Professional soccer referees give more red cards to darker-skinned players but the evidence for prejudice remains uncertain

Authors: Daniel C. Molden[1]*, Maureen A. Craig[1], Ryan Lei[1], and Monica Gamez-Djokic[1]

Affiliations
[1]Northwestern University, Department of Psychology
*Correspondence to: molden@northwestern.edu

Abstract

We examined whether professional soccer referees assign more red cards to players with darker skin-tone. Preliminary analyses assessed confounding or suppressor variables that might obscure such an association and identified player position, and player win-, draw-, and loss-percentages as potential confounds. Multilevel regressions in which the number of red cards players received was predicted by skin tone, these confounds, and the number of games in each player-referee dyad were then conducted, with referee and referee country-of-origin as random intercepts. Results showed that darker-skinned players received more red cards, but that this association was not moderated by mean levels of implicit or explicit prejudice in referees' countries-of-origin. Treating red cards as a binomial event and using multilevel logistic regression supported the same conclusions. Supplementary analyses showed that darker-skinned players received *fewer* yellow cards, suggesting that skin tone may predict whether severe fouls result in ejection rather than merely in a caution.

One Sentence Summary

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

Results

Data Cleaning and Preparation

Players with missing skin-tone ratings were excluded from analyses involving this variable (*N* = 467), but were retained for all other analyses. For the remaining players the two separate coders' skin-tone ratings were averaged into a single index (*M* = 0.29, *SD* = 0.29). Players with missing data on their position (*N* =152) were assigned a value of "unknown" so that they could be retained in analyses involving this variable.

To explore as many potential confounds as possible, several new variables were created. First, a variable was created to capture how *physically imposing* a player might seem. More physically imposing players could be more likely to be perceived as committing a violent act and thus earn more red cards. This variable was calculated by converting player height and weight into a Body-Mass Index (BMI) using the standard formula (weight in kg/[height in m]^2). Neither height nor weight alone adequately capture the extent to which a player might be seen as physically imposing (some players may be tall but slight, whereas others may be heavy but short). Because all players were elite professional athletes, and higher BMIs were more likely to reflect having a larger and more muscular frame rather than being more overweight as in the general population, this index was selected as the best available proxy.

Second, a variable was created to capture *star status*. More popular or famous players might potentially be given more lenient treatment by referees and be less likely to be ejected with a red card. This variable was determined by calculating the average number of goals scored by each player per game they played in the data set. Although this too is imperfect in that it does not properly account for players who are defensive stars and score fewer goals, offensive players are typically the most famous and thus goals scored should generally be a meaningful indicator of this status.

Finally, a variable was created to capture *game situation*. Teams that are losing a match may be more aggressive, and referees may be more attentive for retaliatory actions, than when teams are winning or tied. Although match-level data was not available and it is impossible to determine whether a player was winning or losing in the actual game in which he received or did not receive a red card, it is possible to examine how the frequency with which players were on the winning or losing side within each player-referee dyad relates to red cards received. Therefore, to assess this frequency, separate win, loss, and draw percentages were calculated for each player across the games included in the data set.

. Some data was available on the particular professional leagues and clubs of which players were a member during the 2012-2013 season. However, because in the data file each player was only identified with one club and red card data was drawn over multiple years in which the player may have been with a different club or in a different league, no analyses were possible with these variables. Similarly, although a variable for player age was included in the data, because red cards were examined across a number of different years for each player and there is no way to determine how old each player was during each particular match or when paired with each particular referee, age was not included in any of our analyses.

Basic Analysis Strategy

We focused on the question of whether variation in player skin-tone could potentially explain variation in the number of red cards received, as opposed how well player skin-tone predicts variance in the number of red cards in relation to other relevant predictors. Thus, the first step in our analysis was to identify variables that could serve as potential confounds or surpressor variables and obscure any explanatory relationship between skin tone and red cards. This was done by examining what additional factors were both (a) correlated with player skin-tone, and (b) themselves correlated with the number of red cards received. Any variables that fulfilled both of these criteria were included as covariates in all of the primary analyses. Any additional variables that were correlated with the number of red cards received but were not

correlated with skin tone were not relevant for our explanatory analysis and were therefore omitted from the final model. It should again be noted that this approach means that the present analysis is not designed to answer the question of how good a predictor of red cards skin tone is compared to other factors that might influence this outcome, but it does answer the question of whether there is any evidence that skin tone has some association with red cards received.

