.io/2prib/</a>
	<i>Stage 7</i>
R7.1	Reported rationales for covariate use: <a href="https://osf.io/sea6k/">https://osf.io/sea6k/</a>

**Tables and Figures**

Project Stage	Work Package	Month	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	
1	Building the Dataset		█																				
2	Recruitment and Initial Survey of Data Analysts		█	█																			
3	First round of data analysis				█																		
4	Round-Robin peer evaluations					█																	
5	Second round of data analysis						█																
6	Open discussion and debate, and further analyses							█	█	█	█												
	Write-Up, Edits, Review											█	█	█	█	█							
7	Internal Peer Review of approaches based on Statistical Expertise																	█	█	█	█		
	Write-Up of manuscript																					█	█

*Table 1 Overview of Project Stages and Approximate Duration*

<b>Players</b>	<b>Mean</b>	<b>Stdev</b>	<b>League Country</b>	<b>Players</b>
Height	181.74	6.69	England	564
Weight	75.64	7.10	France	533
Games per Player	71.13	36.17	Germany	489
Yellow Cards	27.41	24.08	Spain	467
Red Cards	0.89	1.26		

<b>Player Skin Color</b>	<b>Rater1</b>	<b>Rater2</b>	<b>Player Position</b>	<b>N</b>
0 (Very light skin)	626	451	Attacking Midfielder	149
0.25	551	693	Center Back	281
0.5	170	174	Center Forward	227
0.75	140	141	Center Midfielder	84
1 (Very dark skin)	98	126	Defensive Midfielder	204
NA	468	468	Goalkeeper	196
			Left Fullback	136
			Left Midfielder	86
			Left Winger	59
			NA	367
			Right Fullback	126
			Right Midfielder	75
			Right Winger	63

*Table 2. Descriptive statistics of player variables.*



Team	Analytic Approach	N covariates	Treatment of Non-Independence	Distribution	Reported Effect Size			Odds Ratio (OR)			
					Unit	Size	95% CI	OR	95% CI		
10	Multilevel regression and logistic regression	3	Variance component	Linear	R	0.01	0.00	0.01	1.03	1.01	1.05
1	Ordinary least squares with robust standard errors, logistic regression	7	Clustered SE	Linear	OR	1.18	0.95	1.41	1.18	0.95	1.41
4	Spearman correlation	3	None	Linear	D	0.10	0.10	0.10	1.21	1.20	1.21
14	Weighted least squares regression with referee fixed-effects and clustered SE	6	Clustered SE	Linear	OR	1.21	0.97	1.46	1.21	0.97	1.46
11	Multiple linear regression	4	None	Linear	D	0.12	0.03	0.22	1.25	1.05	1.49
6	Linear Probability Model	6	Clustered SE	Linear	OR	1.28	0.77	2.13	1.28	0.77	2.13
17	Bayesian logistic regression	2	Variance component	Logistic	OR	0.96	0.77	1.18	0.96	0.77	1.18
15	Hierarchical log-linear modeling	1	None	Logistic	OR	1.02	1.00	1.03	1.02	1.00	1.03
31	Logistic regression	6	Clustered SE	Logistic	OR	1.12	0.88	1.43	1.12	0.88	1.43
30	Clustered robust binomial logistic regression	3	Clustered SE	Logistic	OR	1.28	1.04	1.57	1.28	1.04	1.57
3	Multilevel Binomial Logistic Regression using Bayesian inference	2	Variance component	Logistic	OR	1.31	1.09	1.57	1.31	1.09	1.57
23	Mixed model logistic regression	2	Variance component	Logistic	OR	1.31	1.10	1.56	1.31	1.10	1.56
2	Linear probability model, logistic regression	6	Clustered SE	Logistic	OR	1.34	1.10	1.63	1.34	1.10	1.63
5	Generalized linear mixed models	0	Variance component	Logistic	OR	1.38	1.10	1.75	1.38	1.10	1.75
24	Multilevel logistic regression	3	Variance component	Logistic	OR	1.38	1.11	1.72	1.38	1.11	1.72
28	Mixed effects logistic regression	2	Variance component	Logistic	OR	1.38	1.12	1.71	1.38	1.12	1.71
32	Generalized linear models for binary data	1	Clustered SE	Logistic	OR	1.39	1.10	1.75	1.39	1.10	1.75
8	Negative binomial regression with a log link analysis	0	None	Logistic	OR	1.39	1.17	1.65	1.39	1.17	1.65
25	Multilevel logistic binomial regression	4	Variance component	Logistic	OR	1.42	1.19	1.71	1.42	1.19	1.71
9	Generalized linear mixed effects models with a logit link function	2	Variance component	Logistic	OR	1.48	1.20	1.84	1.48	1.20	1.84
7	Dirichlet process Bayesian clustering	0	None	Miscellaneous	OR	1.71	1.70	1.72	1.71	1.70	1.72
21	Tobit regression	4	Clustered SE	Miscellaneous	R	0.28	0.01	0.56	2.88	1.03	11.47
12	Zero-inflated Poisson regression	2	Fixed effect	Poisson	IRR	0.89	0.49	1.60	0.89	0.49	1.60
26	Three-level hierarchical generalized linear modeling with Poisson sampling	6	Variance component	Poisson	IRR	1.30	1.08	1.56	1.30	1.08	1.56
16	Hierarchical Poisson Regression	2	Variance component	Poisson	IRR	1.32	1.06	1.63	1.32	1.06	1.63
20	Cross-classified multilevel negative binomial model	1	Variance component	Poisson	IRR	1.40	1.15	1.71	1.40	1.15	1.71
13	Poisson Multi-level modeling	1	Variance component	Poisson	IRR	1.41	1.13	1.75	1.41	1.13	1.75
27	Poisson regression	1	None	Poisson	IRR	2.93	0.11	78.66	2.93	0.11	78.66
32	Generalized linear models for binary data	1	Clustered SE	Logistic	OR	1.39	1.10	1.75	1.39	1.10	1.75

