

**Crowdstorming Research: Many analysts, one dataset**  
**Research Protocol**  
**Spring 2014**

**Research Question: Are soccer referees more likely to give red cards to dark skin toned players than light skin toned players?**

**Overview**

In a standard scientific analysis, one analyst or team presents a single analysis of a data set. However, there are often a variety of defensible analytic strategies that could be used on the same data. Variation in those strategies could produce very different results.

We introduce the approach of "crowdstorming a dataset." Multiple independent analysts are recruited to investigate the same hypothesis or hypotheses on the same data set in whatever manner they see as best. The independent analysis strategies produce two datasets of interest: (1) the variation in analysis strategies, and (2) the variation in estimated effects. These two can be partially independent. Different analysis strategies may converge to a very similar estimated effect - indicating robustness despite variation in analysis strategies. Alternatively, the estimated effect may be highly contingent on analysis strategy. In the latter case, there are at least two methods of resolution: (1) consider the central tendency of the estimated effects to be the most accurate, or (2) critically evaluate the analysis strategies to determine whether one or more should be elevated as the preferred analysis.

This approach should be especially useful for complex data sets in which a variety of analytic approaches could be used, and when dealing with controversial issues about which researchers and others have very different priors. If everyone comes up with the same results, then scientists can speak with one voice. If not, the subjectivity and conditionality on analysis strategy is made transparent. Further, when crowdstorming a data set, the potential for errors and suboptimal analyses are reduced.

This first project establishes a protocol for independent simultaneous analysis of a single dataset by multiple teams, and resolution of the variation in analytic strategies and effect estimates among them. Next, we summarize the research question, process for collaboration, and the available dataset. The Open Science Framework project page is [here](#).

## Research Questions

For [this first project](#), we crowdsource the questions of whether soccer referees are more likely to give red cards to dark skin toned players than light skin toned players, and whether this effect is moderated by skin-tone prejudice across cultures. The available dataset provides an opportunity to identify the magnitude of the relationship among these variables. It does not offer opportunity to identify causal relations.

Research Question 1: Are soccer referees more likely to give red cards to dark skin toned players than light skin toned players?

Research Question 2: Are soccer referees from countries high in skin-tone prejudice more likely to award red cards to dark skin toned players?

## Relevant background

For Question 1: Research on assimilation to stereotypes in social perception (Bodenhausen, 1988; Correll et al., 2002; Hugenberg & Bodenhausen, 2003) and cultural preferences for light skin (Maddox & Gray, 2002; Sidanius et al., 2001; Twine, 1998) predicts that darker skin tone will be associated with receiving more red cards. On the other hand, research on accountability (Lerner & Tetlock, 1999), and the debiasing effects of real world professional experience (List, 2003; Levitt & List, 2008) gives reasons to expect no such effect. Although concluding the null is always difficult, our large sample size gives us much greater leeway than usual with regard to concluding no evidence of bias.

For Question 2: Research and theory on the roots of perceptual biases in cultural socialization (Banaji, 2001; Greenwald & Banaji, 1995) suggests growing up in a society that favors light over dark skin should ingrain such prejudices in individual members of that culture. On the other hand, implicit and explicit prejudices measured at the aggregate level of societies may not related to individual-level judgments as these are different levels of analysis and relatively “distant” predictors.

## Related Research

There is some relevant literature looking at other sports, specifically basketball and baseball. Price and Wolfers (2010) demonstrated a same-race bias in NBA foul calls (e.g., White referees call more fouls on Black players) and rebutted the NBA’s criticisms in a follow up paper (Price & Wolfers, 2011). Parsons et al. (2011) and Kim and King (in press) demonstrate racial bias in calls by baseball umpires. Pope, Price, and Wolfers (2013) show that after the publicity around

the original Price and Wolfers paper, the same-race bias shown in NBA referee calls was eliminated. This provides a strong ethical impetus for carrying out the present project. The publicity and controversy surrounding the original Price and Wolfers paper also makes it even more important than usual to get things right when looking for evidence of similar biases among soccer referees.

## **Project Coordination and Authorship**

Raphael Silberzahn and Dan Martin are the project coordinators. Eric Uhlmann is the lead writer and Brian Nosek will supervise the project. The two project coordinators and lead writer will be the first three authors followed by alphabetical listing of all other authors, and then Brian Nosek.

Authorship is earned by completing and submitting a reproducible analysis within the stated timeframe. This includes: (1) the code for the analysis and specification of analysis package required to execute the analysis, (2) a description of the rationale for the analysis strategy, (3) a complete written description of the analysis strategy, and (4) a description of the result including specification of the effect estimate in effect size units ( $d$ ,  $r$ ,  $R^2$  or odds ratio) and 95% confidence interval around the estimate.

## **Planned Timeline**

There are seven phases for this crowdstorming project. In order to meet the timeline, some later phases may commence while earlier phases are in process. For example, some of the report will be written while final data analyses are still in process.

1. **Registration:** Registration via [Google Forms document](#) and with the [Open Science Framework](#): project page is [here](#) (Complete by May 18th, 2014).
2. **1<sup>st</sup> Round Analyses:** First round of Analyses conducted until June 15, EST and analytical approaches are uploaded and shared with other research teams. Initial findings are shared with the project coordinators but not with other research teams.
3. **Round Robin Feedback Round:** Research teams comment and provide suggestions on other teams' research approaches (until June 29, 2014).
4. **2<sup>nd</sup> Round Analyses:** Research teams refine their analytical approach and upload their final analyses (until 20th of July, 2014).
5. **Working Paper:** A working paper presenting and discussing the different results will be circulated to research teams (before August 3rd, 2014) and made available for the wider public (until August 17th, 2014).