In constructing our explanatory regression model, we used multi-level modeling in order to properly account for the multiply-nested structure of the data (Raudenbush & Bryk, 2002). In the original format of the data provided, the red card data for individual players were nested within different player-referee dyads, and the referees were nested within a variety of countries-of-origin. The most appropriate nesting structure for the regression model was evaluated empirically by testing a series of adjusted-means models. In these analyses, the number of red cards received was always regressed only on the number of games in each player-referee dyad, and the variance in these adjusted means accounted for by the different possible grouping variables (i.e., referee and referee country) was estimated. The necessary level of nesting for the final model was determined by retaining any grouping variable that explained significant variance in red cards above and beyond the next simplest level. Additional tests then predicted red cards received from games in the player-referee dyad and coded skin-tone to further evaluate whether modeling the slope of the skin tone as a random effect explained significantly more variance than modeling it as a fixed effect. All tests of significance between different models were evaluated using χ^2 tests of the difference between the -2 log likelihoods of each model with the degrees of freedom equal to the difference between the degrees of freedom in each model (Raudenbush & Bryk, 2002).

Results showed that the individual referees with which players were paired, $\chi^2(1) = 70.10$, $p < .001$, and referee country-of-origin, $\chi^2(1) = 100.70$, $p < .001$, each explained significant variance in the red cards received above and beyond a model in which red cards

received were fixed across these grouping variables. Furthermore, the model in which referee was nested within referee country explained significant variance beyond either of these factors alone, χ^2s(1) > 29.10, *p*s < .001. Therefore, this three-level structure of player nested within referee, nested within referee country was used in all analyses reported below. Further analyses comparing the simple skin-tone slope on red cards in this three-level model as either a fixed or a random effect showed that the random model did not explain any additional variance beyond the fixed model, $\chi^2(4) = 0.0001$, $p = 1.0$. Therefore, this slope was evaluated as a fixed effect in all analyses.

Assessing Potential Confounds

To examine whether any of the available or newly created variables in the data set were indeed potential confounds, a preliminary set of correlations was calculated between players' mean skin tone ratings, their physical stature, their star status, and how often they experienced different game situations. For this specific set of analyses, a new data matrix was constructed that aggregated all of the relevant data at the player level. In addition, a one-way analysis of variance (ANOVA) was conducted with the player-level data set to examine whether, on average, players with different skin tones were equally distributed across different positions. For any variables that significantly related to skin tone, follow-up analyses were conducted to assess whether these variables also predicted the number of red cards received.

Results of the player-level correlations showed that player skin tone was not correlated with physical stature, $r = .03$, $p = .32$, or star status, $r = .03$, $p = .27$. It was however correlated with game situation. Darker-skinned players had slightly, but significantly, lower win percentages, $r = -.09$, $p = .001$, higher draw percentages, $r = .10$, $p < .001$, and marginally higher loss percentages, $r = .05$, $p = .06$. These results further showed that player skin tone was not equally distributed across different positions, $F(12, 1572) = 5.42$, $p < .001$.

In addition, a regression on the full data set in which the number of red cards received was predicted simultaneously by win percentage and draw percentage, controlling for the

number of games in which the player was matched with each particular referee, showed that a greater proportion of games won, β = -.06, t(142878) = 8.31, p < .001, or drawn, β = -.04, t(142878) = 4.56, p < .001, by a player paired with a particular referee were associated with fewer red cards, and a multilevel ANCOVA in which the number of red cards received was predicted by the player's position controlling again for the number of games in each player-referee dyad showed that position was differentially associated with the number of red cards received, F(12, 142868) = 11.92, p < .001. Therefore, win percentage, draw percentage, and position were used as covariates in all of the primary analyses (and because win percentage, draw percentage, and loss percentages sum to 1.0 and are linearly dependent, loss percentage was accounted for as well).