Table 3. Analytical approaches chosen by each team with the number of covariates used and how each team treated the non-independence of the data. Effect sizes reported by each team are listed in their original unit as well as in the converted Odds Ratio format. Effect size units are abbreviated as follows: IRR = incidental risk ratio, OR = odds ratio, D = Cohen's d, R = standardized regression coefficient.

Covariate	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	20	21	23	24	25	26	27	28	30	31	32	% used	
Position	■	■	■	■				■	■	■	■	■	■	■	■				■	■	■	■	■	■	■	■	■	■	■	■	62%
Height	■	■		■		■		■				■		■					■	■		■	■	■	■	■	■	■	■	■	38%
Weight	■	■		■		■						■		■					■	■		■	■	■	■	■	■	■	■	■	38%
Age	■	■		■		■					■	■		■		■			■	■		■	■	■	■	■	■	■	■	■	24%
League Country	■					■					■	■		■					■	■		■	■	■	■	■	■	■	■	■	17%
Goals	■	■		■					■		■			■		■				■	■		■	■	■	■	■	■	■	■	17%
Referee Country	■	■	■	■					■					■						■	■		■	■	■	■	■	■	■	■	17%
Victories	■	■		■					■		■			■							■	■		■	■	■	■	■	■	■	10%
Club	■													■																	7%
Referee	■													■																	7%
Player Cards	■															■	■	■													7%
Player																■	■	■													3%
Referee Cards																■	■	■													3%
Draws										■	■																				3%
N Covariates	7	6	2	3	0	3	0	0	2	3	3	2	1	6	1	2	2	2	1	3	2	3	4	6	1	2	3	4	1		

Table 4. Covariates used by each team. Team numbers are listed on the top and covariates on the left. A shaded box indicates that the corresponding team used the covariate in their final model. The table is ordered by the frequency by which each covariate was used.

