ASSOCIATION FOR
PSYCHOLOGICAL SCIENCE

Empirical Article

Advances in Methods and
Practices in Psychological Science
2018, Vol. 1(3) 337–356

Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/2515245917747646
www.psychologicalscience.org/AMPPS
SAGE

Many Analysts, One Data Set: Making Transparent How Variations in Analytic Choices Affect Results

R. Silberzahn[1], E. L. Uhlmann[2], D. P. Martin[3], P. Anselmi[4], F. Aust[5], E. Awtrey[6], Š. Bahník[7], F. Bai[8], C. Bannard[9], E. Bonnier[10], R. Carlsson[11], F. Cheung[12], G. Christensen[13], R. Clay[14], M. A. Craig[15], A. Dalla Rosa[4], L. Dam[16], M. H. Evans[17], I. Flores Cervantes[18], N. Fong[19], M. Gamez-Djokic[20], A. Glenz[21], S. Gordon-McKeon[22], T. J. Heaton[23], K. Hederos[24], M. Heene[25], A. J. Hofelich Mohr[26], F. Högden[5], K. Hui[27], M. Johannesson[10], J. Kalodimos[28], E. Kaszubowski[29], D. M. Kennedy[30], R. Lei[15], T. A. Lindsay[26], S. Liverani[31], C. R. Madan[32], D. Molden[33], E. Molleman[16], R. D. Morey[34], L. B. Mulder[16], B. R. Nijstad[16], N. G. Pope[35], B. Pope[36], J. M. Prenoveau[37], F. Rink[16], E. Robusto[4], H. Roderique[38], A. Sandberg[24], E. Schlüter[39], F. D. Schönbrodt[25], M. F. Sherman[37], S. A. Sommer[40], K. Sotak[41], S. Spain[42], C. Spörlein[43], T. Stafford[44], L. Stefanutti[4], S. Tauber[16], J. Ullrich[21], M. Vianello[4], E.-J. Wagenmakers[45], M. Witkowiak[46], S. Yoon[19], and B. A. Nosek[3,47]

[1]Organisational Behaviour, University of Sussex Business School; [2]Organisational Behaviour Area, INSEAD Asia Campus; [3]Department of Psychology, University of Virginia; [4]Department of Philosophy, Sociology, Education and Applied Psychology, University of Padua; [5]Department of Psychology, University of Cologne; [6]Department of Management, University of Cincinnati; [7]Department of Management, Faculty of Business Administration, University of Economics, Prague; [8]Department of Management and Marketing, Hong Kong Polytechnic University; [9]Department of Psychology, University of Liverpool; [10]Department of Economics, Stockholm School of Economics; [11]Department of Psychology, Linnaeus University; [12]School of Public Health, University of Hong Kong; [13]Berkeley Institute for Data Science, University of California, Berkeley; [14]Department of Psychology, College of Staten Island, City University of New York; [15]Department of Psychology, New York University; [16]Faculty of Economics and Business, University of Groningen; [17]Division of Neuroscience and Experimental Psychology, University of Manchester; [18]Westat, Rockville, Maryland; [19]Department of Marketing and Supply Chain Management, Temple University; [20]Department of Management and Organizations, Kellogg School of Management, Northwestern University; [21]Department of Psychology, University of Zurich; [22]Washington, D.C.; [23]School of Mathematics and Statistics, University of Sheffield; [24]Swedish Institute for Social Research (SOFI), Stockholm University; [25]Department of Psychology, Ludwig-Maximilians-Universität München; [26]College of Liberal Arts, University of Minnesota; [27]School of Management, Xiamen University; [28]College of Business, Oregon State University; [29]Department of Psychology, Federal University of Santa Catarina; [30]School of Business, University of Washington Bothell; [31]School of Mathematical Sciences, Queen Mary University of London; [32]School of Psychology, University of Nottingham; [33]Department of Psychology, Northwestern University; [34]School of Psychology, Cardiff University; [35]Department of Economics, University of Maryland; [36]Department of Economics, Brigham Young University; [37]Department of Psychology, Loyola University Maryland; [38]Rotman School of Management, University of Toronto; [39]Department of Social Sciences and Cultural Studies, Institute of Sociology, Justus Liebig University, Giessen; [40]United States Military Academy at West Point; [41]Department of Marketing and Management, SUNY Oswego; [42]John Molson School of Business, Concordia University; [43]Lehrstuhl für Soziologie, insb. Sozialstrukturanalyse, Otto-Friedrich-Universität Bamberg; [44]Department of Psychology, University of Sheffield; [45]Department of Psychological Methods, University of Amsterdam; [46]Poznań, Poland; and [47]Center for Open Science, Charlottesville, Virginia

Corresponding Authors:
R. Silberzahn, University of Sussex Business School, Jubilee Building, Brighton BN1 9SL, United Kingdom
E-mail: r.silberzahn@gmail.com

E. L. Uhlmann, INSEAD, Organisational Behaviour Area, 1 Ayer Rajah Ave., 138676 Singapore
E-mail: eric.luis.uhlmann@gmail.com

D. P. Martin, University of Virginia, Department of Psychology, 918 Emmet St. N, Charlottesville, VA 22903
E-mail: dpmartin42@gmail.com

B. A. Nosek, Center for Open Science, 210 Ridge McIntire Rd., Suite 500, Charlottesville, VA 22903-5083
E-mail: nosek@virginia.edu

Abstract

Twenty-nine teams involving 61 analysts used the same data set to address the same research question: whether soccer referees are more likely to give red cards to dark-skin-toned players than to light-skin-toned players. Analytic approaches varied widely across the teams, and the estimated effect sizes ranged from 0.89 to 2.93 (*Mdn* = 1.31) in odds-ratio units. Twenty teams (69%) found a statistically significant positive effect, and 9 teams (31%) did not observe a significant relationship. Overall, the 29 different analyses used 21 unique combinations of covariates. Neither analysts' prior beliefs about the effect of interest nor their level of expertise readily explained the variation in the outcomes of the analyses. Peer ratings of the quality of the analyses also did not account for the variability. These findings suggest that significant variation in the results of analyses of complex data may be difficult to avoid, even by experts with honest intentions. Crowdsourcing data analysis, a strategy in which numerous research teams are recruited to simultaneously investigate the same research question, makes transparent how defensible, yet subjective, analytic choices influence research results.

Keywords

crowdsourcing science, data analysis, scientific transparency, open data, open materials

Received 7/19/17; Revision accepted 11/17/17

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 revealing results from a research study. 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 analytic 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).

Researchers may understand this conceptually, but there is little appreciation for the implications in practice. In some cases, authors use a particular analytic strategy because it is the one they know how to use, rather than because they have a specific rationale for using it. Peer reviewers may comment on and suggest improvements to a chosen analytic strategy, but rarely do those comments emerge from working with the actual data set (Sakaluk, Williams, & Biernat, 2014). Moreover, it is not uncommon for peer reviewers to take the authors' analytic strategy for granted and comment exclusively on other aspects of the manuscript. More important, once an article is published, reanalyses and critiques of the chosen analytic strategy are slow to emerge and rare (Ebrahim et al., 2014; Krumholz & Peterson, 2014; McCullough, McGeary, & Harrison, 2006), in part because of the low frequency with which data are available for reanalysis (Wicherts, Borsboom, Kats, & Molenaar, 2006). The reported results and implications drive the impact of published articles; the analytic 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? In that case, the process of certifying a particular result on the basis of an idiosyncratic analytic strategy might be fraught with unrecognized uncertainty (Gelman & Loken, 2014), and research findings might be less trustworthy than they at first appear to be (Cumming, 2014). Had the authors made different assumptions, an entirely different result might have been observed (Babtie, Kirk, & Stumpf, 2014). In this article, we report an investigation that addressed the current lack of knowledge about how much diversity in analytic choice there can be when different researchers analyze the same data and whether such diversity results in different conclusions. Specifically, we report the impact of analytic decisions on research results obtained by 29 teams that analyzed the same data set to answer the same research question. The results of this project illustrate how researchers can vary in their analytic approaches and how results can vary according to these analytic choices.