Primary Analyses

In our approach, we maintained the structure with which the data was gathered, involving the outcomes of aggregated player-referee dyads. To control for the unequal number of times players were paired with different referees, we included the number of games in which this dyad occurred as another covariate in all analyses. This approach was preferred over creating a proportion, because such a transformation would falsely equate instances in which a player received zero red cards over two games with a particular referee and instances in which a player received zero red cards over 20 games with that referee. Furthermore, even though red cards were a rare event among players who possessed skin tone ratings (occurring in 1.3% of dyads), and multiple red cards within a dyad were even rarer (occurring in .02% of dyads), we initially chose not to analyze this outcome as a binomial event because we believed that this again would not adequately account for the varying rates at which each player received a red card from each referee based on how often the two were paired (but see the supplementary analysis for this alternate approach).

As outlined above, Hypothesis 1 concerning whether skin tone was related to receiving a red card was therefore tested with a multilevel model in which the number of red cards occurring

within each player-referee dyad was predicted by the fixed effects of mean player skin-tone controlling for number of games within the dyad, the win and draw percentage of the player within the dyad when paired with that particularly referee, and the player's position (which was dummy-coded alphabetically). Player-referee dyad was nested within referee and then further within referee country-of-origin, and intercepts were allowed to vary randomly across these levels. Results showed that player skin-tone explained a small, but significant, amount of variance in the number of red cards received, β = .008 [.003, .014], t(121627) = 2.84, p = .005, d = 0.02. Darker-skinned players received slightly more red cards than lighter-skinned players.

Hypothesis 2a concerning whether the association between skin tone and receiving a red card was moderated by the estimated country-level implicit prejudice against people of African descent of the referee's country-of-origin (as assessed by the IAT) was tested by adding this variable to the above multilevel regression, along with an IAT x mean skin-tone interaction term. Results showed no evidence that the country-level IAT explained any variation in the association between player skin-tone and the number of red cards received, β = .0003 [-.005, .005], t(121484) = 0.12, p = .90, d = 0.00. Hypothesis 2b concerning whether the association between skin tone and receiving a red card was moderated by the estimated country-level explicit prejudice against people of African descent of the referee's country-of-origin was tested by adding this variable to the above multilevel regression, along with an explicit prejudice x mean skin-tone interaction term. Results showed no evidence that the country-level explicit prejudice explained any variation in the association between player skin-tone and the number of red cards received, β = .0004 [-.005, .005], t(121484) = 0.17, p = .87, d = 0.00.

Supplementary Analyses

Following these tests of the primary hypotheses, several supplementary analyses were performed. First, all of the above analyses were repeated with yellow cards, and a newly created index of *total bookings* (i.e., total number of penalty cards received = red cards + yellow cards + 2 x yellow-red cards) as the primary dependent variables. Results showed that the

same nesting structure as used for the red card analyses was best supported by the data, and the same variables of position, win percentage, and draw percentage were potential confounds that needed to be included in testing all skin tone effects. The number of games in which a player-referee dyad occurred was also always included as a covariate. The only other significant results found were that darker-skinned players actually received slightly *fewer* yellow cards than lighter-skinned players, β = -.005 [-.0002, -.010], $t(121627) = 2.06$, $p = .04$, $d = 0.01$. This association was also not moderated by the referee's country-of-origin-level implicit or explicit prejudice.

Another set of analyses attempted to examine skin tone as a more direct contrast between players who appeared to be of African versus non-African descent. Because implicit and explicit prejudice were specifically operationalized in these terms, whereas the continuous coding of skin tone from light to dark was not, this approach could potentially provide a more sensitive test of Hypotheses 2a and 2b. A visual inspection of the coders ratings showed that codes of .75 and 1.0 almost exclusively were given to African-appearing players, and codes of 0 and .25 were almost exclusively given to non-African appearing players. However, ratings of .50 were given to some players of African descent and some players of non-African descent (e.g., Asian players). Therefore players with mean skin-tone ratings below .50 were categorized as non-African appearing and players with mean skin-tone ratings above .50 were categorized as African-appearing. Players who scored at .50 ($N = 116$) were dropped from this supplementary analysis. All primary analyses were then repeated exactly as described above substituting in the categorical variable of appearance (-1 = non-African appearing, 1 = African appearing) in place of mean skin-tone ratings.