Question	Mean	SD
Positive relationship likely caused by referee bias	3.37	1.65
Positive relationship likely caused by unobserved variables (e.g., player behavior)	4.21	1.37
Positive relationship but without evidence of cause	5.32	1.47
Positive relationship but it is contingent on a relatively small number of outlier observations	3.18	1.31
Positive relationship but it is contingent on other variables in the dataset (e.g., differences across leagues)	3.84	1.33
Little evidence of a relationship	3.17	1.66
No relationship	2.49	1.28
Negative relationship	1.64	0.80

*Table 5. Mean agreement with potential conclusions that could be drawn about the primary hypothesis tested in the crowdsourced project: whether there is an association between player skin tone and referee red card decisions. Analysts responded to these items on scales ranging from 1 (strongly disagree) to 7 (strongly agree). Note that the complete first item was, “This dataset suggests a positive relationship between darker skin-toned players and frequency of receiving red cards that is likely caused by referee bias.” Items were paraphrased for inclusion in the table.*

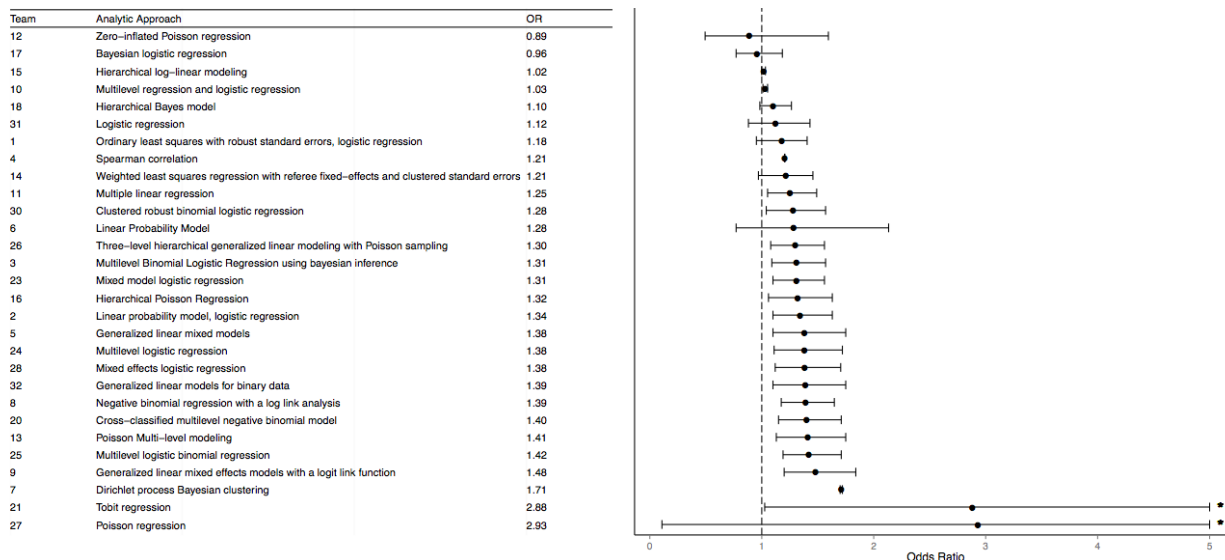


Figure 1. Point estimates and 95% confidence intervals for analysis teams for the primary research question: Are soccer referees more likely to give red cards to dark skin toned players than light skin toned players? Note that the asterisks correspond to a truncated upper bound for Team 21 (11.47) and Team 27 (78.66) to increase the interpretability of this plot.

