

Running Head: Crowdsourcing Analytics

Crowdsourcing data analysis: Do soccer referees give more red cards to dark skin toned players?

Running Head: Crowdsourcing Analytics

Crowdsourcing data analysis: Do soccer referees give more red cards to dark skin toned players?

Authors

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

Contact Authors: RSilberzahn@iese.edu, eric.luis.uhlmann@gmail.com, dpmartin42@gmail.com, nosek@virginia.edu

Affiliations

¹Binghamton University School of Management; ²Brigham Young University; ³Brunel University London, MRC Biostatistics Unit, Cambridge and Imperial College London; ⁴City University of New York; ⁵HEC Paris; ⁶IESE Business School; ⁷Independent; ⁸INSEAD; ⁹Linnaeus University; ¹⁰Loyola University Maryland; ¹¹Ludwig-Maximilians-Universität München; ¹²Michigan State University; ¹³Michigan State University & University of Hong Kong; ¹⁴Northwestern University; ¹⁵Ohio State University; ¹⁶Stockholm School of Economics; ¹⁷Stockholm University; ¹⁸Temple University; ¹⁹The University of Chicago; ²⁰UC Berkeley; ²¹Universidade Federal da Fronteira Sul & Universidade Federal de Santa Catarina; ²²University of Alberta; ²³University of Amsterdam; ²⁴University of Bamberg; ²⁵University of British Columbia; ²⁶University of Cologne; ²⁷University of Giessen; ²⁸University of Groningen; ²⁹University of Liverpool; ³⁰University of Manchester; ³¹University of Minnesota; ³²University of Padua; ³³University of Sheffield; ³⁴University of Toronto; ³⁵University of Virginia; ³⁶Center for Open Science; ³⁷University of Washington; ³⁸University of Washington Bothell; ³⁹University of Würzburg; ⁴⁰University of Zurich; ⁴¹Westat;

Abstract

Twenty-nine teams involving 61 analysts used the same data set to address the same research questions: whether soccer referees are more likely to give red cards to dark skin toned players than light skin toned players and whether this relation is moderated by measures of explicit and implicit bias in the referees' country of origin. Analytic approaches varied widely across teams. For the main research question, 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 significant positive effect and nine teams (31%) observed a non-significant relationship. The causal relationship however remains unclear. No team found a significant moderation between measures of bias of referees' country of origin and red card sanctionings of dark skin toned players. Crowdsourcing data analysis highlights the contingency of results on choices of analytic strategy, and increases identification of bias and error in data and analysis. Crowdsourcing analytics represents a new way of doing science; a data set is made publicly available and scientists at first analyze separately and then work together to reach a conclusion while making subjectivity and ambiguity transparent.

Keywords = crowdsourcing science, research methods, data analysis, scientific transparency, referee decision making, skin tone preferences

Significance Statement: Subjective analyst decisions affect the reliability and robustness of scientific findings. The present study employed a crowdsourcing methodology to investigate, for the first time, the extent to which scientists blind to each other's approach obtain the same results when analyzing the same data set to test the same hypothesis. After the first round of reporting, the 29 teams of analysts reported results with highly varying effect sizes, and moderate consensus. After feedback rounds and discussions, teams submitted their final reports. Analytical strategies still varied, yet 69% of teams reported significant result and 78% of the researchers concluded that the dataset suggests a positive association. Results highlight the value of crowdsourcing data analysis to increase transparency in science.

Creativity is a fundamental part of science. 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 (1–3), it is easy to overlook the fact that the analysis phase is imbued with theory, assumptions, and choice points that could have major implications for the reported outcome. In many cases, there is no single appropriate analysis. Rather, there are many reasonable (and many unreasonable) approaches to evaluating data that bear on a research question (2, 4–6).

This may be understood conceptually, but there is little appreciation for its implications in practice. Peer reviewers may comment and suggest improvements to a chosen analysis strategy, but rarely do those comments come with access to the actual data set (7). 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. Similarly, reviewers may not have appropriate expertise to identify issues with the analytical approach taken. 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 are rare (8–11). 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 (2). Had the authors made different assumptions, an entirely different result might have been observed (12). The present article reports an investigation of the impact of analysis decisions on research results as 29 teams analyze the same data set to evaluate the same research question. The assembled teams are collectively expert on quantitative methods in general. This offers an opportunity to observe a diversity of research strategies to test a research question, and the diversity of outcomes that may result.

Crowds create opportunities to identify and resolve uncertainties in data analysis including: increased likelihood of identifying analytic errors (7), potentially leveraging the wisdom of crowds by using the central tendency of many results (13–15), identifying and debating the appropriateness of different analytic assumptions, and selecting superior analysis strategies from numerous options. Most importantly, perhaps, crowdsourcing data analysis can make transparent the profound implications that different choices in the analytical process have in the research process (3, 16, 17).

Research application of crowdsourcing data analysis: Skin-tone and red cards in soccer

We crowdsourced analysis to evaluate whether soccer referees are more likely to give red cards to dark skin toned players than light skin toned players, and whether this tendency is moderated by cultural preferences for different skin tones. 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.

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 (18–21). 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 (22–24). For instance, negative implicit or explicit 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 (25–27). Further, given that group-based biases frequently result from cultural socialization (28, 29), referees from societies characterized by a strong preference for light over dark skin may be especially prone to exhibit this effect. Importantly, our archival data set provides the opportunity to estimate the magnitude of the relationship between variables, but does not offer the opportunity to identify causal relations.

METHODS

The first three authors and the senior 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 data set and the planned timeline. The project was advertised via the senior author's Twitter account, blogs of prominent academics, and word of mouth.

Data Analysts. Seventy-seven researchers expressed initial interest in participating and were given access to the Open Science Framework project page to obtain the data (<https://osf.io/47tnc/>). 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. 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, 38 hold 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%) have taught at least one undergraduate statistics course, 22 (37%) have taught at least one graduate statistics course, and 24 (39%) have published at least one methodological/statistical article.

Data set. 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. We also took from this source data about the 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 date 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 data set was made available as a list with 146,028 dyads of players and referees (<https://osf.io/47tnc/>).

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 re-scaled 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.

A range of potential independent variables was included in the data concerning the player, the referee, or the dyad. The complete codebook is available at: <https://osf.io/9yh4x/>. For players, data included their typical position, weight, and height at the time of data collection, 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 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.

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 (30, 31). Implicit preference scores for each referee country had been calculated using a skin tone Implicit Association Test (IAT) (32), 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 (33)).