Crowdsourcing Data Analysis: Skin Tone and Red Cards in Soccer

The primary research question tested in this crowdsourced project was whether soccer players with dark skin tone are more likely than those with light skin tone to receive red cards from referees.[1] The decision to give a player a red card results in the player's ejection from the game and has severe consequences because it obliges his team to continue with one fewer player for the remainder of the match. Red cards are given for aggressive behavior, such as a tackling violently, fouling with the intent to deny an opponent a clear goal-scoring opportunity, hitting or spitting on an opposing player, or using threatening and abusive language. However, despite a standard set of rules and guidelines for both players and match officials, referees' decision making is often fraught with ambiguity (e.g., it may not be obvious whether a player committed an intentional foul or was simply 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: Ambiguous behavior tends to be interpreted in line with prior attitudes and beliefs (Bodenhausen, 1988; Correll, Park, Judd, & Wittenbrink, 2002; Frank & Gilovich, 1988; Hugenberg & Bodenhausen, 2003). In societies as diverse as India, China, the Dominican Republic, Brazil, Jamaica, the Philippines, the United States, Chile, Kenya, and Senegal, light skin is seen as a sign of beauty, status, and social worth (Maddox & Chase, 2004; Maddox & Gray, 2002; Sidanius, Pena, & Sawyer, 2001; Twine, 1998). Negative attitudes toward persons with dark skin may lead a referee to interpret an ambiguous foul by such a person as a severe foul and, consequently, to give a red card (Kim & King, 2014; Parsons, Sulaeman, Yates, & Hamermesh, 2011; Price & Wolfers, 2010).

Consider for a moment how you would test this research hypothesis using a complex archival data set including referees' decisions across numerous leagues, games, years, referees, and players and a variety of potentially relevant control variables that you might or might not include in your analysis. Would you treat each red-card decision as an independent observation? How would you address the possibility that some referees give more red cards than others? Would you try to control for the seniority of the referee? Would you take into account whether a referee's familiarity with a player affects the referee's likelihood of assigning a red card? Would you look at whether players in some leagues are more likely to receive red cards compared with players in other leagues, and whether the proportion of players with dark skin varies across leagues and player positions? As these questions suggest, many analytic decisions are required. Moreover, for a given question, different decisions might be defensible and simultaneously have implications for the findings observed and the conclusions drawn. You and another researcher might make different judgment calls (regarding statistical method, covariates included, or exclusion rules) that, prima facie, are equally valid. This crowdsourced project examined the extent to which such good faith, subjective choices by different researchers analyzing a complex data set shape the reported results.

Table 1. Materials Available Online

Project stage and resource	URL
Stage 1	
Project page	https://osf.io/gvm2z/
Codebook	https://osf.io/9yh4x/
Stage 3	
Survey for teams to report their analytic approach	https://osf.io/yug9r/
Summary of each team's analytic approach	https://osf.io/3ifm2/
Stage 4	
Survey evaluating teams' analytic strategies	https://osf.io/evfts/
Round-robin feedback from the survey (in Qualtrics survey-software format)	https://osf.io/ic634/
Stage 5	
Report of all analyses	https://osf.io/qix4g
Stage 6a	
E-mail discussion of the analytic approaches	https://osf.io/8eg94/
Discussion on the appropriateness of the covariates	https://osf.io/2prib/
Stage 7	
Instructions for the peer evaluation	https://osf.io/8e7du/

Disclosures

Data, materials, and online resources

Further information on this study is available online as a project on the Open Science Framework (OSF). Table 1 provides an overview of the materials from each project stage that are available at OSF. The project's main folder at OSF (https://osf.io/gvm2z) provides links to all files, which include the data set (https://osf.io/fv8c3/) and a description of the included variables (https://osf.io/9yh4x/), a numeric overview of results by the various teams at the various project stages (https://osf.io/c9mkx/), graphical overviews of results at the various stages (https://osf.io/j2zth/), and the scripts to obtain each plot (https://osf.io/rgqtx/). The main folder also includes the manuscript for this article and a subarticle by each team detailing its analysis (https://osf.io/qix4g/).

The Supplemental Material available online (http://journals.sagepub.com/doi/suppl/10.1177/2515245917747646) includes a project description, notes on the research process, and the complete text of the surveys sent to the analysis teams. Further, the Supplemental Material documents the analytic approach taken by each team and indicates how these approaches were altered on the basis of peer feedback. In addition, the Supplemental Material includes an overview of results for the primary research question as well as additional analyses (including results for a second research

Project Stage	Work Package	Month																			
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	Building the Data Set	■																			
2	Recruitment and Initial Survey of Data Analysts	■	■																		
3	First Round of Data Analysis			■																	
4	Round-Robin Peer Evaluations				■																
5	Second Round of Data Analysis					■															
6a	Open Discussion and Debate, Further Analyses						■	■	■	■											
6b	Write-Up of Manuscript										■	■	■	■	■						
7	Internal Experts' Peer Review of Approaches															■	■	■	■		
	Revision of Manuscript																			■	■

Fig. 1. Overview of the project's stages.

question that initially was part of this project but was not pursued further because the raw data were inadequate). The Supplemental Material also discusses the limitations of the data set and of including player's club and league country as covariates and provides a link to an IPython notebook illustrating one team's analysis. Finally, the Supplemental Material includes the text of the survey of the analysts' familiarity with the different statistical techniques used and the survey of their assessment of other teams' analytic choices, as well as results of an exploratory analysis undertaken to determine whether convergence regarding the results obtained depended on the analytic approach taken.

Ethical approval

This research was conducted using publicly available archival data and according to ethical standards.

Stages of the Crowdsourcing Process

The project unfolded over several key stages. First, the unique data set used for this project was obtained, documented, and prepared for dissemination to participating analysts (Stage 1). Then, analysts were recruited to participate in the project (Stage 2). The first round of data analysis (Stage 3) was followed by round-robin peer evaluations of each analysis (Stage 4). The second round of data analysis (Stage 5) was followed by an initial discussion of results and debate, which led to further analyses (Stage 6a). When we tried to decide on a common conclusion while writing, editing, and reviewing the manuscript (Stage 6b), further questions emerged, and an internal peer review was started. In this review, each team's approach was evaluated by other analysts who were experts in that technique (Stage 7). The project then concluded with revision of this manuscript. During several of these stages, the analysts' subjective beliefs about the hypothesis being tested were assessed using questionnaires. The timeline of the project is summarized in Figure 1.

Stage 1: building the data set

From a company for sports statistics, we obtained demographic information on all soccer players (N = 2,053) who played in the first male divisions of England, Germany, France, and Spain in the 2012–2013 season. In addition, we obtained data about the interactions of those players with all referees (N = 3,147) whom they encountered across their professional careers. Thus, the interaction data for most players covered multiple seasons of play, from their first professional match until the time that the data were acquired, in June 2014. For players who were new in the 2012–2013 season, the data covered a single season. The data included the number of matches in which each player encountered each referee and our dependent variable, the number of red cards given to each player by each referee. The data set was made available as a list with 146,028 dyads of players and referees.

Photos for 1,586 of the 2,053 players were available from our source. Players for whom no photo was available tended to be relatively new players or those who had just moved up from a team in a lower league. The variable *player's skin tone* was coded by two independent raters blind to the research question. On the basis of the photos, the raters categorized the players on a 5-point scale ranging from 1 (*very light skin*) to 3 (*neither dark nor light skin*) to 5 (*very dark skin*), and these ratings correlated highly ($r = .92$, $\rho = .86$). 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 of effect sizes across the teams of analysts. The raw ratings were rescaled to 0, .25, .50, .75, and 1 to create this new scale.

A variety of potential independent variables were included in the data set (for the complete codebook, see https://osf.io/9yh4x). The data included players' typical position, weight, and height and referees' country of origin. For each dyad, the data included the number of games in which the referee and player encountered each other and the number of yellow and red cards awarded to the player. The records indicated players' ages, clubs, and leagues—which frequently change throughout players' careers—at the time of data collection, not at the specific times the red cards were received (see Table 2 for a summary of some of the player variables). Given the sensitivity of the research topic, referees' identities were protected by anonymization; each referee and each country of referees' origin was assigned a numerical identifier. Our archival data set provided the opportunity to estimate the magnitude of the relationship between player's skin tone and number of red cards received, but did not offer the opportunity to identify causal relations between these variables.

Table 2. Descriptive Statistics for Some of the Player Variables

Variable	Statistic
Height (cm)	$M = 181.74$ ($SD = 6.69$)
Weight (kg)	$M = 75.64$ ($SD = 7.10$)
Number of games	$M = 71.13$ ($SD = 36.17$)
Number of yellow cards	$M = 27.41$ ($SD = 24.08$)
Number of red cards	$M = 0.89$ ($SD = 1.26$)
League country	
England	$n = 564$ players
France	$n = 533$ players
Germany	$n = 489$ players
Spain	$n = 467$ players
Skin color	
0 (very light skin)	Rater 1: $n = 626$ players Rater 2: $n = 451$ players
.25	Rater 1: $n = 551$ players Rater 2: $n = 693$ players
.50	Rater 1: $n = 170$ players Rater 2: $n = 174$ players
.75	Rater 1: $n = 140$ players Rater 2: $n = 141$ players
1 (very dark skin)	Rater 1: $n = 98$ players Rater 2: $n = 126$ players
Not available	Rater 1: $n = 468$ players Rater 2: $n = 468$ players
Player position	
Attacking midfielder	$n = 149$ players
Center back	$n = 281$ players
Center forward	$n = 227$ players
Center midfielder	$n = 84$ players
Defensive midfielder	$n = 204$ players
Goalkeeper	$n = 196$ players
Left fullback	$n = 136$ players
Left midfielder	$n = 86$ players
Left winger	$n = 59$ players
Not available	$n = 367$ players
Right fullback	$n = 126$ players
Right midfielder	$n = 75$ players
Right winger	$n = 63$ players

Stage 2: recruitment and initial survey of data analysts

The first three authors and last author posted a description of the project online (see Supplement 1 in the Supplemental Material available online). This document included an overview of the crowdsourcing project, a description of the data set, and the planned timeline. The project was advertised via Brian Nosek's Twitter account, blogs of prominent academics, and word of mouth.