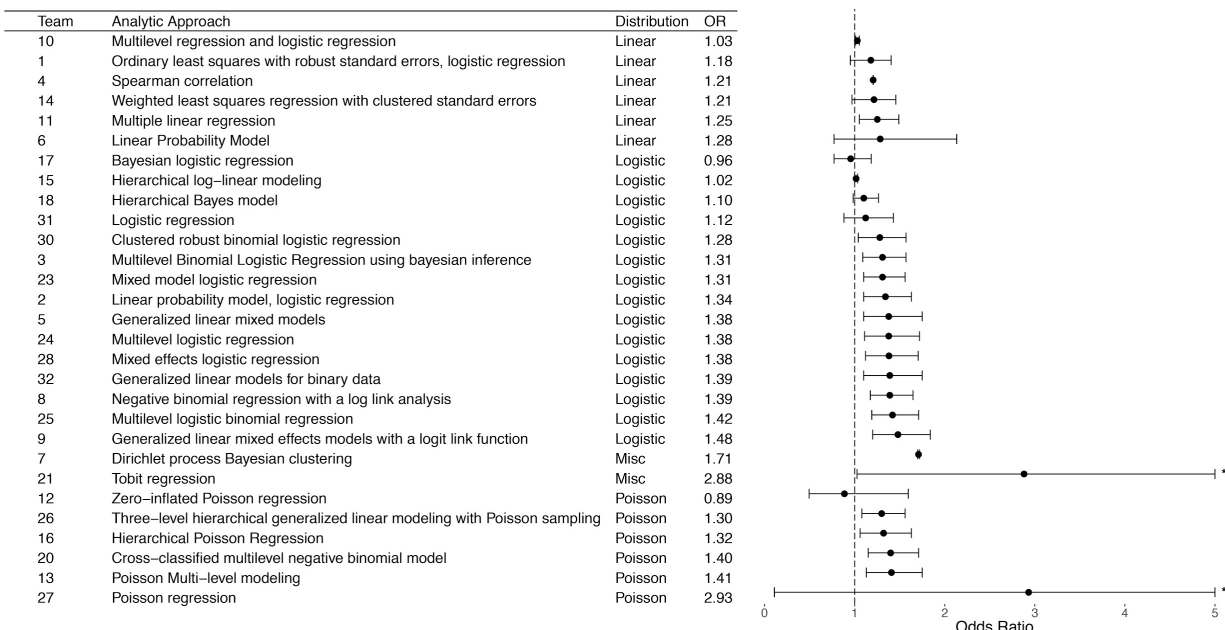


Figure 2. Analytical approaches chosen by each team, clustered by similarity of analytical technique.

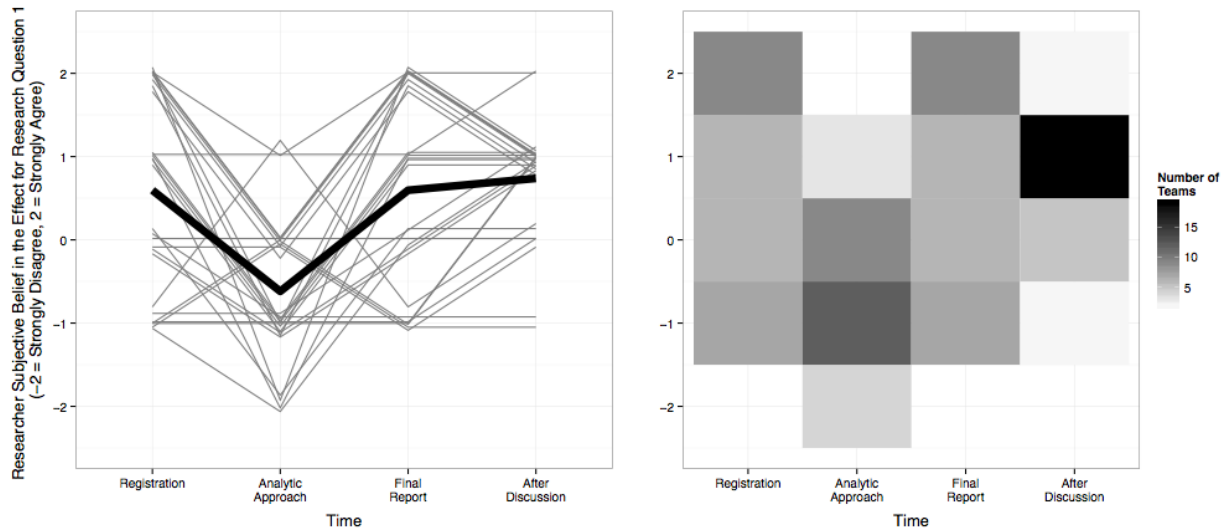


Figure 3. Subjective beliefs across time. The plot on the left reflects team leader beliefs regarding the primary research question: whether player skin tone predicts referee red cards. Each light gray line represents a single team's trajectory throughout the project, and the black trajectory represents the mean value at each time point. Note that each individual trajectory is jittered slightly to increase the interpretability of the plot. The plot on the right represents the consensus (or lack thereof) by plotting the number of team leaders endorsing a particular response category at each time point.