Procedure. At registration we asked team leaders for their present opinion regarding the research questions with a single question for each hypothesis, e.g. "How likely do you think it is

that soccer referees tend to give more red cards to dark skinned players?” with 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. After registration, research teams were given access to the data, 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 about this process). Then, via a standardized Qualtrics survey, teams submitted to the coordinators a structured summary of their analytical approach 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, <https://osf.io/yug9r/> for the the Qualtrics files, and <https://osf.io/3ifm2/> for the full list of analytical approaches).

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 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* (see S3). 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 data set, yet at this point teams were unaware of others’ results (see <https://osf.io/evfts/> and <https://osf.io/ic634/> for the complete survey and round-robin feedback). Each team had the opportunity to learn from others’ analytic approaches, and from the qualitative and quantitative feedback provided by peer reviewers. Here, it became apparent that some teams treated the static variables club or league as relevant for particular referee-player interactions. In the project description it had been unclear that these variables reflected club and league membership only at time of data collection. Project coordinators circulated an e-mail to clarify this (see S2). However, teams were not obliged to change their analytical approach based on the round-robin feedback.

Following peer review, research teams decided their final analysis strategy and the conclusions they drew from the results of their analysis. Participants submitted their final report in a standardized format and also filled out a standardized questionnaire similar to that used in the initial round. All final analysis reports can be found here: <https://osf.io/qix4g>. A summary of the methods employed by each team and a one-sentence description of their findings are presented in S4.

After final analysis reports were compiled and uploaded to the Open Science Framework project space, 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 here <https://osf.io/8eg94/>). 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). Finally, the first three authors and the senior author 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 controversial covariates league and country may be responsible for making some results appear non-significant. A debate emerged regarding whether the inclusion of these covariates was quantitatively defensible (see <https://osf.io/2prib/>). 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. Additionally, we asked these teams to decide whether to keep their prior version or use the results from the updated analysis. The results displayed in the manuscript reflect teams' choices of their final model¹. After this additional round of analysis, the project coordinators updated the paper and invited all team members to make their final edits and approve the manuscript.

RESULTS

From the 79 researchers who initially registered for the crowdstorming project, 33 teams were formed and submitted an initial analytical approach. Of those, 29 teams also submitted a final report. Submitted analytical approaches were diverse, ranging from simple linear regression techniques to complex multilevel regression techniques and Bayesian approaches. Table 1 shows each team's analytic technique, reported effect size, and a number of characteristics describing how their model was specified (e.g., the number of covariates used in the analysis). In total, there were 21 unique combinations of covariates among the 29 teams. 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 were used by three teams each, and four sets of covariates were used by two teams each. All other 15 teams used a combination of covariates, which only their own team used. Table 1 shows variation in analytic strategies for number of covariates ($M = 2.83$ $Stdev = 2.05$), treatment of the non-independent structure of the data, statistical distribution chosen for the

¹ One of the co-authors of the paper, D. Molden, strongly disagreed with the project coordinator's 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 data set 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 detail see <https://osf.io/2prib/>

outcome, and reported effect sizes. More detail regarding specific covariates chosen by each team can be seen in Table 2. Reasons that teams gave for their initial inclusion/exclusion of particular covariates can be found at <https://osf.io/sea6k/>.

For the primary research question, 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. Fig. 1 shows 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 and non-significantly negative) to 2.93 (moderately positive) in odds ratio units², with a median of 1.31. From a null hypothesis significance testing standpoint, twenty teams (69%) found a significant positive effect and nine teams (31%) observed a non-significant relationship. No team reported a significant negative relationship.

After their final analyses, we asked those ten teams who had used the league and club variables in their final analyses to reconduct their analyses without these questionable variables and choose which model to use as their final result. This was done to examine whether results may be potentially confounded by these two variables. The removal of these two covariates overall corresponded to a slight increase in effect size (Median OR = 1.25, MAD = 0.12 to Median OR = 1.32, MAD = 0.07).³ Overview Tables 1 and 2 reflect teams' final model choice.

-- Place Tables 1 and 2 about here --

Overall, teams who employed logistic or Poisson models reported estimates that were somewhat 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 estimated and rates of significance. In total, 15 teams used random effects (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).

² 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, & Rothstein (34). 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 (35).

³ Of the ten reanalyses, three results shifted to being positive and significant, and one (using no covariates) shifted to a negative and significant result. Overall, six of the ten retained their initial models. Three teams explicitly stated that their results only changed marginally and were not contingent on the inclusion of the two covariates in question.

When researchers submitted their final report they also indicated their subjective conclusions about the primary study hypothesis. At that point, 15 team leaders judged it likely or very likely that soccer referees tended to give more red cards to dark skinned players, 7 team leaders judged such an effect to be unlikely and 6 team leaders neither likely nor unlikely.

In additional analyses (S5), we examined correlations between effect sizes and researchers' initial and final subjective beliefs, teams' statistical expertise, and ratings of confidence regarding teams' analytical approaches.

-- Figure 1 about here --

Results for the second research question regarding whether referees from countries high in skin-tone prejudice were more likely to award red cards to dark skin toned players were much more homogeneous. Results for the majority of teams yielded extremely wide confidence intervals and only one team reported a statistically significant effect. Researchers were also concerned about the appropriateness of the available data given that there was very little variation in the country-level estimates. The majority (75% and 72%) were unconfident to somewhat unconfident regarding how appropriate the dataset was for investigating RQ2a and RQ2b. For these reasons results were not discussed in the main text (See S6 for more detailed analyses).

In addition to the Final Submission phase, subjective beliefs were assessed four times during the project (see Fig. 2). Note that this measure was centered in all subsequent analyses to increase interpretability. When we asked researchers at their initial registration (i.e., before they had received the data), overall there was slight agreement that a positive relationship existed between number of red cards and player skin-tone, yet opinions varied greatly ($M = 0.61$, $Stdev = 1.20$). 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 that such a relationship was real ($M = -0.61$, $Stdev = 0.88$). After the round-robin feedback where researchers viewed others' approaches and comments, opinions shifted again. At the point of their final submission, overall slight agreement existed again of the hypothesized relationship at a magnitude similar to the initial beliefs, yet again with substantial variability ($M = 0.61$, $Stdev = 1.20$). Finally, after a group discussion with all approaches and results available, overall agreement increased again very slightly and notably, the variability in beliefs decreased ($M = 0.75$, $Stdev = 0.70$).