## Elaboration of Project Stages

### 1. Registration

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

### 2. 1<sup>st</sup> Round Analyses

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

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

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
























### The Dataset

From a company for sports statistics, we obtained data and profile photos from all soccer players ( $N = 2,053$ ) playing in the first male divisions of England, Germany, France and Spain in the 2012-2013 season and all referees ( $N = 3,147$ ) that these players played under in their professional career (see Figure 1). We created a dataset of player–referee dyads including the number of matches players and referees encountered each other and our dependent variable, the number of red cards given to a player by a particular referee throughout all matches the two encountered each other.

Player's photo was available from the source for 1,586 out of 2,053 players. *Players' skin tone* was coded by two independent raters blind to the research question who, based on their profile photo, categorized players on a 5-point scale ranging from “very light skin” to “very dark skin” with “neither dark nor light skin” as the center value.

Figure 1:

Player overview with list of referees and player-referee statistics, such as matches, goals, and cards.

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

Mauro Boselli



Additionally, implicit bias scores for each referee country were calculated using a race implicit association test (IAT), with higher values corresponding to faster white | good, black | bad associations. Explicit bias scores for each referee country were calculated using a racial thermometer task, with higher values corresponding to greater feelings of warmth toward whites versus blacks. Both these measures were created by aggregating data from many online users in referee countries taking these tests on [Project Implicit](#).

## Data Structure

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

Variable Name:	Variable Description:
playerShort	short player ID
player	player name
club	player club
leagueCountry	country of player club (England, Germany, France, and Spain)

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

- 3. Round Robin Feedback Round:** After submitting their analytical approach, teams are invited to view others' approaches, take inspiration from them and comment and reflect the different strategies. Further details of this process are to be announced.
- 4. 2<sup>nd</sup> Round Analyses:** Based on their initial analyses, and the input received during the Round Robin Feedback round research teams refine their analytical approach and work out their final analyses and conclusion they draw from the data.

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

If everyone reached similar conclusions, scientist can speak with one voice on a socially important issue, which is a nice contribution. If different analysts reach very different results with multiple, defensible approaches, this is also a contribution in highlighting that there is a great deal of subjectivity in science. If errors or suboptimal analyses were uncovered when similar analyses by different analysts were compared, that's a contribution too as scientific errors were avoided through the use of many independent analysts.

There are also some potential drawbacks of crowdstorming that may be worth discussing. The results section will likely become very long because of the need to present the results of so many different analysts. It is also perhaps inefficient to always have many different analysts analyze the same data set to test the same hypothesis. There is limited professional reward for many of those involved, most of whose names are lost in a long author string. In some cases crowdstorming could lead to a “Tower of Babel” problem, where one analytic approach is actually optimal but it is lost amid less optimal (if still defensible) approaches.

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

## References

- Banaji, M. R. (2001). Implicit attitudes can be measured. In H. L. Roediger, III, J. S. Nairne, I. Neath, & A. Surprenant (Eds.), *The nature of remembering: Essays in honor of Robert G. Crowder* (pp. 117-150). Washington, DC: American Psychological Association.
- Bodenhausen, G. V. (1988). Stereotypic biases in social decision making and memory: Testing process models of stereotype use. *Journal of Personality and Social Psychology*, 55, 726-737.
- Correll, J., Park, B., Judd, C.M., & Wittenbrink, B. (2002). The police officer's dilemma: Using ethnicity to disambiguate potentially threatening individuals. *Journal of Personality & Social Psychology*, 83, 1314–1329.
- Greenwald, A. G., & Banaji, M. R. (1995). Implicit social cognition: Attitudes, self-esteem, and stereotypes. *Psychological Review*, 102, 4-27.
- Hugenberg, K., & Bodenhausen, G. V. (2003). Facing prejudice: Implicit prejudice and the perception of facial threat. *Psychological Science*, 14, 640-643.
- Kim, J., & King, B.G. (in press). Seeing stars: Matthew effects and status bias in Major League Baseball umpiring. *Management Science*.
- News: <http://mobile.nytimes.com/2014/03/30/opinion/sunday/what-umpires-get-wrong.html>
- Lerner, J.S., & Tetlock, P.E. (1999). [Accounting for the effects of accountability](#). *Psychological Bulletin*, 125(2), 255-275.
- Levitt, S.D., & List, J.A. (2008). Homo economicus evolves. *Science*, 319, 909–910.
- List, J.A. (2003). [Does market experience eliminate market anomalies?](#) *Quarterly Journal of Economics*, 118(1), 41–71.
- Maddox, K.B. & Gray, S. (2002). Cognitive representations of African Americans: Re-exploring the role of skin tone. *Personality and Social Psychological Bulletin*, 28, 250-259.
- Parsons, C., Sulaeman, J., Yates, M., & Hamermesh, D. (2011). [Strike Three: Discrimination, Incentives, and Evaluation](#). *American Economic Review*, 101, 1410–1435.



Pope, D., Price, J., & Wolfers, J. (2013). [\*Awareness Reduces Racial Bias\*. NBER Working Paper No. 19765.](#)

Price, J., & Wolfers, J. (2010). [\*Racial discrimination among NBA referees\*. \*Quarterly Journal of Economics\*.](#)

Price, J., & Wolfers, J. (2011). [\*Biased Referees?: Reconciling Results with the NBA's Analysis\*. \*Contemporary Economic Policy\*.](#)

Sidanius, J., Peña, Y. & Sawyer, M. (2001). Inclusionary discrimination: Pigmentocracy and patriotism in the Dominican Republic. *Political Psychology*, 22, 827-851.

Twine, F. W. (1998). *Racism in a racial democracy*. New Brunswick, NJ: Rutgers University Press.