### References

- Babtie, A. C., Kirk, P., & Stumpf, M. P. H. (2014). Topological sensitivity analysis for systems biology. *Proceedings of the National Academy of Sciences of the United States of America*. <http://doi.org/10.1073/pnas.1414026112>
- Bakker, M., van Dijk, A., & Wicherts, J. M. (2012). The rules of the game called Psychological Science. *Perspectives on Psychological Science: A Journal of the Association for Psychological Science*, 7, 543–554. <http://dx.doi.org/10.1177/1745691612459060>
- Benjamin, D., Berger, J., Johannesson, M., Nosek, B., Wagenmakers, E.J., Richard Berk, R., et al. (2017). Redefine statistical significance. Unpublished manuscript. Available at: <https://psyarxiv.com/mky9j>
- 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. <http://dx.doi.org/10.1037/0022-3514.55.5.726>
- Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). Converting among effect sizes. In M. Borenstein, L. V. Hedges, J. P. T. Higgins, & H. R. Rothstein (Eds.), *Introduction to meta-analysis*. Chichester, UK: John Wiley & Sons.
- Carp, J. (2012a). On the plurality of (methodological) worlds: Estimating the analytic flexibility of fMRI experiments. *Frontiers in Neuroscience*, 6, 149. <http://dx.doi.org/10.3389/fnins.2012.00149>
- Carp, J. (2012b). The secret lives of experiments: methods reporting in the fMRI literature. *NeuroImage*, 63, 289–300. <http://dx.doi.org/10.1016/j.neuroimage.2012.07.004>
- 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 and*

*Social Psychology*, 83, 1314–1329. <http://dx.doi.org/10.1037/0022-3514.83.6.1314>

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science*, 25, 7–29.

Ebersole, C. R., Atherton, O. E., Belanger, A. L., Skulborstad, H. M., Allen, J. M., Banks, J. B., Baranski, E., Bernstein, M. J., Bonfiglio, D. B. V., Boucher, L., Brown, E. R., Budiman, N. I., Cairo, A. H., Capaldi, C. A., Chartier, C. R., Chung, J. M., Cicero, D. C., Coleman, J. A., Conway, J. G., Davis, W. E., Devos, T., Fletcher, M. M., German, K., Grahe, J. E., Hermann, A. D., Hicks, J. A., Honeycutt, N., Humphrey, B., Janus, B., Johnson, D. J., Joy-Gaba, J. A., Juzeler, H., Keres, A., Kinney, D., Kirshenbaum, J., Klein, R. A., Lucas, R. E., Lustgraaf, C. J. N., Martin, D., Menon, M., Metzger, M., Moloney, J. M., Morse, P. J., Prislín, R., Razza, T., Re, D. E., Rule, N. O., Sacco, D. F., Sauerberger, K., Shriver, E., Shultz, M., Siemsen, C., Sobocko, K., Sternglanz, R. W., Summerville, A., Tskhay, K. O., van Allen, Z., Vaughn, L. A., Walker, R. J., Weinberg, A., Wilson, J. P., Wirth, J. H., Wortman, J., & Nosek, B. A. (in press). Many Labs 3: Evaluating participant pool quality across the academic semester via replication. *Journal of Experimental Social Psychology*.

Ebrahim, S., Sohani, Z. N., Montoya, L., Agarwal, A., Thorlund, K., Mills, E. J., & Ioannidis, J.

P. A. (2014). Reanalyses of randomized clinical trial data. *JAMA: The Journal of the American Medical Association*, 312, 1024–1032.

<http://dx.doi.org/doi:10.1001/jama.2014.9646>

Frank, M. G., & Gilovich, T. (1988). The dark side of self- and social perception: Black uniforms and aggression in professional sports. *Journal of Personality and Social Psychology*, 54, 74–85.

Gelman, A., & Loken, E. (2014). The statistical crisis in science. *American Scientist*, 102, 460.