Seventy-seven researchers expressed initial interest in participating and were given access to the OSF project page to obtain the data. Individual analysts were welcome to form teams, and most did. For the sake of consistency, in this article we use the term *team* also for those few individuals who chose to work on their own. Thirty-three teams submitted a report in the first round (Stage 3), and 29 teams submitted a final report. The analysis presented in this article focuses on the submissions of those 29 teams. In total, the final project involved 61 data analysts plus the four authors who organized the project. A demographic survey revealed that the team leaders worked in 13 different countries and came from a variety of disciplinary backgrounds, including psychology, statistics, research methods, economics, sociology, linguistics, and management. At the time that the first draft of this manuscript was written, 38 of the 61 data analysts (62%) held a Ph.D. (62%), and 17 (28%) had a master's degree. The analysts came from various ranks and included 8 full professors (13%), 9 associate professors (15%), 13 assistant professors (21%), 8 postdocs (13%), and 17 doctoral students (28%). In addition, 27 participants (44%) had taught at least one undergraduate statistics course, 22 (36%) had taught at least one graduate statistics course, and 24 (39%) had published at least one methodological or statistical article.

In addition to collecting data on the analysts' demographic characteristics, we asked the team leaders for their opinion regarding the research question. For example, using a 5-point Likert scale from 1 (*very unlikely*) to 5 (*very likely*), they answered the question "How likely do you think it is that soccer referees tend to give more red cards to dark-skinned players?" This question was asked again at several points in the research project to track beliefs over time: when analysts submitted their analytic approach, when they submitted their final analyses, and after the group discussion of all the teams' results.

Stage 3: first round of data analysis

After registering and answering the subjective-beliefs survey for the first time, the research teams were given access to the data. Each team then decided on its own analytic approach to test the primary research question and analyzed the data independently of the other teams (see Item 1 in Supplement 2 for further details). Then, via a standardized Qualtrics survey, the teams submitted to the coordinators structured summaries of their analytic approach, including information about data transformations, exclusions, covariates, the statistical techniques used, the software used, and the results (see Supplement 3 for the text of the survey materials sent to the team leaders; the Qualtrics files and descriptions of the individual teams' analytic approaches are available at https://osf.io/yug9r/ and https://osf.io/3ifm2/, respectively). The teams were also asked about their beliefs regarding the primary research question.

Stage 4: round-robin peer evaluations of overall analysis quality

For the first three stages of the project, the teams were expected to work independently of each other. However, beginning with Stage 4, they were encouraged to discuss and debate their respective approaches to the data set. In Stage 4, after descriptions of the results were removed, 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 the analysts were instructed to provide feedback on at least the first three approaches that they examined. They were asked to provide qualitative feedback as well as a confidence rating ("How confident are you that the described approach below is suitable for analyzing the research questions?") on a 7-point scale from 1 (*unconfident*) to 7 (*confident*). On average, each team received feedback from about five 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. Thus, each team received peer-review commentaries about their own analytic strategy and the other teams' analytic strategies. Notably, these commentaries came from reviewers who were highly familiar with the data set, yet at this point the teams were unaware of others' results (for the complete survey and round-robin feedback, see https://osf.io/evfts/ and https://osf.io/ic634/, respectively). 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 other teams' estimated effect sizes. This phase offered the teams an opportunity to improve the quality of their analyses and, if anything, ought to have promoted convergence in analytic strategies and outcomes.

Stage 5: second round of data analysis

Following the peer review, the teams had the opportunity to change their analytic strategies and draw new conclusions (see Supplement 4 for a list of the initial and final approaches of each team). They submitted formal reports in a standardized format and also filled out a standardized questionnaire similar to that used in Stage 2. Their subjective beliefs about the primary research question were also assessed in this questionnaire. Notably, the teams were not forced to present a single effect size without robustness checks. Rather, they were encouraged to present results in the way they would in a published article, with formal Method and Results sections. Some teams adopted a model-building approach and reported the results of the model that they felt was the most appropriate one. The fact that not every team did this represents yet another subjective, yet defensible analytic choice. All the teams' reports are available on the OSF, at https://osf.io/qix4g. Supplement 5 presents a brief summary of each team's methods and a one-sentence description of each team's findings, and Supplement 11 provides an illustration of one team's process.

Stage 6: open discussion and debate, further analyses, and drafting a report on the project

After the formal analysis, the reports were compiled and uploaded to the OSF project. A summary e-mail sent to all the teams invited them to review the reports and discuss as a group the analytic strategies and what to conclude regarding the primary research question. Team members engaged in a substantive e-mail discussion regarding the variation in findings and analytic strategies (the full text of this discussion can be found at https://osf.io/8eg94/). For example, one team found a strong influence of five outliers on their results. Other teams performed additional analyses to investigate whether their results were similarly driven by a few outliers (interestingly, they were not). Limitations of the data set were also discussed (see Supplement 9). At this stage, a final assessment of subjective beliefs was conducted; this survey also presented a series of possible statements summarizing the outcome of this project and asked the analysts to rate their agreement with each one. The first three authors and last author then wrote a first draft of this manuscript, and all the team members were invited to jointly edit and extend the draft using Google Docs.

When the analysts scrutinized each other's results, it became apparent that differences in results may have been due not only to variations in statistical models, but also to variations in the choice of covariates. Doing a preliminary reanalysis, the leader of Team 10 discovered that including league and club as covariates may have been responsible for the nonsignificant results obtained by some teams. A debate emerged regarding whether the inclusion of these covariates was quantitatively defensible given that the data on league and club were available for the time of data collection only and these variables likely changed over the course of many players' careers (see the discussion at https://osf.io/2prib/). The project coordinators therefore asked the 10 teams that had included these variables in their final models to rerun their models without these covariates (see Supplement 10). Additionally, these teams were allowed to decide whether they wanted to revise their final models to exclude these covariates.[2] The results reported in this article reflect the teams' choices of their final models.

Stage 7: more granular peer assessments of analysis quality

The discussion and debate about analytic choices motivated the project coordinators to initiate a more fine-grained assessment of each of the final analyses to identify potential flaws that might account for any variability in the reported results. Therefore, after the methods and results of all the teams were known, the analysts participated in an additional internal peer-review assessment. First, they indicated their familiarity with each approach used by each team, on a 5-point scale ranging from 1 (*very unfamiliar*) to 5 (*very familiar*; see Supplement 12). For some techniques, most of the analysts responded "familiar" or "very familiar" (e.g., 34 in the case of multiple regression). For other techniques, relatively few analysts did so (e.g., 3 in the case of Bayesian clustering with the Dirichlet process). On the basis of their expertise, the coordinators then assigned each analyst one to three analytic strategies to assess in greater depth (i.e., strategies involving techniques that the analyst reported being familiar or very familiar with). No researcher was assigned to review the approach of his or her own team.

From comments the analysts made in the earlier rounds of analysis (Stages 3–6), the coordinators derived a list of seven potential statistical concerns regarding analytic decisions that were made (see Supplement 13). For example, an analysis may have unnecessarily excluded a large number of cases or may not have adequately accounted for the number of games played. The analysts were asked to report whether the assigned analytic strategies had failed to take into account each of these seven issues (on a 5-point scale ranging from 1, *strongly disagree*, to 5, *strongly agree*). Note that lower scores indicated that more obstacles were avoided, and higher scores indicated that more issues were left unaddressed. For each assigned strategy, the survey also included an open-ended question asking whether there was an additional analytic issue that might have biased the results, and another item asked the analysts to rate their agreement that this additional issue affected the validity of the approach (same 5-point scale). The final question asked the analysts to rate how convinced they were that the approach in question successfully addressed most of the potential statistical concerns (1= *very unconvinced*, 5 = *very convinced*).

Main Findings From the Project

How much did results vary between different teams using the same data to test the same hypothesis?

Table 3 shows each team's final analytic technique, model specifications for treatment of nonindependence, and reported effect size.[3] The analytic techniques chosen ranged from simple linear regression to complex multilevel regression and Bayesian approaches. The teams also varied greatly in their decisions regarding which covariates to include (see https://osf.io/sea6k/ for the rationales the teams provided). Table 4 shows that the 29 teams used 21 unique combinations of covariates. Apart from the variable games (i.e., the number of games played under a given referee, which was used by all the teams, just one covariate (player position, 69%) was used in more than half of the teams' analyses, and three were used in just one analysis. Three teams chose to use no covariates, and another 3 teams included player position as the only covariate in their analysis. Four sets of variables were used by 2 teams each, and each of the remaining 15 teams used a unique combination of covariates.