Results showed that, African-appearance did not significantly explain any variance in the number of red cards received, β = .014 [-.002, .030], $t(112697) = 1.73$, $p = .08$, $d = 0.01$, but it did significantly explain variance in the number of yellow cards received, β = -.029 [-.042, -.015], $t(112697) = 4.22$, $p < .001$, $d = 0.03$, and the total number of bookings, β = -.022 [-.035, -.009],

$t(112697) = 3.24$, $p < .002$, $d = 0.02$. African-appearing players received fewer yellow cards and fewer total bookings. In this set of analyses, country-of-origin-level implicit prejudice, $\beta = -.004$ [-.017, .009], $t(112569) = 0.61$, $p = .54$, $d = 0.00$, and explicit prejudice, $\beta = -.003$ [-.017, .010], $t(112569) = 0.47$, $p = .64$, $d = 0.00$, of the referees did not moderate the number of red cards received. However, implicit prejudice did moderate the association between appearing African and the number of yellow cards received, $\beta = -.012$ [-.023, -.0006], $t(112569) = 2.06$, $p < .04$, $d = 0.01$, as well as the total bookings, $\beta = -.012$ [-.023, -.0006], $t(112569) = 2.05$, $p = .04$, $d = 0.01$. In both cases, follow-up analyses showed that the higher levels of implicit prejudice in the referees' country-of-origin, the fewer yellow cards and total bookings African-appearing players received. Explicit prejudice did not moderate the number of yellow cards, $\beta = -.008$ [-.019, .004], $t(112569) = 1.26$, $p = .21$, $d = 0.01$, or the total bookings, $\beta = -.008$ [-.019, .004], $t(112569) = 1.36$, $p = .17$, $d = 0.01$ received.

Finally, although we chose not to transform the number of red cards received to a binomial variable for the reasons noted above, the distribution of red cards is extremely non-normal and could bias the analyses. To check the robustness of these analyses, we therefore repeated the tests of the primary hypotheses after converting red cards to a binomial variable within each player-referee dyad and conducting multilevel logistic regressions with the same set of covariates. Results of this analysis supported the same conclusions: skin tone explained significant variance in the likelihood of receiving a red card $\beta = .275$ [.09, .46], $z = 2.99$, $p = .003$, $OR = 1.32$ [1.09, 1.58], but this association was not moderated by referee's country-of-origin-level implicit prejudice, $\beta = .484$ [-6.15, .7.11], $z = 0.14$, $p = .89$, $OR = 1.62$ [0.001, 1224], or referee's country-of-origin-level explicit prejudice, $B = .39$ [-0.56, 1.35], $z = 0.81$, $p = .42$, $OR = 1.48$ [0.57, 3.86].

Conclusions

On the whole, results showed an extremely small, but statistically robust association of player skin tone with the number of red cards given by professional soccer referees

independent of possible confounds involving players' physical stature, the position they play, their goal-scoring prowess, or their likelihood of being on the winning or losing team. Darker-skinned players received more red cards. However, there was no evidence that referees from countries with higher national levels of implicit or explicit prejudice were any more likely to assign red cards to darker-skinned player. Thus, there was no direct link between the skin-tone effect and prejudice in this data.

Several important qualifications to these results must be noted. First, country-level prejudice scores are an extremely crude proxy for the prejudice levels of the professional soccer referees in this sample. Thus, the absence of evidence for the influence of prejudice in this data does not truly allow any conclusions about the role of prejudice in the referees' decisions. Second, although darker-skinner players received more red cards, they received fewer yellow cards, and there was no connection between skin tone and the total number of bookings (i.e., fouls that warranted penalty cards) received. One possible implication of these effects is that while skin-tone may not be related to whether a referee decides that a foul is serious enough to warrant a booking, it does influence whether a more severe foul is decided to warrant a red card resulting in an immediate ejection or a yellow card resulting in a caution. That is, for darker-skinned players, more of these severe fouls may result in red cards and fewer may result in yellow cards, which would explain the current pattern of results. Finally, there are several other possible confounds that could not be adequately assessed in this data set. Some professional soccer clubs generally play a more aggressive and physical style, which should result in more red cards for players on those clubs. It is possible that these clubs happen to have darker-skinned players, which could explain the observed association. In the absence of controls for these types of variables, and of any direct evidence for the role of prejudice, any conclusions that professional soccer referees show subtle attitudinal biases against players with darker skin when assigning red cards would be premature.

References

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (Vol. 1). Sage.