Running head: MANY ANALYSTS, ONE DATASET

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

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.-J.[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

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

Abstract

Twenty-nine teams involving 61 analysts used the same dataset to address the same research question: whether soccer referees are more likely to give red cards to dark skin toned players than light skin toned players. Analytic approaches varied widely across teams, and estimated effect sizes ranged from 0.89 to 2.93 in odds ratio units, with a median of 1.31. Twenty teams (69%) found a statistically significant positive effect and nine teams (31%) observed a non-significant relationship. Crowdsourcing data analysis, a strategy by which numerous research teams are recruited to simultaneously investigate the same research question, makes transparent how variations in analytical choices affect results.

Keywords: crowdsourcing science, data analysis, scientific transparency

Many analysts, one dataset:

Making transparent how variations in analytical choices affect results

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

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

reported results and implications drive the impact of published articles; the analysis strategy is pushed to the background.

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

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

The primary research question tested in the crowdsourced project was whether soccer players with dark skin tone are more likely than light skin toned players to receive red cards from referees. The decision to give a player a red card results in the ejection of the player from the game and has severe consequences as it obliges his team to continue with one less player for the remainder of the match. Red cards are given for aggressive behavior such as a violent tackle, a foul intended to deny an opponent a clear goal scoring opportunity, hitting or spitting on an opposing player, or threatening and abusive language. However, despite a standard set of rules

and guidelines for both players and match officials, referee decisions are often fraught with ambiguity (e.g., was that an intentional foul or was the player only going for the ball?). It is inherently a judgment call on the part of the referee as to whether a player's behavior merits a red card.

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

Methods

The first three authors and last author posted a description of the project online (see S2 of the Supplementary Materials). This document included an overview of the research question, a description of the dataset and the planned timeline. The project was advertised via Brian Nosek's Twitter account, blogs of prominent academics, and word of mouth.

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.

Dataset. From a company for sports statistics, we obtained player demographics from all soccer players (N = 2,053) playing in the first male divisions of England, Germany, France, and Spain in the 2012-2013 season. 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 dataset 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 rescaled to be bounded by 0 (*very light skin*) and 1 (*very dark skin*) prior to the final analysis to ensure consistency among effect sizes between teams and to reflect the largest possible effect.

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, and for referees, their country of origin. For each dyad, data included the number of games referees and players encountered each other and the number of yellow and red cards awarded. The variables of age, club, and league-- which frequently change throughout a player's career-- were only available for players at the time of data collection, not at the time of receiving the particular red card sanctioning. To protect their identities given the sensitivity of the research topic, referees were anonymized and listed by a numerical identifier for each referee and for each country of origin. Importantly, our archival dataset provides the opportunity to estimate the magnitude of the relationship between variables (i.e., player skin tone and referee red card decisions), but does not offer the opportunity to identify causal relations.

Procedure. Analysts' subjective beliefs about the research hypothesis were tracked from the beginning to the end of the project. At registration we asked team leaders for their present opinion regarding the research question, by asking "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 S1 for further details). 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*. Each team received feedback from an average of about 5 other teams (M = 5.32, SD = 2.87).

The qualitative and quantitative feedback was aggregated into a single report and shared with all team members. As such, each team received peer review commentaries about their own and other teams' analysis strategies. Notably, these commentaries came from reviewers that were highly familiar with the dataset, yet at this point teams were unaware of others' results (see https://osf.io/evfts/ and https://osf.io/ic634/ for the complete survey and round-robin feedback). Each team therefore had the opportunity to learn from others' analytic approaches, and from the qualitative and quantitative feedback provided by peer reviewers, but did not have access to each others' estimated effect sizes. This phase offered opportunity to improve the quality of analyses and, if anything, ought to have promoted convergence in analysis strategies and outcomes.

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. The results of analyses involving analysts' subjective beliefs about the research hypothesis across time are reported in S5.

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). Limitations of the dataset were also discussed (S9). Finally, the first three authors and last 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 covariates league and club may be responsible for making some results appear non-significant. A debate emerged regarding whether the inclusion of these covariates was quantitatively defensible given that league and club were only available at the time of data collection and likely changed over the course of many players' careers (see 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 (S10). Additionally, we asked these teams to decide whether to keep their prior version or use the results from the updated analysis.[1] The results displayed in the manuscript reflect teams' choices of their final model. 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.