After the discussion, and before seeing the draft of this report, most teams agreed moderately that the data showed a positive relationship between number of red cards and player skin-tone. In this final survey, a set of supplementary items assessing agreement with more nuanced beliefs (e.g., "There is little evidence for an effect," "The effect is positive and due to referee bias") revealed greatest endorsement (78% agreement) of the position that "The effect is

positive and the mechanism is unknown” ($M = 5.32$, $SD = 1.47$ on a scale ranging from 1 = *strongly disagree* to 7 = *strongly agree*; see S7 for more details).

-- Figure 2 about here --

DISCUSSION

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 effect size is also present in the *same data* contingent on choices and assumptions in the analysis process. We observed variation in the effect estimates of whether soccer referees gave more red cards to dark skin toned players. We also observed convergence on the discrete judgment of whether there was a positive effect in the data. These crowdsourcing results illustrate both the contingency of effects as a function of analytic choices, and the opportunity for converging beliefs through shared examination and evaluation of a research question using a shared data set.

The median result ($OR = 1.31$) indicated that the odds were 31% higher for players rated as having the darkest skin tone to receive a red card when compared to players rated as having the lightest skin tone. Assuming a 40 game season, these results suggest that the probability of receiving at least one red card over a season is 15.2% for a player with the darkest skin tone and 11.8% for a player with the lightest skin tone.⁴ The estimated effects ranged from 0.88 to 2.93 in odds ratio units (1.0 indicates a null effect), with zero teams finding a negative effect, nine teams finding no 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% likelihood of reporting a positive result and a 31% likelihood of reporting a null effect.

What is the correct answer?

With 29 analyses of the same question using the same data, 28 more than the typical research article, it is reasonable to expect a clear conclusion about which is the right method and whether a statistically significant effect would be found. One approach for attaining consensus is the central tendency of all analytic approaches. Indeed, this is a core logic of wisdom of the crowds (13–15, 36). The error in each individual case is canceled out in the aggregate estimate. Yet, this approach ignores variations in quality or assumptions behind analytical approaches.

⁴ Because lighter skin toned players played more games than darker skin toned players, the referee-player dyads were disaggregated into referee-player-game units. In this case, with a median odds ratio of 1.31, the prevalence of red cards is 0.314% for referee-player-game units with players with the lightest skin tone rating, and 0.411% for units with players with the darkest skin tone rating. Over n games, the probability of a given player to receive at least one red card is equal to $1 - (\text{probability never receiving a card})^n$. Assuming 40 games, this probability for a player with the darkest skin tone is $1 - (1 - 0.00411)^{40}$, while this probability for a player with the lightest skin tone is $1 - (1 - 0.00314)^{40}$.

Crowdsourcing research may benefit from a more deliberate approach, one in which analytical approaches are discussed and refined.

The peer review process between initial and final analysis strategy provided opportunity for collective improvement across the analytic strategies. Peer review can filter out some strategies perceived as less defensible thus improving the aggregate set. Additionally, on open group discussion of the diversity of analytic approaches, a single, superior method could have become evident. In a lengthy email discussion among researchers, a few approaches were mentioned to be rather clearly inappropriate, and a number of approaches were each recognized as defensible (for the full text of the email discussion see <https://osf.io/8eg94/>). One discussion lead to the conclusion that statistical significance in some models was affected by the inclusion of two covariates.

Despite the variation in analytical approaches, we do observe some evidence for a conclusion. Following discussion, the analysts converged toward agreement that there is a small, statistically significant relationship between player skin tone and receiving red cards, the cause of which is unknown.

Data set shortcomings and future directions for research on referee decisions

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. Further, our crowdsourcing project attracted participants from a wide range of academic disciplines. Disciplines vary in their focus on identifying an empirical effect (controlling for as few covariates as possible) or a theoretical effect (controlling for as many covariates as possible). If this research had been conducted among teams within a single discipline, there could have been more homogeneity regarding the selection of covariates. Moreover, the longitudinal nature of the data set and the dyadic focus of referee-player interactions made it possible to observe a multitude of interactions and potential red card decisions of referees. Yet, the structure of the data meant that some other potentially meaningful variables were not available. Future research should examine the effect with data collected at the game level and variables such as club and league collected over time, which was a shortcoming of the current data set.

Another major limitation is that data on explicit and implicit skin tone preferences 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 data set was not adequate to answer the question effectively. 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.

Limitations and challenges of crowdsourcing analytics

Based on our experiences with this first crowdsourcing project, we see great opportunities for crowdsourcing research and are mindful of limitations and challenges that need to be overcome in future projects applying this new method.

Crowdsourcing of data analysis is inefficient in that numerous analysts conduct multiple rounds of data analysis to answer a single research question. But, consider that inefficiency in comparison to the status quo in which a research question is examined and reported using a single analysis strategy. Conventional practice makes little accommodation for the possible contingency of the results on the analytic method (2, 6). Moreover, misspecification of results via analysis strategy is virtually undetectable without an ethic of open data and community review of analytic strategies. It is conceivable that the relative inefficiency trade-offs would actually produce a net benefit by having many independent analysts for a complex data set compared to the currently prevalent practice of individual analysis teams providing stand-alone analyses of privately held data. Further, the use of 29 independent teams helped us illustrate the variation in analytic strategies and conclusions, but - in practice - fewer independent teams may be needed to assess robustness of conclusions. A related challenge, then, is how variation in analytical strategies can be channeled into meaningful conclusions. In our case, while beliefs tended to converge towards the end of the project, final reports still included analytical choices (e.g. statistical methods and inclusion of particular covariates) that lack consensus among teams. Future crowdsourcing projects could investigate process variations on independence of teams' choices and consensus building on final model decisions.

Another limitation involves the present system of professional rewards. Professional incentives are limited for authorship in a lengthy author string rather than getting all or most of the credit for individual or small team contributions. Encouragingly, however, 61 data analysts from all over the world volunteered for the present project. The scientific appeal of crowdsourcing may, at least for some, outweigh the material rewards of participation. Or, even more optimistically, people may recognize that professional reputation is less a function of how many names are on any given article, and more a function of the quality of one's contributions to the research -- large or small.