Table 3. Analytic Approaches and Results for Each Team

Team	Distribution	Treatment of nonindependence	Number of covariates	Analytic approach	OR
1	Linear	Clustered standard errors	7	Ordinary least squares regression with robust standard errors, logistic regression	1.18 [0.95, 1.41]
6	Linear	Clustered standard errors	6	Linear probability model	1.28 [0.77, 2.13]
14	Linear	Clustered standard errors	6	Weighted least squares regression with clustered standard errors	1.21 [0.97, 1.46]
4	Linear	None	3	Spearman correlation	1.21 [1.20, 1.21]
11	Linear	None	4	Multiple linear regression	1.25 [1.05, 1.49]
10	Linear	Variance component	3	Multilevel regression and logistic regression	1.03 [1.01, 1.05]
2	Logistic	Clustered standard errors	6	Linear probability model, logistic regression	1.34 [1.10, 1.63]
30	Logistic	Clustered standard errors	3	Clustered robust binomial logistic regression	1.28 [1.04, 1.57]
31	Logistic	Clustered standard errors	6	Logistic regression	1.12 [0.88, 1.43]
32	Logistic	Clustered standard errors	1	Generalized linear models for binary data	1.39 [1.10, 1.75]
8	Logistic	None	0	Negative binomial regression with a log link	1.39 [1.17, 1.65]
15	Logistic	None	1	Hierarchical log-linear modeling	1.02 [1.00, 1.03]
3	Logistic	Variance component	2	Multilevel logistic regression using Bayesian inference	1.31 [1.09, 1.57]
5	Logistic	Variance component	0	Generalized linear mixed models	1.38 [1.10, 1.75]
9	Logistic	Variance component	2	Generalized linear mixed-effects models with a logit link	1.48 [1.20, 1.84]
17	Logistic	Variance component	2	Bayesian logistic regression	0.96 [0.77, 1.18]
18	Logistic	Variance component	2	Hierarchical Bayes model	1.10 [0.98, 1.27]
23	Logistic	Variance component	2	Mixed-model logistic regression	1.31 [1.10, 1.56]
24	Logistic	Variance component	3	Multilevel logistic regression	1.38 [1.11, 1.72]
25	Logistic	Variance component	4	Multilevel logistic binomial regression	1.42 [1.19, 1.71]
28	Logistic	Variance component	2	Mixed-effects logistic regression	1.38 [1.12, 1.71]
21	Miscellaneous	Clustered standard errors	3	Tobit regression	2.88 [1.03, 11.47]
7	Miscellaneous	None	0	Dirichlet-process Bayesian clustering	1.71 [1.70, 1.72]
12	Poisson	Fixed effect	2	Zero-inflated Poisson regression	0.89 [0.49, 1.60]
27	Poisson	None	1	Poisson regression	2.93 [0.11, 78.66]
13	Poisson	Variance component	1	Poisson multilevel modeling	1.41 [1.13, 1.75]
16	Poisson	Variance component	2	Hierarchical Poisson regression	1.32 [1.06, 1.63]
20	Poisson	Variance component	1	Cross-classified multilevel negative binomial model	1.40 [1.15, 1.71]
26	Poisson	Variance component	6	Hierarchical generalized linear modeling with Poisson sampling	1.30 [1.08, 1.56]

Note: Values in brackets are 95% confidence intervals (CIs). Each team's observed effect size is presented in this table as an odds ratio, but some of the teams reported effect sizes in other units that were converted to odds ratios. Those originally reported effect sizes are as follows—Team 4: Cohen's *d* = 0.10, 95% CI = [0.10, 0.10]; Team 11: Cohen's *d* = 0.12, 95% CI = [0.03, 0.22]; Team 10: β = 0.01, 95% CI = [0.00, 0.01]; Team 21: β = 0.28, 95% CI = [0.01, 0.56]; Team 12: incidental risk ratio (IRR) = 0.89, 95% CI = [0.49, 1.60]; Team 27: IRR = 2.93, 95% CI = [0.11, 78.66]; Team 13: IRR = 1.41, 95% CI = [1.13, 1.75]; Team 16: IRR = 1.32, 95% CI = [1.06, 1.63]; Team 20: IRR = 1.40, 95% CI = [1.15, 1.71]; Team 26: IRR = 1.30, 95% CI = [1.08, 1.56].

What were the consequences of this variability in analytic approaches?

Figure 2 shows each team's estimated effect size, along with its 95% confidence interval (CI). As this figure and Table 3 show, the estimated effect sizes ranged from 0.89 (slightly negative) to 2.93 (moderately positive) in odds-ratio (OR) units; the median estimate was 1.31. The confidence intervals for many of the estimates overlap, which is expected because they are based on

Table 4. Covariates Included by Each Team

	Team																												
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	Percentage of teams
Player position	X	X	X			X			X	X	X	X	X	X					X	X	X	X	X	X		X	X	X	X	69%
Player's height	X	X		X		X								X						X		X	X	X			X	X		38%
Player's weight	X	X		X		X						X		X								X	X	X			X	X		38%
Player's league country[a]	X										X				X						X			X		X		X		24%
Player's age	X					X					X			X		X				X				X						24%
Goals scored by player		X									X					X							X		X			X		21%
Player's club	X					X								X														X		14%
Referee's country		X	X	X					X																					14%
Referee	X					X								X																10%
Player's number of victories		X								X														X						10%
Number of cards received by player																	X	X												7%
Player																		X												3%
Number of cards awarded by referee																	X													3%
Number of draws										X																				3%
Number of covariates	7	6	2	3	0	6	0	0	2	3	4	2	1	6	1	2	2	2	1	3	2	3	4	6	1	2	3	6	1	

Note: The covariates are listed in order of the frequency with which they were included in teams' analytic approaches. The number of games a player had played was essential to the analysis, was used by all teams, and is thus not listed here as a separate covariate.

[a] Team 9 mistakenly labeled referee's country as league country.

Team	Analytic Approach	Odds Ratio
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	OLS Regression With Robust Standard Errors, Logistic Regression	1.18
4	Spearman Correlation	1.21
14	WLS 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	Hierarchical Generalized Linear Modeling With Poisson Sampling	1.30
3	Multilevel 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	1.39
20	Cross-Classified Multilevel Negative Binomial Model	1.40
13	Poisson Multilevel Modeling	1.41
25	Multilevel Logistic Binomial Regression	1.42
9	Generalized Linear Mixed-Effects Models With a Logit Link	1.48
7	Dirichlet-Process Bayesian Clustering	1.71
21	Tobit Regression	2.88
27	Poisson Regression	2.93

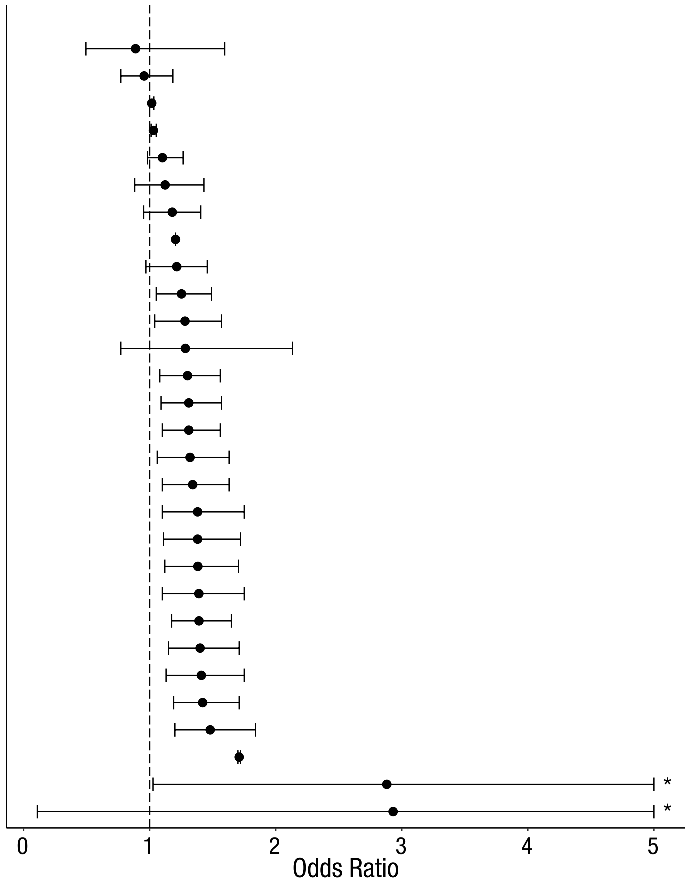

Fig. 2. Point estimates (in order of magnitude) and 95% confidence intervals for the effect of soccer players' skin tone on the number of red cards awarded by referees. Reported results, along with the analytic approach taken, are shown for each of the 29 analytic teams. The teams are ordered so that the smallest reported effect size is at the top and the largest is at the bottom. The asterisks indicate upper bounds that have been truncated to increase the interpretability of the plot; the actual upper bounds of the confidence intervals were 11.47 for Team 21 and 78.66 for Team 27. OLS = ordinary least squares; WLS = weighted least squares.

the same data. Twenty teams (69%) found a significant positive relationship, $p < .05$, and nine teams (31%) found a nonsignificant relationship. No team reported a significant negative relationship.