- Hugenberg, K., & Bodenhausen, G. V. (2003). Facing prejudice: implicit prejudice and the perception of facial threat. *Psychological Science, 14*, 640–643.  
<http://dx.doi.org/10.1046/j.0956->
- Hunter, J. (2012). Post-publication peer review: opening up scientific conversation. *Frontiers in Computational Neuroscience, 6*, 63.
- Johnson, V.E. (2013). Revised standards for statistical evidence. *Proceedings of the National Academy of Sciences, 110(48)*, 19313–19317.
- Kim, J. W., & King, B. G. (2014). Seeing stars: Matthew effects and status bias in Major League Baseball umpiring. *Management Science, 60*, 2619–2644.  
<http://dx.doi.org/10.1287/mnsc.2014.1967>
- Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B., Jr., Bahník, Š., Bernstein, M. J., Bocian, K., Brandt, M. J., Brooks, B., Brumbaugh, C. C., Cemalcilar, Z., Chandler, J., Cheong, W., Davis, W. E., Devos, T., Eisner, M., Frankowska, N., Furrow, D., Galliani, E. M., Hasselman, F., Hicks, J. A., Hovermale, J. F., Hunt, S. J., Huntsinger, J. R., IJzerman, H., John, M., Joy-Gaba, J. A., Kappes, H. B., Krueger, L. E., Kurtz, J., Levitan, C. A., Mallett, R., Morris, W. L., Nelson, A. J., Nier, J. A., Packard, G., Pilati, R., Rutchick, A. M., Schmidt, K., Skorinko, J. L., Smith, R., Steiner, T. G., Storbeck, J., Van Swol, L. M., Thompson, D., van't Veer, A., Vaughn, L. A., Vranka, M., Wichman, A., Woodzicka, J. A., & Nosek, B. A. (2014). Investigating variation in replicability: A "many labs" replication project. *Social Psychology, 45*, 142-152.
- Krumholz, H. M., & Peterson, E. D. (2014). Open access to clinical trials data. *JAMA: The Journal of the American Medical Association, 312*, 1002–1003.  
<http://dx.doi.org/10.1001/jama.2014.9647>



- Maddox, K. B., & Chase, S. G. (2004). Manipulating subcategory salience: exploring the link between skin tone and social perception of Blacks. *European Journal of Social Psychology*, *34*, 533–546. <http://dx.doi.org/10.1002/ejsp.214>
- Maddox, K. B., & Gray, S. A. (2002). Cognitive Representations of Black Americans: Reexploring the role of skin tone. *Personality and Social Psychology Bulletin*, *28*, 250–259. <http://dx.doi.org/10.1177/0146167202282010>
- McCullough, B. D., McGeary, K. A., & Harrison, T. D. (2006). Do economics journal archives promote replicable research? *Canadian Journal of Economics*, *41*, 1406–1420. <http://dx.doi.org/10.1111/j.1540-5982.2008.00509.x>
- Parsons, C. A., Sulaeman, J., Yates, M. C., & Hamermesh, D. S. (2011). Strike three: Discrimination, incentives, and evaluation. *The American Economic Review*, *101*, 1410–1435. <http://www.jstor.org/stable/23045903>
- Price, J., & Wolfers, J. (2010). Racial discrimination among NBA referees. *The Quarterly Journal of Economics*, *125*, 1859–1887.
- Sakaluk, J. K., Williams, A. J., & Biernat, M. (2014). Analytic review as a solution to the misreporting of statistical results in Psychological Science. *Perspectives on Psychological Science: A Journal of the Association for Psychological Science*, *9*, 652–660. <http://dx.doi.org/10.1177/1745691614549257>
- Schweinsberg, M. et al. (2017). *Crowdsourcing data analysis 2: Gender, status, and science*. Crowdsourced research project in progress.
- Sidanius, J., Pena, Y., & Sawyer, M. (2001). Inclusionary Discrimination: Pigmentocracy and patriotism in the Dominican Republic. *Political Psychology*, *22*, 827–851. <http://dx.doi.org/10.1111/0162-895X.00264>

