Supplement 5: Analysts' subjective beliefs regarding the primary research hypothesis

Tracking subjective beliefs across time. Analysts' subjective beliefs about the theoretical hypothesis were assessed four times during the project. This measure was centered in all subsequent analyses to increase interpretability. Subjective beliefs exhibited variability across time (see Fig. S5). When we asked researchers at their initial registration (i.e., before they had received the data), there was slight agreement on average that a positive relationship existed between number of red cards and player skin-tone, yet opinions varied greatly (M = 0.61, Stdev = 1.20). We asked the same question again after researchers accessed the data and submitted their analytical approach. At that point, the slight initial agreement had turned into slight disagreement regarding whether a relationship existed (M = -0.61, Stdev = 0.88). At the point of the submission of their final analyses, overall slight agreement existed again of the hypothesized relationship at a magnitude similar to the initial beliefs, yet again with substantial variability (M = 0.61, Stdev = 1.20). Finally, after a group discussion with all approaches and results available for collective review, overall agreement increased slightly and, notably, variability in beliefs decreased (M = 0.75, Stdev = 0.70), suggesting convergence in beliefs over time.

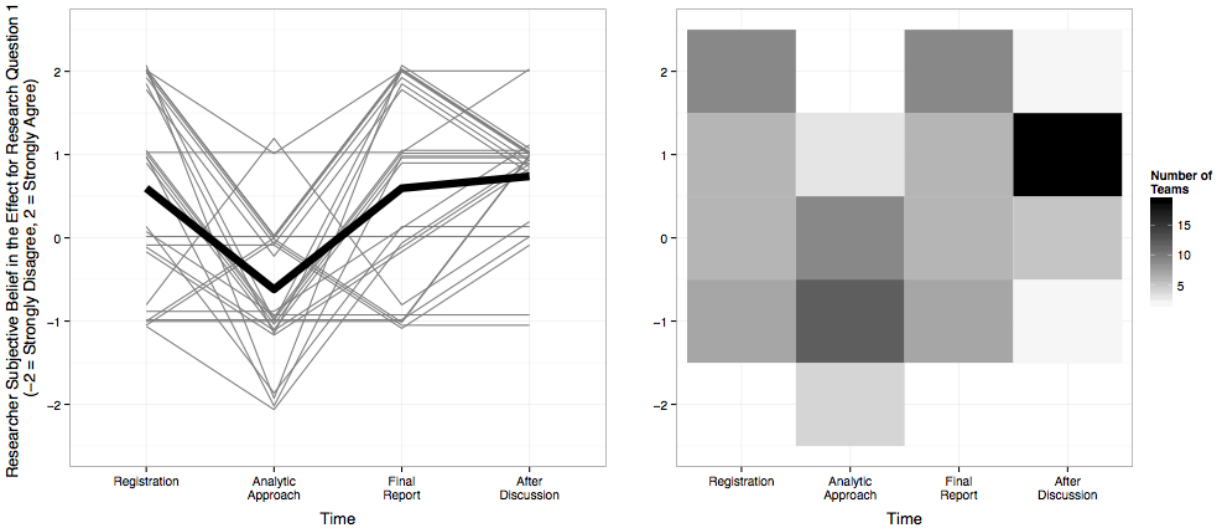

Fig. S5. The plot on the left reflects team leader beliefs regarding the primary research question: whether player skin tone predicts referee red cards. Each light gray line represents a single team's trajectory throughout the project, and the black trajectory represents the mean value at each time point. Note that each individual trajectory is jittered slightly to increase the interpretability of the plot. The plot on the right represents the consensus (or lack thereof) by plotting the number of team leaders endorsing a particular response category at each time point.

Nuanced beliefs about the research question. In the fourth and final survey we administered items assessing more nuanced beliefs about our primary research question (i.e., whether there is an association between player skin tone and referee red card decisions). These included items such as "The effect is positive and due to referee bias" and "There is little evidence for an effect." Analysts responded to these items on scales ranging from 1 (*strongly disagree*) to 7 (*strongly agree*). The items, means, and standard deviations are reported below. Note that the complete first item was, "This dataset suggests a positive relationship between darker skin-toned players and frequency of receiving red cards that is likely caused by referee bias." Items were paraphrased for inclusion in the table.

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

By the end of the project, a majority of teams agreed that the data showed a positive relationship between number of red cards and player skin-tone. As seen above, the greatest endorsement (78% agreement) was given to the statement "The effect is positive and the mechanism is unknown" (M = 5.32, SD = 1.47).

Correlations between subjective beliefs and effect size estimates. Of further interest was whether subjective beliefs that the primary research hypothesis is true were related to the results a team obtained. Self-reported beliefs regarding research question 1 at each stage were correlated with the final reported effect size using Spearman's rho, with the following magnitudes across the four timepoints (and corresponding 95% CIs): 0.14 [-0.25, 0.49], -0.20 [-0.53, 0.19], 0.43 [0.07, 0.69], 0.41 [0.04, 0.68]. Because both the magnitude of the effect and the estimate precision varied by team, Spearman's rho correlations were also calculated between the lower bound of the final reported effect size and self-reported beliefs regarding the primary research question, with the following magnitudes across the four timepoints (and corresponding 95% CIs): 0.29 [-0.09, 0.60], -0.10 [-0.46, 0.28], 0.52 [0.18, 0.75], 0.58 [0.26, 0.78].

Analysts' beliefs at registration regarding whether dark skin toned players were more likely to receive red cards were weakly related to the observed effect size of their final report (rho = 0.14 [-0.25, 0.49]). However, beliefs changed considerably throughout the research process, and analysts' *post*-analysis belief in the hypothesis was related to their effect estimate and lower bound (*rho* = 0.41 and *rho* = 0.58, respectively), also suggesting updating of beliefs

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