What were the results obtained with the different types of analytic approaches used?

Teams that employed logistic or Poisson models tended to report estimates that were larger than those of teams that used linear models (see the effect sizes in Fig. 3, in which the teams are clustered according to the distribution used for analyses). Fifteen teams used logistic models, and 11 of these teams found a significant effect (median OR = 1.34; median absolution deviation, or MAD = 0.07). Six teams used Poisson models, and 4 of these teams found a significant effect (median OR = 1.36, MAD = 0.08). Of the 6 teams that used linear models, 3 found a significant effect (median OR = 1.21, MAD = 0.05). The final 2 teams used models classified as miscellaneous, and both of these teams reported significant effects (ORs = 1.71 and 2.88, respectively).

The teams also varied in their approaches to handling the nonindependence of players and referees, and this variability also influenced both median estimates of the effect size and the rates of significant results. In total, 15 teams estimated a fixed effect or variance component for players, referees, or both; 12 of these teams reported significant effects (median OR = 1.32, MAD = 0.12). Eight teams used clustered standard errors, and 4 of these teams found significant effects (median OR = 1.28, MAD = 0.13). An additional 5 teams did not account for this artifact, and 4 of these teams reported significant effects (median OR = 1.39, MAD = 0.28). The remaining team used fixed effects for the referee variable and reported a nonsignificant result (OR = 0.89).

Did the analysts' beliefs regarding the hypothesis change over time?

Analysts' subjective beliefs about the research hypothesis were assessed four times during the project: at initial registration (i.e., before they had received the data), after they had accessed the data and submitted their analytic approach, at the time final analyses were submitted, and after a group discussion of all the teams' approaches and results. Responses were centered at 0 for analyses to increase interpretability (thus, the range was from −2, for *very unlikely*, to +2, for *very likely*). Subjective beliefs changed over time (see Fig. 4). At initial registration, there was slight agreement, on average, that the number of red cards was positively related to players' skin tone, yet opinions varied greatly (M = 0.46, SD = 0.84). At the next assessment, the slight initial agreement had turned into slight disagreement (M = −0.61, SD = 0.88). When the teams submitted their final analyses, they again slightly agreed that there was a relationship; the magnitude of agreement was similar to what it had been initially, but again there was substantial variability (M = 0.61, SD = 1.20). Finally, after the group discussion, overall agreement increased slightly, and, notably, variability decreased (M = 0.75, SD = 0.70), which suggests some convergence in beliefs over time. The right-hand plot in Figure 4 shows the number of teams who endorsed each level of agreement at each of the four assessments. Beliefs converged over time, such that that toward the end of the project, more teams agreed that skin tone affected the number of red cards received.

The fourth and final survey assessed more nuanced beliefs about the primary research question. All the analysts were asked to respond individually to this survey. The new items included, for example, "The effect is positive and due to referee bias" and "There is little evidence for an effect." The analysts responded to these items on scales ranging from 1 (*strongly disagree*) to 7 (*strongly agree*). Summary statistics for this survey are reported in Table 5. By the end of the project, a majority of the analysts agreed that the data showed a positive relationship between the number of red cards received and players' skin tone but were unclear regarding the underlying mechanism. The level of agreement was highest (78%) for the statement "The effect is positive and the mechanism is unknown" (M = 5.32, SD = 1.47).

What was the association between analysts' subjective beliefs regarding the hypothesis and the results obtained?

Of particular interest was whether subjective beliefs about the truth of the primary research hypothesis were related to the results the teams obtained. One might anticipate a confirmation bias, that is, that the analysts found what they initially expected to find. Alternatively, they might have rationally updated their beliefs in response to the empirical results they obtained, even if those results contradicted their initial expectations.

The team leaders' self-reports regarding the primary research question at each of the four assessments of beliefs were correlated with the final reported effect size, and the magnitude of this association increased across time: $\rho = .14$, 95% CI = [−.25, .49]; $\rho = -.20$, 95% CI = [−.53, .19]; $\rho = .43$, 95% CI = [.07, .69]; and $\rho = .41$,

Team	Analytic Approach	Distribution	Odds Ratio
10	Multilevel Regression and Logistic Regression	Linear	1.03
1	OLS Regression With Robust Standard Errors, Logistic Regression	Linear	1.18
4	Spearman Correlation	Linear	1.21
14	WLS Regression With Clustered Standard Errors	Linear	1.21
11	Multiple Linear Regression	Linear	1.25
6	Linear Probability Model	Linear	1.28
17	Bayesian Logistic Regression	Logistic	0.96
15	Hierarchical Log-Linear Modeling	Logistic	1.02
18	Hierarchical Bayes Model	Logistic	1.10
31	Logistic Regression	Logistic	1.12
30	Clustered Robust Binomial Logistic Regression	Logistic	1.28
3	Multilevel Logistic Regression Using Bayesian Inference	Logistic	1.31
23	Mixed-Model Logistic Regression	Logistic	1.31
2	Linear Probability Model, Logistic Regression	Logistic	1.34
5	Generalized Linear Mixed Models	Logistic	1.38
24	Multilevel Logistic Regression	Logistic	1.38
28	Mixed-Effects Logistic Regression	Logistic	1.38
32	Generalized Linear Models for Binary Data	Logistic	1.39
8	Negative Binomial Regression With a Log Link	Logistic	1.39
25	Multilevel Logistic Binomial Regression	Logistic	1.42
9	Generalized Linear Mixed-Effects Models With a Logit Link	Logistic	1.48
7	Dirichlet-Process Bayesian Clustering	Misc	1.71
21	Tobit Regression	Misc	2.88
12	Zero-Inflated Poisson Regression	Poisson	0.89
26	Hierarchical Generalized Linear Modeling With Poisson Sampling	Poisson	1.30
16	Hierarchical Poisson Regression	Poisson	1.32
20	Cross-Classified Multilevel Negative Binomial Model	Poisson	1.40
13	Poisson Multilevel Modeling	Poisson	1.41
27	Poisson Regression	Poisson	2.93

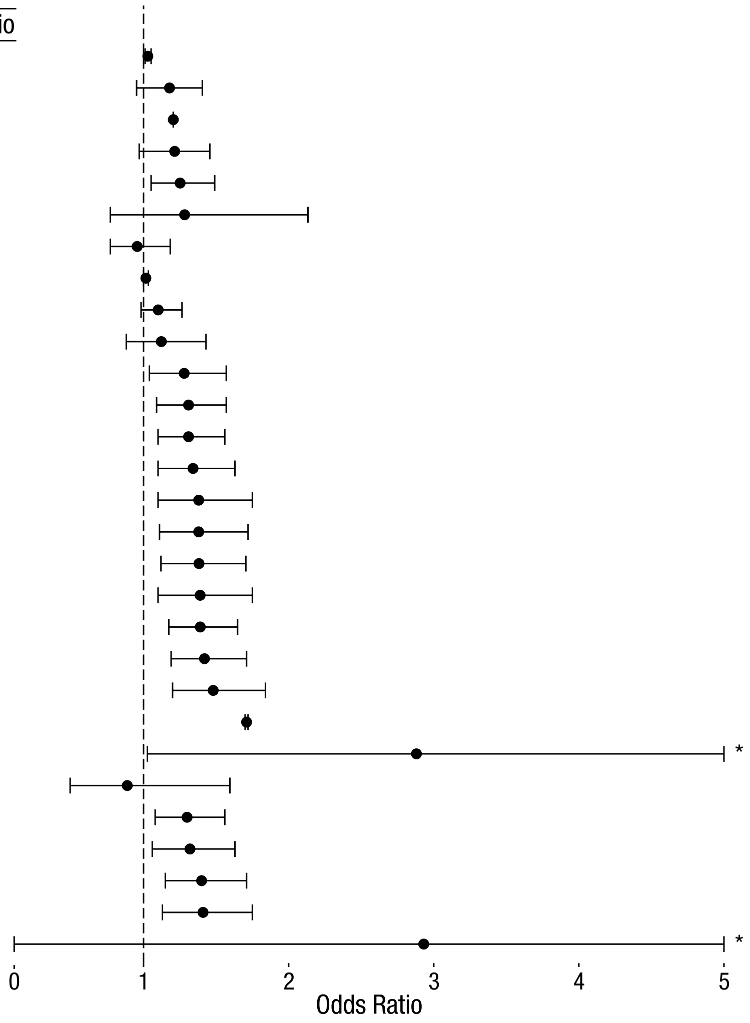

Fig. 3. Point estimates (clustered by analytic approach) and 95% confidence intervals for the effect of soccer players' skin tone on the number of red cards awarded by referees. Reported results, along with the analytic approach taken, are shown for each of the 29 analytic teams. The teams are clustered according to the distribution used in their analyses; within each cluster, the teams are listed in order of the magnitude of the reported effect size, from smallest at the top to largest at the bottom. The asterisks indicate upper bounds that have been truncated to increase the interpretability of the plot (see Fig. 2). OLS = ordinary least squares; WLS = weighted least squares; Misc = miscellaneous.