- Simonsohn, U. (2013). Just post it: The lesson from two cases of fabricated data detected by statistics alone. *Psychological Science*, *24*(10), 1875-1888.
- Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*, 1359–1366. <http://dx.doi.org/10.1177/0956797611417632>
- Simonsohn, U., Simmons, J., & Nelson, L. (2016). *Specification curve: Descriptive and inferential statistics for all plausible specifications*. Unpublished manuscript.
- Steege, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science*, *11*(5), 702 –712.
- Twine, F. W. (1998). *Racism in a racial democracy: The maintenance of White supremacy in Brazil*. New Brunswick, NJ: Rutgers University Press.
- Viera, A. J. (2008). Odds ratios and risk ratios: what’s the difference and why does it matter? *Southern Medical Journal*, *101*, 730–734.  
<http://dx.doi.org/10.1097/SMJ.0b013e31817a7ee4>
- Wagenmakers, E.-J., Wetzels, R., Borsboom, D., van der Maas, H. L. J., & Kievit, R. A. (2012). An agenda for purely confirmatory research. *Perspectives on Psychological Science: A Journal of the Association for Psychological Science*, *7*, 632–638.  
<http://dx.doi.org/10.1177/1745691612463078>
- Wicherts, J. M., Borsboom, D., Kats, J., & Molenaar, D. (2006). The poor availability of psychological research data for reanalysis. *American Psychologist*, *61*, 726–728.  
<http://dx.doi.org/10.1037/0003-066X.61.7.726>

**ONLINE SUPPLEMENTS**

Supplement 1: Publicly posted project description	40
Supplement 2: Additional notes on the research process	48
Supplement 3: Complete surveys sent to analysis teams	49
Supplement 4: Changes in analytic approaches based on peer feedback	55
Supplement 5: Final results	56
Supplement 6: Additional analyses of research team expertise and statistical model choice	60
Supplement 7: Research Questions 2a and 2b	61
Supplement 8: Author contribution forms from analysis teams	64
Supplement 9: Limitations of the dataset	65
Supplement 10: Club and league as covariates	66
Supplement 11: IPython notebook visualisation of the dataset	67
Supplement 12: Survey of familiarity with each analytic approach	68
Supplement 13: Peer review of final analytical choices for specific issues	70
Supplement 14: Exploratory analyses in search of converging results	74

**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](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. **1<sup>st</sup> 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. **2<sup>nd</sup> 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](#).

### **2.1<sup>st</sup> 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	SI	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 Belgoy		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

Mauro Boselli



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.



<b>Variable Name:</b>	<b>Variable Description:</b>
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. 2<sup>nd</sup> 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](mailto:mail@raphael.rs)) and Dan ([dpmartin42@gmail.com](mailto: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](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:

`{://SurveyLink?d=Take the Survey}`

Or copy and paste the URL below into your internet browser:

`{://SurveyURL}`

-----

`{://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 responses 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**

<b>Team</b>	<b>Method</b>
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.