A related challenge is the potential diffusion of responsibility. With a very large team, individual researchers may feel less responsible for the quality and accuracy of the overall research project, therefore increasing the risk of error in the report. We sought to minimize this possibility with (a) a highly modularized approach so that each team “owned” and was responsible for a small portion of the total project, (b) transparency so that team members’ contributions, strategies, and code were available to others to maximize accountability and detectability of error, and (c) leadership that took responsibility for integrating the modular pieces, soliciting feedback, resolving disagreement, and promoting project quality (37). Yet the challenge remains how teams can be incentivized throughout the project to adapt their approaches based on feedback and stay involved and actively contribute throughout. In our project, representatives from 21 teams provided comments to the manuscript, others wrote e-mails suggesting edits, and all teams that were approached to conduct further analyses responded promptly and volunteered to do so. Deep scrutiny and suggestions for edits came rather from a smaller group of roughly ten teams leaders. Crowdsourcing projects in general thus face the challenge to find ways to incentivize and identify researchers’ involvement throughout and decide on how to deal with situations in which justified critique on analytical approaches is not taken into account.

Finally, in some cases crowdsourcing data analysis could lead to scientific confusion, such as when one analytic approach is actually ideal but is lost amid many less optimal approaches. Again, however, with complex data sets such as this one, it seems unlikely that a single team that collects the data and conducts the analysis in isolation is likely to identify and use the single most optimal method. Crowdsourcing may help identify and surface superior methods over the alternatives. For simpler data sets there may be more initial consensus between analysts on the best approach to use, and therefore comparatively greater convergence in results.

Conclusion

Crowdsourcing analytics represents a new way of doing science. A data set is made publicly available and scientists at first conduct their analyses independently and then work together to reach a conclusion regarding the most appropriate methodology while making ambiguity, subjectivity, and disagreement transparent. If a consensus in results emerges, scientists can speak with one voice. Conversely, if different analysts obtain distinct results with each employing defensible approaches, the subjectivity of science is highlighted and embraced as part of the reality and challenge of acquiring knowledge.

Author Contribution Statement

The first two authors 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 questions regarding skin tone influencing referee decisions and moderation by skin tone preferences. RS and DPM collected the referee decisions data and prepared the data set for analysis. BAN designed the implicit and explicit skin tone preferences measures and collected the relevant data. 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 created the figures. 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.

Tables and Figures

| Team | Analytic Approach | N covariates | Treatment of Non-Independence | Distribution | Reported Effect Size | | | Odds Ratio (OR) | | |
|------|---|--------------|-------------------------------|---------------|----------------------|------|------------|-----------------|------------|--|
| | | | | | Unit | Size | 95% CI | OR | 95% CI | |
| 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 | |
| 2 | Linear probability model, logistic regression | 6 | Clustered SE | Logistic | OR | 1.34 | 1.10 1.63 | 1.34 | 1.10 1.63 | |
| 3 | Multilevel Binomial Logistic Regression using Bayesian inference | 2 | Random effect | Logistic | OR | 1.31 | 1.09 1.57 | 1.31 | 1.09 1.57 | |
| 4 | Spearman correlation | 3 | None | Linear | D | 0.10 | 0.10 0.10 | 1.21 | 1.20 1.21 | |
| 5 | Generalized linear mixed models | 0 | Random effect | Logistic | OR | 1.38 | 1.10 1.75 | 1.38 | 1.10 1.75 | |
| 6 | Linear Probability Model | 6 | Clustered SE | Linear | OR | 1.28 | 0.77 2.13 | 1.28 | 0.77 2.13 | |
| 7 | Dirichlet process Bayesian clustering | 0 | None | Miscellaneous | OR | 1.71 | 1.70 1.72 | 1.71 | 1.70 1.72 | |
| 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 | |
| 9 | Generalized linear mixed effects models with a logit link function | 2 | Random effect | Logistic | OR | 1.48 | 1.20 1.84 | 1.48 | 1.20 1.84 | |
| 10 | Multilevel regression and logistic regression | 3 | Random effect | Linear | R | 0.01 | 0.00 0.01 | 1.03 | 1.01 1.05 | |
| 11 | Multiple linear regression | 4 | None | Linear | D | 0.12 | 0.03 0.22 | 1.25 | 1.05 1.49 | |
| 12 | Zero-inflated Poisson regression | 2 | Fixed effect | Poisson | IRR | 0.89 | 0.49 1.60 | 0.89 | 0.49 1.60 | |
| 13 | Poisson Multi-level modeling | 1 | Random effect | Poisson | IRR | 1.41 | 1.13 1.75 | 1.41 | 1.13 1.75 | |
| 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 | |
| 15 | Hierarchical log-linear modeling | 1 | Random effect | Logistic | OR | 1.02 | 1.00 1.03 | 1.02 | 1.00 1.03 | |
| 16 | Hierarchical Poisson Regression | 2 | Random effect | Poisson | IRR | 1.32 | 1.06 1.63 | 1.32 | 1.06 1.63 | |
| 17 | Bayesian logistic regression | 2 | Random effect | Logistic | OR | 0.96 | 0.77 1.18 | 0.96 | 0.77 1.18 | |
| 18 | Hierarchical Bayes model | 2 | Random effect | Logistic | OR | 1.10 | 0.98 1.27 | 1.10 | 0.98 1.27 | |
| 20 | Cross-classified multilevel negative binomial model | 1 | Random effect | Poisson | IRR | 1.40 | 1.15 1.71 | 1.40 | 1.15 1.71 | |
| 21 | Tobit regression | 4 | Clustered SE | Miscellaneous | R | 0.28 | 0.01 0.56 | 2.88 | 1.03 11.47 | |
| 23 | Mixed model logistic regression | 2 | Random effect | Logistic | OR | 1.31 | 1.10 1.56 | 1.31 | 1.10 1.56 | |
| 24 | Multilevel logistic regression | 3 | Random effect | Logistic | OR | 1.38 | 1.11 1.72 | 1.38 | 1.11 1.72 | |
| 25 | Multilevel logistic binomial regression | 4 | Random effect | Logistic | OR | 1.42 | 1.19 1.71 | 1.42 | 1.19 1.71 | |
| 26 | Three-level hierarchical generalized linear modeling with Poisson sampling | 6 | Random effect | Poisson | IRR | 1.30 | 1.08 1.56 | 1.30 | 1.08 1.56 | |
| 27 | Poisson regression | 1 | None | Poisson | IRR | 2.93 | 0.11 78.66 | 2.93 | 0.11 78.66 | |
| 28 | Mixed effects logistic regression | 2 | Random effect | Logistic | OR | 1.38 | 1.12 1.71 | 1.38 | 1.12 1.71 | |
| 30 | Clustered robust binomial logistic regression | 3 | Clustered SE | Logistic | OR | 1.28 | 1.04 1.57 | 1.28 | 1.04 1.57 | |
| 31 | Logistic regression | 6 | Clustered SE | Logistic | OR | 1.12 | 0.88 1.43 | 1.12 | 0.88 1.43 | |
| 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 1. This table shows the 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 | 2 | 3 | 3 | 2 | 1 | 6 | 1 | 2 | 2 | 2 | 1 | 3 | 2 | 3 | 4 | 6 | 1 | 2 | 3 | 4 | 1 | | | |

Table 2. This overview shows the 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.