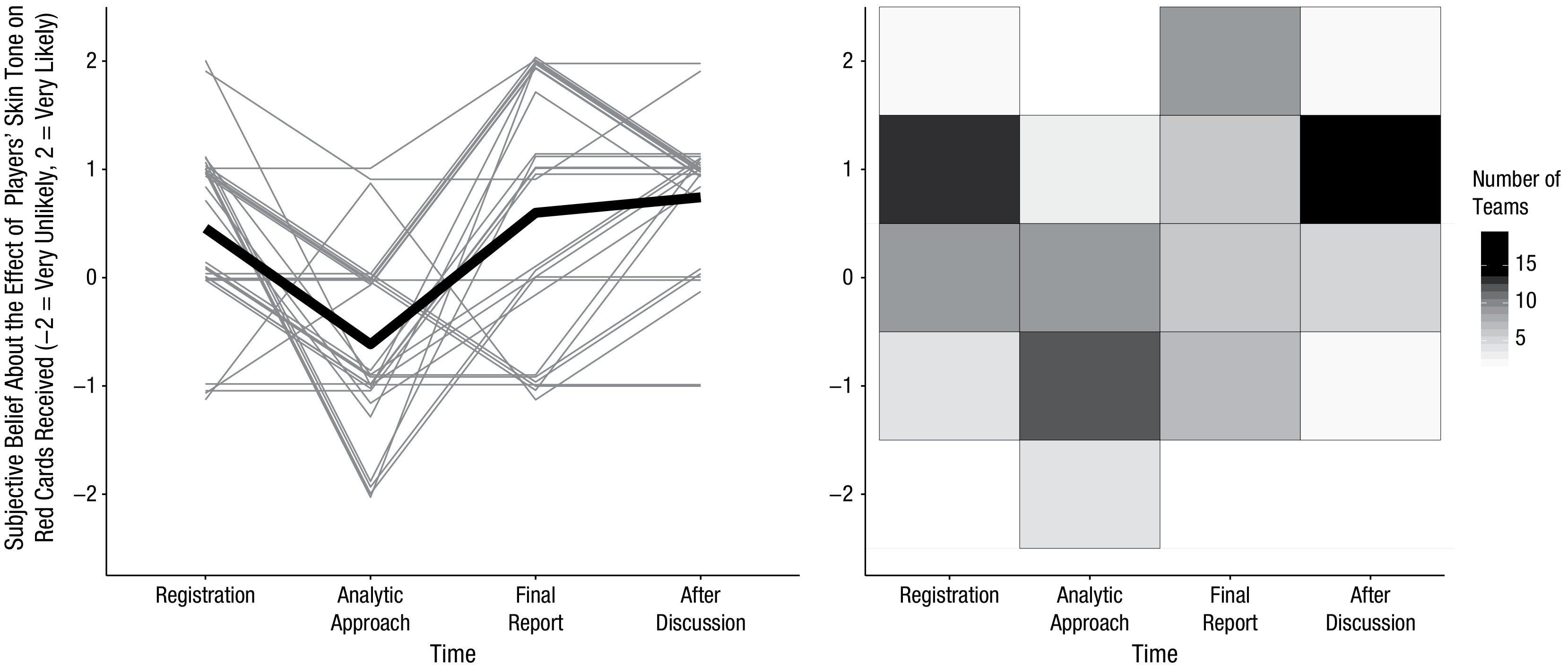

Fig. 4. The teams' subjective beliefs about the primary research question across time. For each of the four subjective-beliefs surveys, the plot on the left shows each team leader's response to the question asking whether players' skin tone predicts how many red cards they receive. The heavy black line represents the mean response at each time point. Each individual trajectory is jittered slightly to increase the interpretability of the plot. The plot on the right shows the number of team leaders who endorsed each response option at each time point.

Table 5. Analysts' Mean Agreement With Potential Conclusions That Could Be Drawn From the Data

Conclusion	Mean	*SD*
Positive relationship likely caused by referee bias	3.37	1.65
Positive relationship likely caused by unobserved variables (e.g., players' behavior)	4.21	1.37
Positive relationship but the cause is unknown	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 data set (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

Note: The results shown are from the final survey. Each item concerned whether there is a relationship between players' skin tone and the number of red-card decisions they receive. The response scale ranged from 1 (*strongly disagree*) to 7 (*strongly agree*). The items have been paraphrased for inclusion in the table.

95% CI = [.04, .68], respectively. Because both the magnitude of the estimated effect and the precision of the estimate varied by team, we also correlated the lower bound of the 95% CI and responses to this question and obtained the following correlations across the four time points: ρ = .29, 95% CI = [−.09, .60]; ρ = −.10, 95% CI = [−.46, .28]; ρ = .52, 95% CI = [.18, .75]; and ρ = .58, 95% CI = [.26, .78], respectively.

In short, the analysts' beliefs at registration regarding whether players with darker skin tone were more likely to receive red cards were not significantly related to the final effect sizes reported, but beliefs changed considerably throughout the research project, and as a result, the analysts' *post*-analysis beliefs were significantly related to both the reported effect-size estimates and the lower bounds of the 95% CIs for these estimates. These results suggest that there was some updating of beliefs based on the empirical results. Although the sample size was small (*N* = 29), the overall results are more consistent with rational updating of beliefs based on the evidence than with confirmation bias.

Does the analysts' expertise explain the variability in results?

An important question is whether the variability in the analytic choices made and results found by the teams resulted from teams with the greatest statistical expertise making different choices than the other teams. A related question is whether teams whose members had more quantitative expertise showed greater convergence in their estimated effect sizes. To answer these questions, we dichotomized the teams into two groups using latent class analysis. The first group (*n* = 9) was more likely to have a team member who had a Ph.D. (100% vs. 53%), was a 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 or statistical publication (78% vs. 47%). Seventy-eight percent of the teams in this first group reported effects that were statistically significant (median OR = 1.39, MAD = 0.13), whereas 68% of the teams with less expertise reported a significant effect (median OR = 1.30, MAD = 0.13). Analyses of the effects of the team's quantitative expertise on their choice of statistical models is provided in Supplement 6. Note, however, that teams in both latent classes exhibited considerable variability in whether they found a significant effect, and the two classes had similar degrees of dispersion in their effect-size estimates. Thus, overall, statistical expertise may have had some influence on analytic approaches and estimated effect sizes, but does not explain the high variability in these choices or in the results obtained.

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

We also examined whether the peer evaluations of the overall quality of each analytic approach were associated with the reported results. During the round-robin feedback phase, when the methods (but not results) for each team were known, the analysts rated their confidence in the suitability of other teams' analytic plans. The final effect sizes reported by teams whose analytic approach received higher confidence ratings (no rating lower than 4; median OR = 1.31, MAD = 0.15) did not differ from the reported effect sizes of those teams that received lower confidence ratings (median OR = 1.28, MAD = 0.12). Thus, there was little evidence that the variability in estimated effect sizes observed across teams was attributable to a subset of analyses that were lower than the others in quality overall.

Do the peer assessments of specific statistical issues explain the variability in results?

Toward the end of the crowdsourcing process, each team's final analytic approach was evaluated by other analysts who had particular expertise in that approach. These experts assessed the extent to which the assigned approaches addressed each of seven statistical issues and also rated their overall confidence in the approaches.

On average, each approach was assessed by 2.55 experts; 16 were reviewed by 3 experts, and 13 were reviewed by 2 experts. The average rating of agreement that statistical issues had not been addressed was 2.18 (*SD* = 0.55) on a scale from 1 to 5 (lower numbers indicate fewer unaddressed analytic issues).