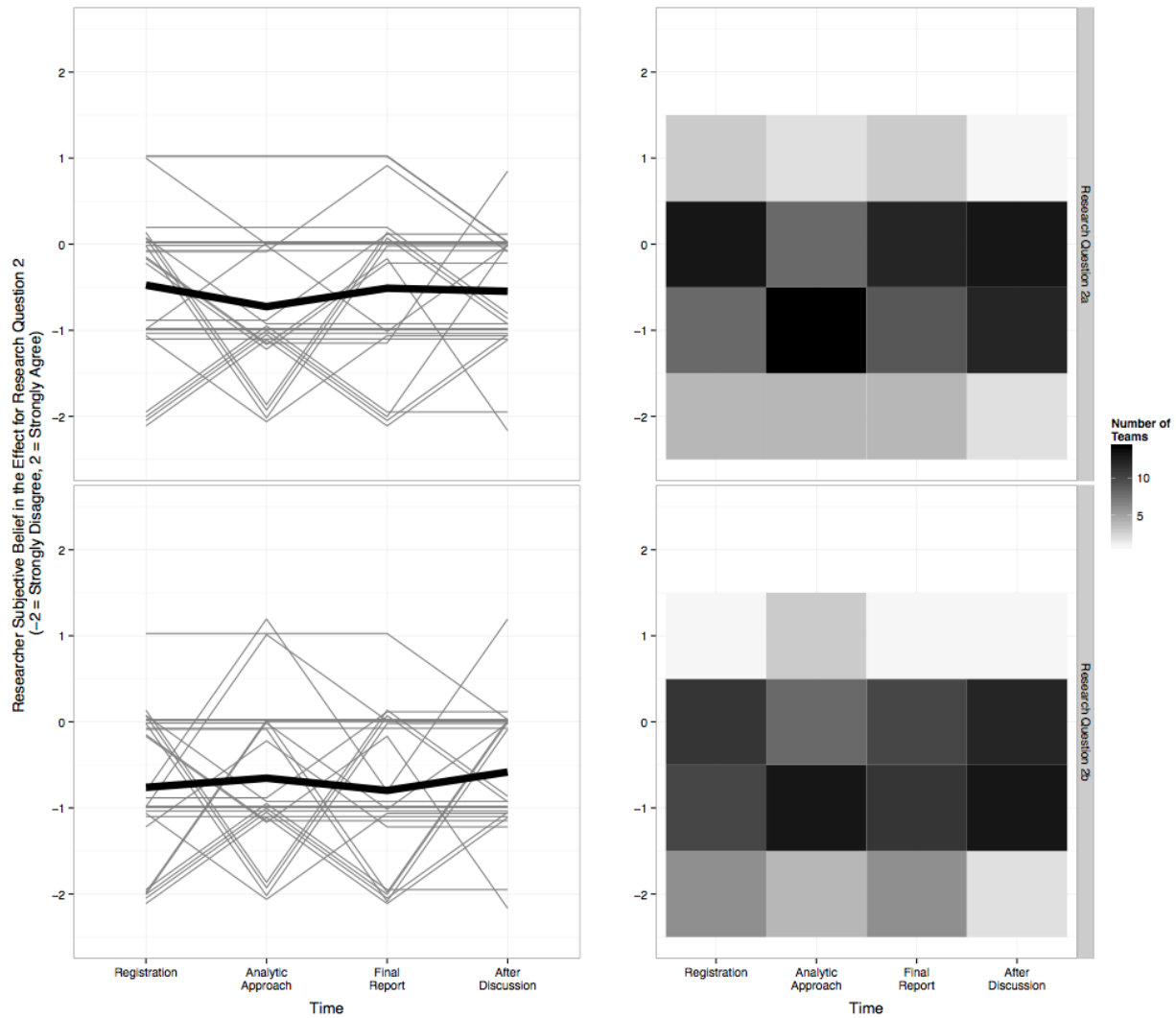
For the country of each referee, we included average scores of implicit and explicit preferences for light vs. dark skin tone that had been gathered in independent research by Project Implicit (Nosek et al., 2007; Nosek, Banaji, & Greenwald, 2002). Implicit preference scores for each referee country 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 also Marini 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 reflect 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**

<b>Team</b>	<b>Name</b>	<b>Contribution</b>
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 Witkowski	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	Analysis, 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](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)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Linear probability model, logistic regression (2)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Multilevel Binomial Logistic Regression using Bayesian inference (3)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Spearman correlation (4)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Generalized linear mixed models (5)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Linear Probability Model (6)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Dirichlet process Bayesian clustering (7)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Negative binomial regression with a log link analysis (8)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Generalized linear mixed effects models with a logit link function(9)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Multilevel regression and logistic regression (10)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Multiple linear regression (11)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Zero-inflated Poisson regression (12)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Poisson Multi-level modeling (13)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Weighted least squares regression with referee fixed-effects and clustered SE (14)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Hierarchical log-linear modeling (15)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Hierarchical Poisson Regression (16)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Bayesian logistic regression (17)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>



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

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		OR	Min	Max
			Issues	Confidence			
<b>0-1</b>	<b>20</b>	<b>1</b>	<b>1.06</b>	<b>5.00</b>	<b>1.40</b>	<b>1.15</b>	<b>1.71</b>
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
<b>2-3</b>	<b>28</b>	<b>2</b>	<b>1.54</b>	<b>5.00</b>	<b>1.38</b>	<b>1.12</b>	<b>1.71</b>
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
<b>&gt;3</b>	<b>25</b>	<b>4</b>	<b>1.83</b>	<b>4.67</b>	<b>1.42</b>	<b>1.19</b>	<b>1.71</b>
>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*