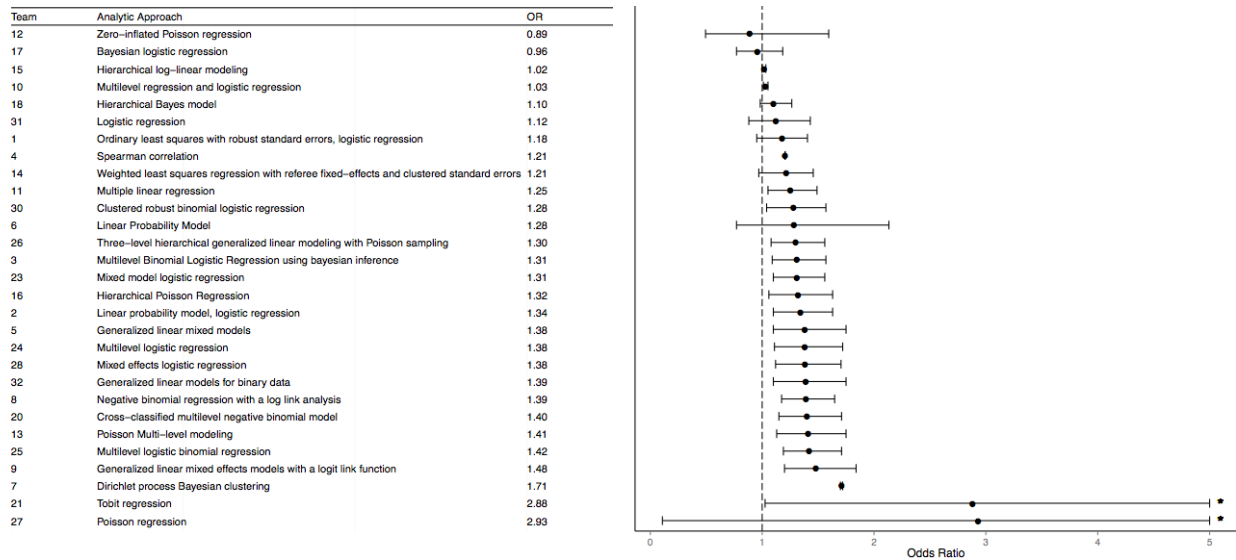


Fig. 1. Point estimates and 95% confidence intervals for analysis teams for Research Question 1: Are soccer referees more likely to give red cards to dark skin toned players than light skin toned players? Note that the asterisks correspond to a truncated upper bound for Team 21 (11.47) and Team 27 (78.66)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

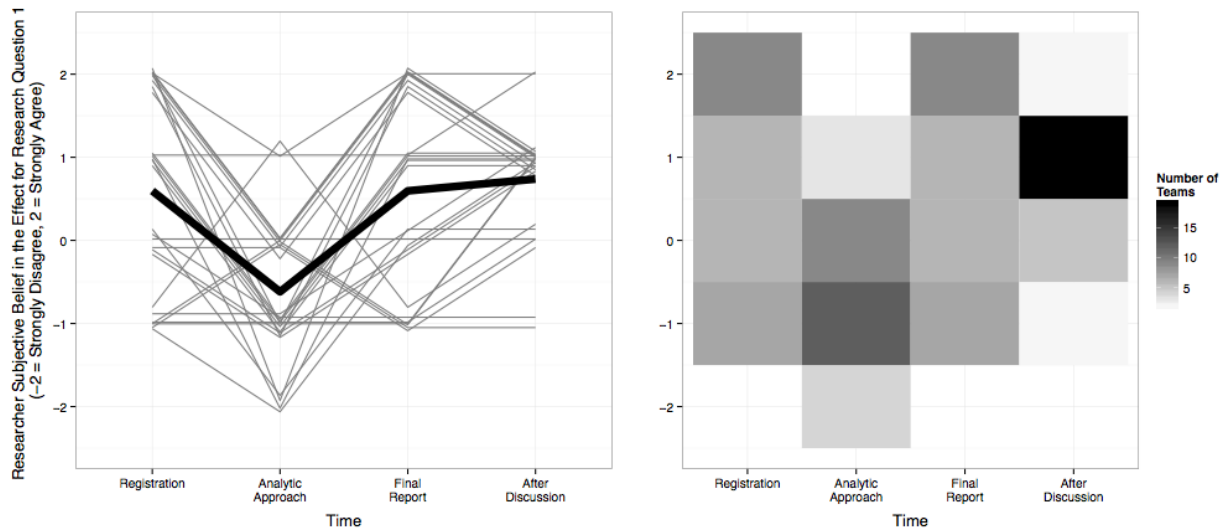


Fig. 2. The plot on the left reflects team leader beliefs regarding the primary research question: whether player skin tone predicts referee red cards. Each light gray line represents a single team's trajectory throughout the project, and the black trajectory represents the mean value at each time point. 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