The experts tended to be more convinced by approaches in which fewer problematic issues remained, as indicated by a negative correlation between the average rating across the seven statistical issues and the experts' rating of confidence (*r* = −.75, 95% CI = [−.60, −.86]). However, ratings for the analytic issues were unrelated to the OR for the relationship between darker skin tone and number of red cards received (*r* = .06, 95% CI = [−.35, .31]). Likewise, experts' overall confidence in each analytic approach was unrelated to the OR for the relationship between skin tone and red cards (*r* = −.03, 95% CI = [−.39, .60]). Overall, analyses revealed relatively little evidence that analytic approaches with identifiable statistical problems accounted for the variability in results across teams (e.g. by producing abnormally large or small effect sizes). Supplement 14 reports exploratory analyses aimed at determining whether certain kinds of analyses exhibited more convergence across teams than others did.

Implications for the Scientific Endeavor

It is easy to understand that effects can vary across independent tests of the same research hypothesis when different sources of data are used. Variation in measures and samples, as well as random error in assessment, naturally produce variation in results. Here, we have demonstrated that as a result of researchers' choices and assumptions during analysis, variation in estimated effect sizes can emerge even when analyses use the same data. The independent teams' estimated effects for the primary research question ranged from 0.89 to 2.93 in OR units (1.0 indicates a null effect); no teams found a negative effect, 9 found no significant relationship, and 20 found a positive effect. If a single team, selected randomly from the present teams, had conducted the study using the same data set, there would have been a 69% probability of a positive estimated effect size and a 31% probability of a null effect.

This variability in results cannot be readily accounted for by differences in expertise. Analysts with high and lower levels of quantitative expertise both exhibited high levels of variability in their estimated effect sizes. Further, analytic approaches that received highly favorable evaluations from peers showed the same variability in final effect sizes as did analytic approaches that were less favorably rated. This was true for two different measures of quality: peer ratings of overall quality and experts' ratings of whether specific statistical issues had been addressed.

The problem of analysis-contingent results is distinct from the problems introduced by p*-hacking, the garden of forking paths, and reanalyses of original data*

The main contribution of this article is in directly demonstrating the extent to which good-faith, yet subjective, analytic choices can have an impact on research results. This problem is related to, but distinct from, the problems associated with *p*-hacking (Simonsohn, Nelson, & Simmons, 2014), the garden of forking paths (Gelman & Loken, 2014), and reanalyses of original data used in published reports.

p*-hacking.* As originally defined by Simonsohn et al. (2014), *p*-hacking is either consciously or unconsciously exploiting researcher degrees of freedom in order to achieve statistical significance. For instance, they wrote that "researchers may file merely the subsets of analyses that produce nonsignificant results. We refer to such behavior as *p-hacking*" (p. 534). Thus, *p*-hacking is driven by the implicit or explicit goal to obtain statistically significant support for a particular conclusion. Although the specific decisions made in the process of *p*-hacking may be independently justifiable, it is not justifiable to choose an analytic strategy on the basis of whether it provides a desired result. Few editors would accept a manuscript, even one based on a series of prima facie defensible analytic choices, if the researchers admitted that they had made their analytic choices so as to reach the *p* < .05 criterion.

In the current crowdsourcing project, all the teams knew that their analyses would be shown to other analysts and made public, and the perceived need to achieve a significant result for publishability was lessened by the nature of the project. Although distinct from *p*-hacking, highly defensible analytic decisions made without direct incentives to achieve statistical significance can still produce wide variability in effect-size estimates. In the case of the hypothesized relationship between players' skin tone and referees' red-card decisions, the findings collectively suggest a positive correlation, but this can be glimpsed only through the fog of varying subjective analytic decisions.

The garden of forking paths. Gelman and Loken's (2014) concept of a garden of forking paths focuses not on selection from among different analytic options in order to achieve significant results (as in *p*-hacking), but rather on testing for significance after patterns in the data have been observed. Such data-contingent analyses do capitalize heavily (perhaps unintentionally) on chance, because patterns that emerge randomly are subjected to significance tests whose validity requires a priori predictions. This practice leads to "researcher degrees of freedom without fishing, [and] consists of computing a single test based on the data, but in an environment where a different test would have been performed given different data" (Gelman & Loken, 2014, p. 460).

The analysis-contingent results we examined in the current project reveal an issue that is broader than the issue of forking paths: Variability in effect sizes can occur even when the researcher has not looked for patterns in the data first and tested for significance only after the fact. For example, the analysts were asked to test a specific relationship between players' skin tone and referees' red-card decisions. This arguably limited opportunities for a garden-of-forking-paths process, which might have taken the form of examining relationships between players' various group-based characteristics (skin tone, ethnicity, per capita gross domestic product of country of origin), on the one hand, and various referee decisions (red cards, yellow cards, stoppage time, offside calls, disallowed goals), on the other, and then running formal significance tests only for the relationships that looked as if they might be meaningful.

Moreover, imagine if the 29 teams had been required to preregister their analysis plans before observing the data (Wagenmakers et al., 2012). Preregistration solves the problems of forking paths and *p*-hacking by removing the flexibility of data-contingent analyses and reducing the opportunity to present post hoc tests as a priori (Wagenmakers et al., 2012). However, preregistration would not have prevented the observed variability in effect-size estimates across the teams in this study. Outcomes can vary as a result of different, defensible analytic decisions whether they are made post hoc or a priori.

Reanalyzing data used in published reports. Making data from published reports more accessible to facilitate reanalyses and postpublication peer review (Hunter, 2012; Simonsohn, 2013; Wicherts et al., 2006) is important for science, but also does not make fully transparent the contingency of observed findings on analytic decisions. For example, few scientists would bother to write (and even fewer editors would publish) a commentary presenting new analyses and results unless they suggest a conclusion different from the one in the original publication. This creates perverse incentives for both original authors and commenters. Original authors have strong incentives to find positive results so that their work will be published, and commenters have strong incentives to find different (usually negative) results for the same reason. Thus, published commentaries will almost inevitably differ from original articles in their analytic approaches and conclusions, which introduces a strong selection bias.

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

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

How much variability in results is too much?

Scientists can have comparatively more faith in a finding when there is less variability in analytic approaches taken to investigating the targeted phenomenon and in results obtained using different methods. In a follow-up to this project, Crowdsourcing Data Analysis 2, a group of more than 40 analysts have independently analyzed a complex data set to test hypotheses regarding the effects of gender and status on intellectual debates. This new crowd of analysts are reporting radically dispersed effect sizes, and in some cases significant effects in opposite directions for the same hypothesis tested with the same data. In such extreme cases of little to no convergence in results, the crowdsourcing process suggests that the scientific community should have no faith that the hypothesis is true, even if one or two teams find significant support with a defensible analysis—results that might have been publishable on their own. In the present project on referees' decisions, the degree of convergence in results was relatively high by comparison, as more than two thirds of the teams found support for the hypothesis and the vast majority of teams obtained effect-size estimates in the predicted direction.

There will almost always be variability in a measured effect depending on analytic choices. As transparency about this variability increases with data-posting requirements and additional crowdsourced projects, scientists and policymakers will need to make ultimately subjective decisions about how much consistency is enough (and not enough) to conclude an effect most likely exists. Similar subjective and continually debated decisions have had to be made about the cutoff for statistical significance (Benjamin et al., 2017; Johnson, 2013). Setting cutoffs may be particularly challenging for policymakers because it is their responsibility to make decisions, and the ideal information on which to base a decision would include both whether an effect exists and how large it is. For example, some economic interventions might have both societally positive and societally negative effects, and policymakers will want to have precise estimates of all these effects to evaluate the trade-offs. Policymakers and practitioners may require greater convergence in effect-size estimates than scientists, for whom establishing a directional effect is often sufficient for building theory. We believe that crowdsourcing data-analysis initiatives will help policymakers by improving estimation of confidence and uncertainty. Crowdsourced analysis, combined with preregistered investigations and replications, will provide more informed benchmarks regarding the contingency of observed findings on characteristics of the sample and setting, procedures followed, and analytic decisions.

Generalizability to other data sets

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

Of course, more than one such investigation is needed to determine how contingent research results are on analytic decisions more generally. This demonstration is thus limited to being a case example; its conclusions are plausible, but have untested generalizability. For example, the project coordinators framed a specific research question for the analysts (Does players' skin tone correlate with referees' red-card decisions?), which may have artificially reduced the variability in estimated effect sizes. The research question could have been posed more broadly (Is there evidence of bias against minority groups in referees' decisions?), or the key outcome measure (e.g., number of yellow cards, number of red cards, stoppage time) could have been left up to each research team. This research question is being examined in Crowdsourcing Data Analysis 2, on the roles of gender and status in intellectual debates. In this follow-up project, analysts are also choosing how to operationalize each construct (e.g., is academic status best measured by citation counts, job rank, school rank, or some combination of these?). As noted earlier, the variability in effect-size estimates is even greater in this second project than in the present initiative. Systematic investigation via crowdsourcing will facilitate more general conclusions about how contingent research results are on analytic choices, and what characteristics of the research question, data set, and analyses serve as moderating variables.