[1] One of the co-authors of the present 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 dataset that spanned multiple years, as the present one did. He further argued that these conditions rendered the decision to use these variables a major analytic mistake, not a defensible analytic choice. For more details see https://osf.io/2prib/

Results

Twenty-nine independent teams of researchers submitted analytical approaches and refined these throughout the crowdsourcing project. Table 1 shows each team's analytic technique, model specifications and reported effect size.[2] Analytic techniques ranged from simple linear regression techniques to complex multilevel regression techniques and Bayesian approaches. Teams also varied highly in their decisions regarding which covariates to include (see https://osf.io/sea6k/). Table 2 shows that the 29 teams used 21 unique combinations of covariates. Apart from the variable 'games', which was used by all teams, just one covariate (player position, 62%) was used in more than half of the analytic strategies and three were used in just one analysis. Two sets of covariates were used by three teams each, and four sets of covariates were used by two teams each. The remaining 15 teams used a unique combination of covariates.

What were the consequences of this variability in analytic approaches? Researchers' conclusions likewise 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 negative) to 2.93 (moderately positive) in odds ratio units (S1), with a median of 1.31. From a null hypothesis significance testing standpoint,

[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, and Rothstein (2009). Additionally, because the prevalence of red cards is so low, we make the "rare disease" assumption by assuming that the risk ratios yielded in analyses adopting a Poisson regression framework yield a fair approximation to the odds ratio (Viera, 2008).

twenty teams (69%) found a significant positive relationship and nine teams (31%) observed a non-significant relationship. No team reported a significant negative relationship.

-- Place Tables 1 and 2 about here –

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

Teams also varied in their approaches to handling the non-independence of players and referees, which resulted in variability regarding both median estimates and rates of significance. In total, 15 teams estimated a variance component for players and/or referees (12/15 significant, Median OR = 1.32, MAD = 0.12), eight teams used clustered standard errors (4/8 significant, Median OR = 1.28, MAD = 0.13), five teams did not account for this artifact (4/5 significant, Median OR = 1.39, MAD = 0.28), and one team used fixed effects for the referee variable (0/1 significant, OR = 0.89).

An important question is whether the variability in the analytic choices made and results found by each team (Fig. 1) simply results from teams with the greatest statistical expertise making different choices than the remaining teams. Relatedly, teams whose members have more quantitative expertise may show greater convergence in their estimated effect sizes. To examine

these questions further, we dichotomized teams into two groups using latent class analysis. The first group ($N = 9$) was more likely to have a team member who: had a PhD (100% vs. 53%), was professor at a university (100% vs. 37%), had taught a graduate statistics course more than twice (100% vs. 0%), and had at least one methodological/statistical publication (78% vs. 47%). 78% of teams with high ratings of general statistical expertise reported effects that were statistically significant (median OR = 1.39, MAD = 0.13) whereas 68% of teams with less expertise reported a significant effect (median OR = 1.30, MAD = 0.13). Further analyses of the effects of quantitative expertise on choice of statistical models is provided in S6. Note however that both teams high and comparatively lower in expertise exhibited considerable variability in whether they found a significant effect, and had a similar degree of dispersion in their effect size estimates. Thus, overall, statistical expertise did have some influence on analytic approaches and estimated effect sizes, but this is far from sufficient to explain the high variability in these choices or in the conclusions they supported.

An alternative approach to examining the role of expertise in the results found by each team is to look at how the peer-evaluations of each analytic approach were associated with the conclusions supported. During the round robin feedback phase, each analytical plan received ratings of peers' confidence regarding the suitability of the approach. The final effect sizes from teams whose analytic approach received very positive peer evaluations (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). Thus little evidence emerged that the variability in estimated effect sizes observed across teams was attributable to a subset of lower quality analyses.[3]

-- Figure 1 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 researchers' choices and assumptions in the analysis process. Independent teams estimated effects for the primary research question ranging from 0.89 to 2.93 in odds ratio units (1.0 indicates a null effect), with zero teams finding a negative effect, nine teams finding no significant relationship, and twenty teams finding a positive effect. If, as in virtually all other research projects, a single team had conducted the study, selecting randomly from the present teams, there would have been a 69% probability of reporting a positive result and a 31% probability of reporting a null effect.