1. Bakker M, van Dijk A, Wicherts JM (2012) The Rules of the Game Called Psychological Science. *Perspect Psychol Sci* 7:543–554.
2. Gelman A, Loken E (2014) The Statistical Crisis in Science. *Am Sci* 102:460.
3. Simmons JP, Nelson LD, Simonsohn U (2011) False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychol Sci* 22:1359–1366.
4. Carp J (2012) The secret lives of experiments: methods reporting in the fMRI literature. *Neuroimage* 63:289–300.
5. Carp J (2012) On the plurality of (methodological) worlds: estimating the analytic flexibility of FMRI experiments. *Front Neurosci* 6:149.
6. Wagenmakers E-J, Wetzels R, Borsboom D, van der Maas HLJ, Kievit RA (2012) An Agenda for Purely Confirmatory Research. *Perspect Psychol Sci* 7:632–638.
7. Sakaluk JK, Williams AJ, Biernat M Analytic Review as a Solution to the Misreporting of Statistical Results in Psychological Science. *Perspect Psychol Sci* 9:652–660.
8. Ebrahim S et al. (2014) Reanalyses of randomized clinical trial data. *JAMA* 312:1024–1032.
9. Krumholz HM, Peterson ED (2014) Open access to clinical trials data. *JAMA* 312:1002–1003.
10. McCullough BD, McGeary KA, Harrison TD (2006) Do Economics Journal Archives Promote Replicable Research? *Canadian Journal of Economics* 41:1406–1420.
11. Wicherts JM, Borsboom D, Kats J, Molenaar D (2006) The poor availability of psychological research data for reanalysis. *Am Psychol* 61:726–728.
12. Babbie AC, Kirk P, Stumpf MPH (2014) Topological sensitivity analysis for systems biology. *Proc Natl Acad Sci U S A*. Available at: <http://dx.doi.org/10.1073/pnas.1414026112>.
13. Galton F (1907) Vox Populi. *Nature* 75:450–451.
14. Mannes AE, Soll JB, Larrick RP (2014) The wisdom of select crowds. *J Pers Soc Psychol* 107:276–299.
15. Surowiecki J (2004) *The Wisdom of Crowds: Why the Many Are Smarter than the Few and How Collective Wisdom Shapes Business, Economies, Societies, and Nations* (Doubleday Books, New York).

16. Nosek BA, Bar-Anan Y (2012) Scientific Utopia: I. Opening Scientific Communication. *Psychol Inq* 23:217–243.
17. Nosek BA, Spies JR, Motyl M (2012) Scientific Utopia: II. Restructuring Incentives and Practices to Promote Truth Over Publishability. *Perspect Psychol Sci* 7:615–631.
18. Maddox KB, Chase SG (2004) Manipulating subcategory salience: exploring the link between skin tone and social perception of Blacks. *Eur J Soc Psychol* 34:533–546.
19. Maddox KB, Gray SA (2002) Cognitive Representations of Black Americans: Reexploring the Role of Skin Tone. *Pers Soc Psychol Bull* 28:250–259.
20. Sidanius J, Pena Y, Sawyer M (2001) Inclusionary Discrimination: Pigmentocracy and Patriotism in the Dominican Republic. *Polit Psychol* 22:827–851.
21. Twine FW (1998) *Racism in a Racial Democracy: The Maintenance of White Supremacy in Brazil* (Rutgers University Press).
22. Bodenhausen GV (1988) Stereotypic biases in social decision making and memory: testing process models of stereotype use. *J Pers Soc Psychol* 55:726–737.
23. Correll J, Park B, Judd CM, Wittenbrink B (2002) The police officer's dilemma: using ethnicity to disambiguate potentially threatening individuals. *J Pers Soc Psychol* 83:1314–1329.
24. Hugenberg K, Bodenhausen GV (2003) Facing prejudice: implicit prejudice and the perception of facial threat. *Psychol Sci* 14:640–643.
25. Kim JW, King BG (2014) Seeing Stars: Matthew Effects and Status Bias in Major League Baseball Umpiring. *Manage Sci* 60:2619–2644.
26. Parsons CA, Sulaeman J, Yates MC, Hamermesh DS (2011) Strike Three: Discrimination, Incentives, and Evaluation. *Am Econ Rev* 101:1410–1435.
27. Price J, Wolfers J (2010) Racial discrimination among NBA referees. *The Quarterly Journal of Economics* 125:1859–1887.
28. Banaji MR, Roediger HL, Nairne JS (2001) in *The nature of remembering: Essays in honor of Robert G. Crowder*, eds Banaji MR, Roediger HL, Nairne J S, Neath I, Surprenant A (American Psychological Association, Washington, DC), pp 117–150.
29. Greenwald AG, Banaji MR (1995) Implicit social cognition: attitudes, self-esteem, and stereotypes. *Psychol Rev* 102:4–27.
30. Nosek BA et al. (2007) Pervasiveness and correlates of implicit attitudes and stereotypes. *European Review of Social Psychology* 18:36–88.

31. Nosek BA, Banaji M, Greenwald AG (2002) Harvesting implicit group attitudes and beliefs from a demonstration web site. *Group Dyn* 6:101–115.
32. Greenwald AG, McGhee DE, Schwartz JL (1998) Measuring individual differences in implicit cognition: the implicit association test. *J Pers Soc Psychol* 74:1464–1480.
33. Marini M et al. (2013) Overweight people have low levels of implicit weight bias, but overweight nations have high levels of implicit weight bias. *PLoS One* 8:e83543.
34. Borenstein M, Hedges LV, Higgins JPT, Rothstein HR (2009) Converting among effect sizes. *Introduction to Meta-Analysis*, eds Borenstein M, Hedges LV, Higgins JPT, Rothstein HR (John Wiley & Sons, Ltd, Chichester, UK).
35. Viera AJ (2008) Odds ratios and risk ratios: what's the difference and why does it matter? *South Med J* 101:730–734.
36. Lorge I, Fox D, Davitz J, Brenner M (1958) A survey of studies contrasting the quality of group performance and individual performance, 1920-1957. *Psychol Bull* 55:337–372.
37. Open Science Collaboration (2014) The Reproducibility Project: A Model of Large-Scale Collaboration for Empirical Research on Reproducibility. *Implementing Reproducible Computational Research (A Volume in The R Series)*, eds Stodden V, Leisch F, Peng R (Taylor & Francis, New York, NY), pp 299–323.

ONLINE SUPPLEMENT

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

Crowdstorming 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 data set. 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 "crowdstorming a dataset." Multiple independent analysts are recruited to investigate the same hypothesis or hypotheses on the same data set 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 data sets 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 crowdstorming a data set, 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 crowdstorming project. In order to meet the timeline, some later phases may commence while earlier phases are in process. For example, some of the report will be written while final data analyses are still in process.

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

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

Elaboration of Project Stages

1. Registration

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

2. 1st Round Analyses

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

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

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
























The Dataset

From a company for sports statistics, we obtained data and profile photos from all soccer players ($N = 2,053$) playing in the first male divisions of England, Germany, France and Spain in the 2012-2013 season and all referees ($N = 3,147$) that these players played under in their professional career (see Fig. 1). 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. 1: Player overview with list of referees and player-referee statistics, such as matches, goals, and cards.

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

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](https://projectimplicit.org/).

Data Structure

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

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

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

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

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

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

References

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. The two project coordinators created independent clean data sets from the original source. Both data sets were checked against each other for accuracy and spot checks with the original source revealed no differences, thus this updated data set 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. When we examined analytical approaches, 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 data set, 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 and pointed to benefits of sharing analytical approaches before settling on particular results.