There are also constraints on the useful application of crowdsourcing strategies. For example, the flexibility in analytic choices and thus their impact on estimated effect sizes is likely to be greatest when data sets are complex (e.g., longitudinal data sets with missing data, many potential covariates, levels of nesting). It remains an empirical question how contingent results are on analytic choices in the case of comparatively simple experimental studies with two to four conditions and few measured variables. There may still be enough choice points (outlier exclusions and statistical transformations, such as in the case of skewed data), even when researchers analyze a relatively simple data set, to introduce considerable variability in results based on those choices (Steegen, Tuerlinckx, Gelman, & Vanpaemel, 2016).

Notably, the robustness of experimental laboratory effects can also be assessed via replications repeating the same experimental design with new research participants (Ebersole et al., 2016; Klein et al., 2014). Crowdsourcing data analysis is particularly relevant for data sets that have many analytic choice points and that cannot easily be independently replicated with new observations. Crowdsourcing may also add a great deal of value when controversial research questions are being addressed or when there are many competing theoretical predictions to be adjudicated empirically.

Recommendations for individual researchers and teams

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

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

Conclusion

The observed results from analyzing a complex data set can be highly contingent on justifiable, but subjective, analytic decisions. Uncertainty in interpreting research results is therefore not just a function of statistical power or the use 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 analyzing data 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.

Action Editor

Daniel J. Simons served as action editor for this article.

Author Contributions

The first and second authors contributed equally to the project. E. L. Uhlmann proposed the idea of crowdsourcing data analysis and wrote the initial project outline. R. Silberzahn, E. L. Uhlmann, D. P. Martin, and B. A. Nosek developed the research protocol. R. Silberzahn and E. L. Uhlmann developed the specific research question regarding the influence of skin tone on referees' decisions. R. Silberzahn and D. P. Martin collected the data on referees' decisions and skin tone and prepared the data set for analysis. R. Silberzahn and D. P. Martin coordinated the different stages of the crowdsourcing process. All the other authors worked in teams to analyze the data, give feedback, and produce individual reports. A detailed list of contributions from each team is provided in Supplement 8 of the Supplemental Material. R. Silberzahn and D. P. Martin combined and analyzed the results of the different teams. E. L. Uhlmann outlined the manuscript and wrote the first draft of the abstract, introduction, and discussion of the implications of the findings. R. Silberzahn wrote the first draft of the description of the methods and the Supplemental Materials. R. Silberzahn and D. P. Martin wrote the first draft of the section reporting the main findings. D. P. Martin and R. Silberzahn created the figures and tables. B. A. Nosek heavily revised the manuscript, gave critical comments, and provided overall project supervision. All the authors reviewed the manuscript, and many provided crucial comments and edits that were incorporated.

Declaration of Conflicting Interests

The author(s) declared that there were no conflicts of interest with respect to the authorship or the publication of this article.

Funding

D. P. Martin was supported by the Institute of Education Sciences, U.S. Department of Education (Grant No. R305B090002). The contribution of R. D. Morey and E.-J. Wagenmakers was supported by a grant from the European Research Council (Grant No. 283876). M. Johannesson received funding from the Jan Wallander and Tom Hedelius Foundation (Grant No. P2015-0001:1), as well as from the Swedish Foundation for Humanities and Social Sciences (Grant No. NHS14-1719:1). S. Liverani was supported by a Leverhulme Trust Early Career Fellowship (Grant No. ECF-2011-576). C. R. Madan was supported by a Canadian Graduate Scholarship, Doctoral-level, from the Natural Sciences and Engineering Research Council of Canada (Grant No. CGSD2-426287-2012). T. Stafford was supported by a Leverhulme Trust Research Project Grant (Grant No. RPG2013-326).

Supplemental Material

Additional supporting information can be found at http://journals.sagepub.com/doi/suppl/10.1177/2515245917747646

Open Practices

All data and materials have been made publicly available via the Open Science Framework and can be accessed at https://osf.io/47tnc/ and https://osf.io/gvm2z/. The complete Open Practices Disclosure for this article can be found at http://journals

.sagepub.com/doi/suppl/10.1177/2515245917747646. This article has received badges for Open Data and Open Materials. More information about the Open Practices badges can be found at http://www.psychologicalscience.org/publications/badges.

Notes

1. This project also included a second research question: whether country-level preferences for light versus dark skin predict the red-card decisions of referees from the countries for which we had data on such preferences. In brief, the teams found little to no evidence that referees' decisions were moderated by explicit or implicit skin-tone preferences. However, data on individual referees' skin-tone preferences were not available; this variable was a measure of preferences based on aggregated data from referees' nations of origin, and the majority of the analysts judged the available data set to be inadequate to test this potential moderator. Detailed results are reported in Supplement 7 in the Supplemental Material available online.

2. One of the coauthors of this article, D. Molden, strongly disagreed with the project coordinators' decision to allow teams to choose to retain these covariates in any final analyses. He argued that the high rate of movement of players between clubs and leagues 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.

3. Because the majority of teams used analyses that favored reporting odds ratios, we chose this effect size as the common effect size. For teams that performed standard linear regression analyses, we used traditional conversion formulas (from Borenstein, Hedges, Higgins, & Rothstein, 2009) for both Cohen's *d* and standardized regression weights (assumed to be correlation coefficients). Additionally, because the prevalence of red cards is so low, we made the "rare disease" assumption by assuming that the risk ratios yielded in analyses adopting a Poisson regression framework were fair approximations to odds ratios (Viera, 2008).

References

Babtie, A. C., Kirk, P., & Stumpf, M. P. H. (2014). Topological sensitivity analysis for systems biology. *Proceedings of the National Academy of Sciences, USA*, *111*, 18507–18512. doi: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*, *7*, 543–554. doi:10.1177/1745691612459060

Benjamin, D., Berger, J., Johannesson, M., Nosek, B., Wagenmakers, E.-J., Berk, R., . . . Johnson, V. (2017). *Redefine statistical significance*. Retrieved from https://psyarxiv.com/mky9j

Bodenhausen, G. V. (1988). Stereotypic biases in social decision making and memory: Testing process models of stereotype use. *Journal of Personality and Social Psychology*, *55*, 726–737. doi: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 *Introduction to meta-analysis* (pp. 45–50). Chichester, England: John Wiley & Sons.

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

Carp, J. (2012b). The secret lives of experiments: Methods reporting in the fMRI literature. *NeuroImage*, *63*, 289–300. doi: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. doi:10.1037/0022-3514.83.6.1314

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

Ebersole, C. R., Atherton, O. E., Belanger, A. L., Skulborstad, H. M., Allen, J. M., Banks, J. B., . . . Nosek, B. A. (2016). Many Labs 3: Evaluating participant pool quality across the academic semester via replication. *Journal of Experimental Social Psychology*, *67*, 68–82.

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. *Journal of the American Medical Association*, *312*, 1024–1032. doi:10.1001/jama.2014.9646

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

Gelman, A., & Loken, E. (2014). The statistical crisis in science. *American Scientist*, *102*, 460–465.

Hugenberg, K., & Bodenhausen, G. V. (2003). Facing prejudice: Implicit prejudice and the perception of facial threat. *Psychological Science*, *14*, 640–643. doi:10.1046/j.0956-7976.2003.psci_1478.x

Hunter, J. (2012). Post-publication peer review: Opening up scientific conversation. *Frontiers in Computational Neuroscience*, *6*, Article 63. doi:10.3389/fncom.2012.00063

Johnson, V. E. (2013). Revised standards for statistical evidence. *Proceedings of the National Academy of Sciences, USA*, *110*, 19313–19317.

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

Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B., Jr., Bahník, Š., Bernstein, M. J., . . . Nosek, B. A. (2014). Investigating variation in replicability: A "Many Labs" replication project. *Social Psychology*, *45*, 142–152.

Krumholz, H. M., & Peterson, E. D. (2014). Open access to clinical trials data. *Journal of the American Medical Association*, *312*, 1002–1003. doi: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. doi: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. doi: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. doi: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. *American Economic Review*, *101*, 1410–1435.

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*, *9*, 652–660. doi:10.1177/1745691614549257

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

Simonsohn, U. (2013). Just post it: The lesson from two cases of fabricated data detected by statistics alone. *Psychological Science*, *24*, 1875–1888.

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). *P*-curve: A key to the file-drawer. *Journal of Experimental Psychology: General*, *143*, 534–547.

Simonsohn, U., Simmons, J., & Nelson, L. (2016). *Specification curve: Descriptive and inferential statistics for all plausible specifications*. Unpublished manuscript, Operations, Information and Decisions Department, University of Pennsylvania.

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. doi:10.1177/0956797611417632

Steegen, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science*, *11*, 702–712.

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

Viera, A. J. (2008). Odds ratios and risk ratios: What's the difference and why does it matter? *Southern Medical Journal*, *101*, 730–734. doi: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. doi: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. doi:10.1037/0003-066X.61.7.726