This variability in results could not be accounted for by differences in expertise. Analysts with high and comparatively lower levels of quantitative expertise both exhibited high levels of variability in their estimated effect sizes. Further, analytic approaches that received highly

[3] This project also examined whether preferences for light vs. dark skin predict the red card decisions of referees from those countries. In brief, little to no evidence emerged that referee decisions were moderated by explicit or implicit skin tone preferences. Data for skin tone preferences was however not available from individual referees, only from referees' nation of origin, and the majority of analysts judged the available dataset to be inadequate to test this potential moderator. Detailed results are reported in S7.

favorable evaluations from peers showed the same variability in final effect sizes as analytic approaches that were less favorably rated.

The observed results from a complex dataset can be contingent on justifiable, but subjective, analytic decisions. This is striking because the present research question— the relationship between player skin tone and referee red card decisions— was clear with a limited number of predictors. Compared to many research questions in neuroscience, economics, and biology, this is a research problem of relatively modest complexity. And yet the process of translating this question from natural language to statistical models gave rise to many different assumptions and choices that clearly influenced the conclusions. This raises the possibility that many research projects contain hidden uncertainty due to the wide range of analytic choices available to the researchers.

In the present research, it seems unlikely that research teams would employ questionable research practices in order to achieve significance (Gelman & Loken, 2014; Simmons et al., 2011; Wagenmakers et al., 2012). Indeed, all teams knew that their process would be observed and public, and the perceived need to achieve a significant result for publishability was lessened by the nature of the project. Nonetheless, substantial variation is observed in outcomes based on the many reasonable decisions that analysts make, even given identical data and research questions. Uncertainty in interpreting research results is not just a function of statistical power or the presence of questionable research practices, it is also a function of the many reasonable decisions that researchers must make in order to conduct the research. This does not mean that data analysis and drawing research conclusions is a subjective enterprise with no connection to

reality. It does mean that many subjective decisions are part of the research process and can affect the outcomes. The best defense against subjectivity in science is to expose it. Transparency in data, methods, and process gives the rest of the community opportunity to see the decisions, question them, offer alternatives, and test these alternatives in further research.

Author Contribution Statement

The first and second author contributed equally to the project. EU proposed the idea of crowdsourcing data analysis and wrote the initial project outline. RS, EU, DPM, and BAN developed the research protocol. RS and EU developed the specific research question regarding skin tone influencing referee decisions. RS and DPM collected the referee decisions data and prepared the dataset for analysis. RS and DPM coordinated the different stages of the crowdsourcing process. All other authors worked in teams to analyze data, give feedback and produce individual reports. A detailed list of contributions for each team is provided in S8 of the Supplementary Materials. RS and DPM combined and analyzed the results of the different teams. EU outlined the paper and wrote the first draft of the abstract, introduction, and discussion. RS wrote the first draft of the methods and online supplement. RS and DPM wrote the first draft of the results section. DPM 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. Daniel P. Martin was supported by the Institute of Education Sciences, U.S. Department of Education (Grant No. R305B090002).

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	Variance component	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	Variance component	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	Variance component	Logistic	OR	1.48	1.20	1.84	1.48	1.20	1.84
10	Multilevel regression and logistic regression	3	Variance component	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	Variance component	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	Variance component	Logistic	OR	1.02	1.00	1.03	1.02	1.00	1.03
16	Hierarchical Poisson Regression	2	Variance component	Poisson	IRR	1.32	1.06	1.63	1.32	1.06	1.63
17	Bayesian logistic regression	2	Variance component	Logistic	OR	0.96	0.77	1.18	0.96	0.77	1.18
18	Hierarchical Bayes model	2	Variance component	Logistic	OR	1.10	0.98	1.27	1.10	0.98	1.27
20	Cross-classified multilevel negative binomial model	1	Variance component	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	Variance component	Logistic	OR	1.31	1.10	1.56	1.31	1.10	1.56
24	Multilevel logistic regression	3	Variance component	Logistic	OR	1.38	1.11	1.72	1.38	1.11	1.72
25	Multilevel logistic binomial regression	4	Variance component	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	Variance component	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	Variance component	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	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.