3. 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 Crowdstorming Research Project. We are excited to have you in the team! I am sending you below some further information, which will help us work together. Raphael Silberzahn and Dan Martin are the project coordinators. Eric Uhlmann is the lead writer and Brian Nosek will supervise the project. Raphael (mail@raphael.rs) and Dan (dpmartin42@gmail.com) are your first points of contact for any question you may have. More information about the project itself, as well as a timeframe and further information are in our google document:

https://docs.google.com/document/d/1uCF5wmbcL90qvrk_J27fWAvDcDNrO9o_APkicwRkOKc/edit We will update this document over time but will also inform you via e-mail of major changes. At this point you may likely ask what the next steps are.

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

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

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

2. Analytical Approach Collection E-Mail

Dear \${m://FirstName},

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

Best regards,
Raphael, Dan, Eric and Brian

Follow this link to submit your analytical approach:

[\\${l://SurveyLink?d=Take the Survey}](#)

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

[\\${l://SurveyURL}](#)

[\\${l://OptOutLink?d=To%20opt%20out%20from%20the%20crowdstorming%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 crowdstorming 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 you and your team for making this Crowdstorming 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: Final Results

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

Summary of Methods

| Team | Method |
|------|---|
| 1 | We use a variety of different regressions. First, we use ordinary least squares with robust standard errors and control for various things such as height, weight, age. We also add in fixed effects for league country, position, club, and referee. In addition, we employ a logistic regression to compare with our OLS regressions. |
| 2 | Linear probability model, logistic regression |
| 3 | Multilevel Binomial Logistic Regression using bayesian inference. |
| 4 | Spearman correlation |
| 5 | Generalized linear mixed models |
| 6 | Linear Probability Model |
| 7 | Dirichlet process Bayesian clustering |
| 8 | - |
| 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
 10 players, an effect that could not be explained by the average racial biases of the
 referee's countries.

11 Professional soccer referees give more red cards (and fewer yellow cards) to
 12 darker-skinned players, but this behavior is not associated with prejudice levels
 in the referees' country-of-origin

13 There was statistical support for a unique bivariate relation between the skin tone
 14 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.

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

16 Our analysis supports the hypothesis that referees are more likely to give red
 17 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.

18 Whether the club of the player is controlled for is important for the results of the
 19 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".

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

22 Evidence from Poisson regression analysis indicates that darker skin tone soccer
 23 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.

24 After removing seven outliers –0.3% of the complete data set– a Bayesian
 25 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.

26 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.

27 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.

28 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.

29 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.

30 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.

31 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 5: Additional Analyses Regarding Research Question 1

We were interested to know whether subjective beliefs that the primary research hypothesis is true were related to the results a team obtained. 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 timepoints (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 research question 1, with the following magnitudes across the four timepoints (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 regarding the hypothesis for RQ1 at registration were weakly related to the teams' observed effect size of their final report ($\rho = 0.14 [-0.25, 0.49]$). However, beliefs changed considerably throughout the research process, and analysts' *post-analysis* belief in the hypothesis was likewise related to their effect estimate and lower bound ($\rho = 0.41$ and $\rho = 0.58$, respectively), also suggesting updating of beliefs based on empirical evidence. This

analysis does not take into account the extent to which the aggregate findings across teams influenced beliefs regarding individual results.

During the round-robin feedback phase, each analytical plan received ratings of peers' confidence that the approach was suitable. These ratings yielded a small correlation with final effect size estimates ($\rho = 0.10$ [-0.28, 0.46]). The final effect sizes from teams whose analytic approach received very positive confidence ratings (median OR = 1.31, MAD = 0.15) did not differ from those of teams who received lower peer evaluations (median OR = 1.28, MAD = 0.12).

A final question is whether the high level of dispersion we observe in estimated effect sizes (Fig. 1) is due to less expert teams obtaining different results from teams with greater statistical expertise. Relatedly, teams whose members have more quantitative expertise may show greater convergence in their estimated effect sizes. To examine these questions further, teams were dichotomized 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%). A greater number of teams with high ratings of general statistical expertise reported effects that were statistically significant (group 1: 78% significant, median OR = 1.39, MAD = 0.13; group 2: 68% significant, median OR = 1.30, MAD = 0.13).

Supplement 6: Additional Analyses Regarding Research Question 2a and 2b

Research question 2a examines whether national level implicit preferences for light vs. dark skin predict referee card decisions, which research question 2b does the same with explicit preferences. In total, 75% and 72% of respondents were unconfident to somewhat unconfident regarding how appropriate the data set was for answering either research question 2a or 2b, respectively. Only 32% of respondents felt the same way regarding the primary research questions. 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.

Additionally, 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. Because of this, we chose to not include the aggregated results for these research questions in the main text.

See Fig. S6 below for team's beliefs regarding the effects for research question 2a and 2b.