Team	Analytic Approach	OR
12	Zero–inflated Poisson regression	0.89
17	Bayesian logistic regression	0.96
15	Hierarchical log–linear modeling	1.02
10	Multilevel regression and logistic regression	1.03
18	Hierarchical Bayes model	1.10
31	Logistic regression	1.12
1	Ordinary least squares with robust standard errors, logistic regression	1.18
4	Spearman correlation	1.21
14	Weighted least squares regression with clustered standard errors	1.21
11	Multiple linear regression	1.25
30	Clustered robust binomial logistic regression	1.28
6	Linear Probability Model	1.28
26	Three–level hierarchical generalized linear modeling with Poisson sampling	1.30
3	Multilevel Binomial Logistic Regression using bayesian inference	1.31
23	Mixed model logistic regression	1.31
16	Hierarchical Poisson Regression	1.32
2	Linear probability model, logistic regression	1.34
5	Generalized linear mixed models	1.38
24	Multilevel logistic regression	1.38
28	Mixed effects logistic regression	1.38
32	Generalized linear models for binary data	1.39
8	Negative binomial regression with a log link analysis	1.39
20	Cross–classified multilevel negative binomial model	1.40
13	Poisson Multi–level modeling	1.41
25	Multilevel logistic binomial regression	1.42
9	Generalized linear mixed effects models with a logit link function	1.48
7	Dirichlet process Bayesian clustering	1.71
21	Tobit regression	2.88
27	Poisson regression	2.93

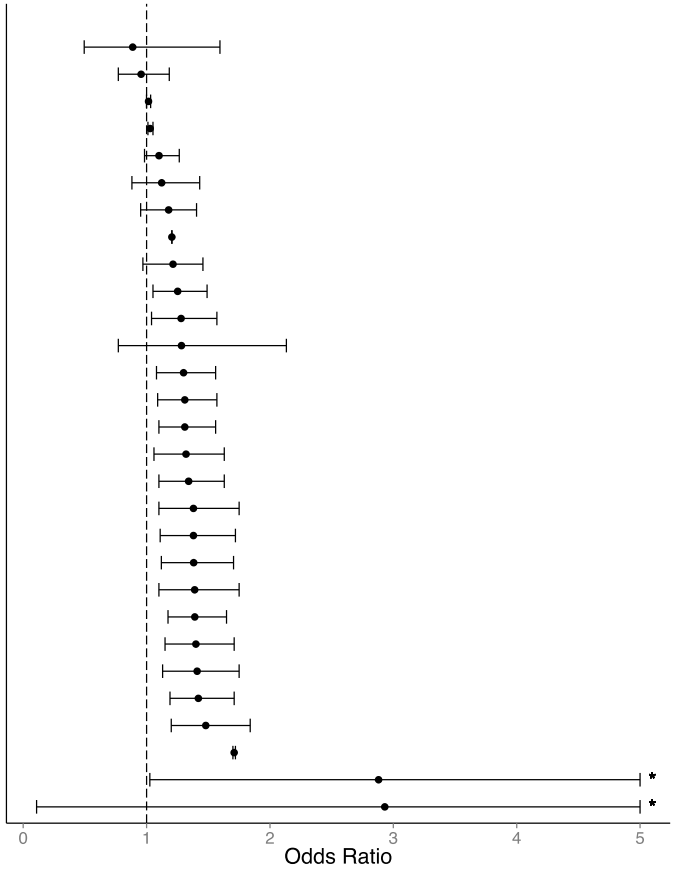

Fig. 1. Point estimates and 95% confidence intervals for analysis teams for the primary research question: Are soccer referees more likely to give red cards to dark skin toned players than light skin toned players? Note that the asterisks correspond to a truncated upper bound for Team 21 (11.47) and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

References

Babtie, A. C., Kirk, P., & Stumpf, M. P. H. (2014). Topological sensitivity analysis for systems biology. *Proceedings of the National Academy of Sciences of the United States of America*. http://doi.org/10.1073/pnas.1414026112

Bakker, M., van Dijk, A., & Wicherts, J. M. (2012). The rules of the game called Psychological Science. *Perspectives on Psychological Science: A Journal of the Association for Psychological Science*, *7*, 543–554. http://dx.doi.org/10.1177/1745691612459060

Bodenhausen, G. V. (1988). Stereotypic biases in social decision making and memory: testing process models of stereotype use. *Journal of Personality and Social Psychology*, *55*, 726–737. http://dx.doi.org/10.1037/0022-3514.55.5.726

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). Converting among effect sizes. In M. Borenstein, L. V. Hedges, J. P. T. Higgins, & H. R. Rothstein (Eds.), *Introduction to meta-analysis*. Chichester, UK: John Wiley & Sons.

Carp, J. (2012a). On the plurality of (methodological) worlds: Estimating the analytic flexibility of FMRI experiments. *Frontiers in Neuroscience*, *6*, 149. http://dx.doi.org/10.3389/fnins.2012.00149

Carp, J. (2012b). The secret lives of experiments: methods reporting in the fMRI literature. *NeuroImage*, *63*, 289–300. http://dx.doi.org/10.1016/j.neuroimage.2012.07.004

Correll, J., Park, B., Judd, C. M., & Wittenbrink, B. (2002). The police officer's dilemma: using ethnicity to disambiguate potentially threatening individuals. *Journal of Personality and Social Psychology*, *83*, 1314–1329. http://dx.doi.org/10.1037/0022-3514.83.6.1314

Ebrahim, S., Sohani, Z. N., Montoya, L., Agarwal, A., Thorlund, K., Mills, E. J., & Ioannidis, J. P. A. (2014). Reanalyses of randomized clinical trial data. *JAMA: The Journal of the American Medical Association*, *312*, 1024–1032. http://dx.doi.org/doi:10.1001/jama.2014.9646

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

Hugenberg, K., & Bodenhausen, G. V. (2003). Facing prejudice: implicit prejudice and the perception of facial threat. *Psychological Science*, *14*, 640–643. http://dx.doi.org/10.1046/j.0956-

Kim, J. W., & King, B. G. (2014). Seeing stars: Matthew effects and status bias in Major League Baseball umpiring. *Management Science*, *60*, 2619–2644. http://dx.doi.org/10.1287/mnsc.2014.1967

Krumholz, H. M., & Peterson, E. D. (2014). Open access to clinical trials data. *JAMA: The Journal of the American Medical Association*, *312*, 1002–1003. http://dx.doi.org/10.1001/jama.2014.9647

Maddox, K. B., & Chase, S. G. (2004). Manipulating subcategory salience: exploring the link between skin tone and social perception of Blacks. *European Journal of Social Psychology*, *34*, 533–546. http://dx.doi.org/10.1002/ejsp.214

Maddox, K. B., & Gray, S. A. (2002). Cognitive Representations of Black Americans: Reexploring the role of skin tone. *Personality and Social Psychology Bulletin*, *28*, 250–259. http://dx.doi.org/10.1177/0146167202282010

McCullough, B. D., McGeary, K. A., & Harrison, T. D. (2006). Do economics journal archives

promote replicable research? *Canadian Journal of Economics*, *41*, 1406–1420. http://dx.doi.org/10.1111/j.1540-5982.2008.00509.x

Parsons, C. A., Sulaeman, J., Yates, M. C., & Hamermesh, D. S. (2011). Strike three: Discrimination, incentives, and evaluation. *The American Economic Review*, *101*, 1410–1435. http://www.jstor.org/stable/23045903

Price, J., & Wolfers, J. (2010). Racial discrimination among NBA referees. *The Quarterly Journal of Economics*, *125*, 1859–1887.

Sakaluk, J. K., Williams, A. J., & Biernat, M. (2014). Analytic review as a solution to the misreporting of statistical results in Psychological Science. *Perspectives on Psychological Science: A Journal of the Association for Psychological Science*, *9*, 652–660. http://dx.doi.org/10.1177/1745691614549257

Sidanius, J., Pena, Y., & Sawyer, M. (2001). Inclusionary Discrimination: Pigmentocracy and patriotism in the Dominican Republic. *Political Psychology*, *22*, 827–851. http://dx.doi.org/10.1111/0162-895X.00264

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*, 1359–1366. http://dx.doi.org/10.1177/0956797611417632

Twine, F. W. (1998). *Racism in a racial democracy: The maintenance of White supremacy in Brazil*. New Brunswick, NJ: Rutgers University Press.

Viera, A. J. (2008). Odds ratios and risk ratios: what's the difference and why does it matter? *Southern Medical Journal*, *101*, 730–734.

http://dx.doi.org/10.1097/SMJ.0b013e31817a7ee4

Wagenmakers, E.-J., Wetzels, R., Borsboom, D., van der Maas, H. L. J., & Kievit, R. A. (2012). An agenda for purely confirmatory research. *Perspectives on Psychological Science*, *7*, 632–638. http://dx.doi.org/10.1177/1745691612463078

Wicherts, J. M., Borsboom, D., Kats, J., & Molenaar, D. (2006). The poor availability of psychological research data for reanalysis. *American Psychologist*, *61*, 726–728. http://dx.doi.org/10.1037/0003-066X.61.7.726