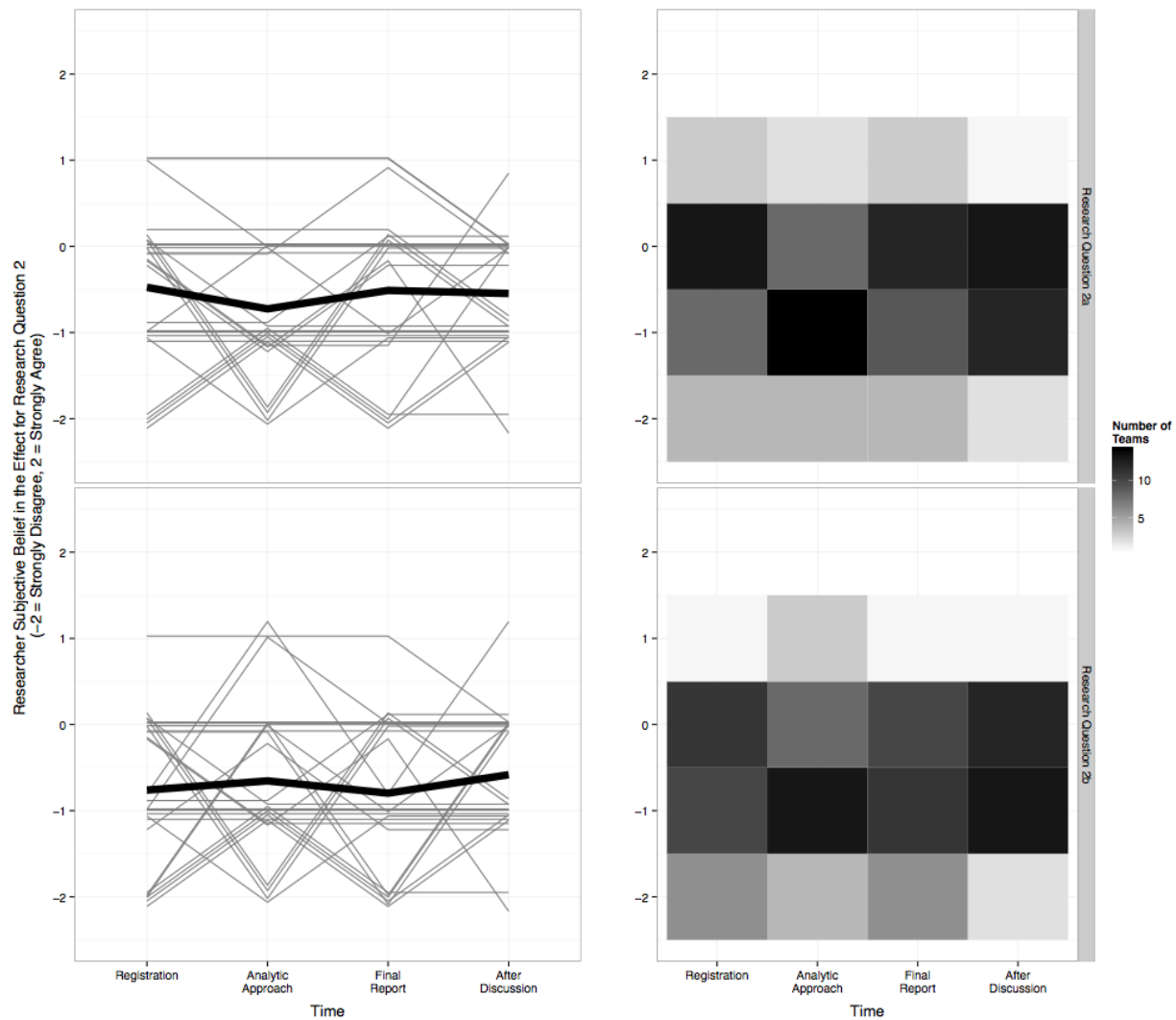


Fig. S6. The top panels reflect team leaders' beliefs regarding research question 2a (whether national level implicit preferences for light vs. dark skin predict referee red card decisions). The bottom two panels reflect team leader beliefs for research question 2b (whether national level explicit skin 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.

Supplement 7: Nuanced Beliefs about Research Question 1

In the fourth and final survey we administered items assessing more nuanced beliefs about research question 1 (i.e., 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*). The items, means, and standard deviations are reported below. 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.

| 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 |

Supplement 8: Author Contribution Forms from Analysis Teams

| Team | Name | Contribution |
|------|----------------|--|
| 1 | Nolan G. | Analysis and writing |
| 1 | Bryson | Analysis and writing |
| 2 | Garret | |
| 3 | Erikson | |
| 4 | Christopher R. | Analysis, interpretation, writing |
| 5 | Johannes | Coordinated project, planned analyses, conducted analyses, wrote report |
| 5 | Elmar | Coordinated project, planned analyses, discussed report |
| 5 | Christoph | Planned analyses, conducted analyses, discussed report |
| 5 | Andreas | Planned analyses, disaggregated data, conducted analyses, checked R script |
| 6 | Jonathan | Sole researcher |
| 7 | Silvia | |
| 8 | S. Amy | |
| 8 | D.M. | |
| 9 | Felix D. | Data Analysis, Writing |
| 9 | Moritz | Data Analysis, Writing |
| 10 | Daniel | Helped design analyses; conducted analyses; wrote the report |
| 10 | Maureen | Helped design analyses; conducted analyses |
| 10 | Ryan | Helped design analyses |
| 10 | Monica | Helped design analyses |
| 11 | Jason M. | Analyses and write-up |
| 11 | Martin F. | Analyses and write-up |
| 12 | Eli | |
| 13 | Alicia J. | Analysis plan, analysis, writing report |
| 13 | Thomas A. | Analysis plan, writing report |
| 14 | Anna | Analysis plan, analysis, writing report |
| 14 | Evelina | Analysis plan, analysis, writing report |
| 14 | Karin | Analysis plan, analysis, writing report |
| 14 | Magnus | Analysis plan, writing report |
| 15 | Michelangelo | Analyzed data; Wrote report |
| 15 | Egidio | Analyzed data |
| 15 | Pasquale | Analyzed data |
| 15 | Luca | Analyzed data |
| 15 | Anna | Analyzed data |
| 16 | Russ | All |
| 17 | Eric-Jan | Conceptualizing the analyses, writing |

| | | |
|----|-------------|--|
| 17 | Richard D. | Conceptualizing the analyses, conducting the analyses, writing |
| 18 | Maciej | |
| 20 | Felix | Collection of data on players' position; Data analysis; Interpretation of the results; Draft the final results |
| 20 | Kent | Collection of data on players' position; Interpretation of the results; Provide feedback on written drafts |
| 21 | Laetitia B. | coordination, feedback on other teams, and final writing |
| 21 | Lammertjan | Performing (and informing on) Tobit regressions, feedback on the other teams |
| 21 | Eric | Initial analyses |
| 21 | Bernard A. | Initial analyses and deciding on final analyses |
| 21 | Floor | Advise, input on analyses and writing, feedback on other teams, and fellow-coordination |
| 21 | Susanne | Advise, feedback on other teams |
| 23 | Tom | analysis coordination, analysis; writing up |
| 23 | Mathew H. | visualisation; writing up |
| 23 | TimH. | analysis, frequentist models; writing up |
| 23 | Colin | analysis, Bayesian models |
| 24 | Štěpán | |
| 25 | Seth | analysis plan, data preparation, analysis, report writing and editing |
| 25 | Kristin | analysis planning, analysis, report write-up |
| 26 | Feng | |
| 26 | Hadiya | |
| 27 | Shauna | Design and execution of analysis plan; write up. |
| 28 | Frederik | Data analysis, reporting of results |
| 28 | Fabia | Data analysis, reporting of results |
| 30 | Rickard | Single author |
| 31 | Sangsuk | data analysis, write up |
| 31 | Nathan | data analysis |
| 32 | Ismael | |

Supplement 9: IPython notebook visualisation of the dataset

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

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

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