Multiple Regression Analyses Reveal High Probability of Soccer Referee Bias Based on Player Skin Tone

Authors: Jason Prenoveau[1*], Martin F. Sherman[1*]

Affiliations

[1]Loyola University Maryland.

*Correspondence to: jmprenoveau@loyola.edu or msherman@loyola.edu.

Abstract

The current statistical analyses (multiple linear and logistic regressions) were designed to determine: (1) whether there would be a relation between the skin tone color of soccer players in European leagues and the likelihood of a referee penalizing them with a red card, and (2) whether measures of the implicit and explicit racial biases of referees' countries would moderate the above-mentioned relation. After accounting for shared variability between covariates (i.e., players' position, height, weight, age, games played, and goals scored) and player skin tone color and red cards, results revealed support for a bivariate relation between player skin tone color and red cards given, but no support that racial biases acted as moderators of this relation.

One Sentence Summary

There was statistical support for a unique bivariate relation between the skin tone color of a player and the player's receiving red cards, but there was no support for either implicit or explicit biases of the referee's country acting as a moderator variable of the above mentioned relation.

Results

Relation Between Player Skin Tone and Red Cards Received

Final approach with player club variable removed. The approach was exactly the same as described below in the Final Approach section except that the variable of player club was not considered as a covariate as requested by the project leaders. Please see the Final Approach section below for additional details of the analyses conducted.

Player position was examined as a covariate because it was significantly related to both average player skin tone, $F(11, 1421) = 5.33$, $p < .001$, and total red cards, $F(11, 1674) = 8.27$, $p < .001$. Total games played was examined as a potential covariate because it was significantly related to both average player skin tone, $r(1583) = -.06$, $p = .03$, and total red cards, $r(2051) = .40$, $p < .001$. Player height was examined as a potential covariate because it was significantly related to both average player skin tone, $r(1580) = -.06$, $p = .02$, and total red cards, $r(2031) = .07$, $p = .003$. Total red cards was significantly related to player age, $r(2051) = .35$, $p < .001$, player weight, $r(1971) = .09$, $p < .001$, and total goals scored, $r(2051) = .19$, $p < .001$, so these variables were also examined as potential covariates.

A multiple linear regression analysis was conducted with total red cards received as the outcome variable, average player skin tone as the predictor variable of main interest, and all of the variables listed above included as covariates. A simultaneous regression analysis was used with forced entry of average player skin tone and all potential covariates. Whereas the overall model indicated that the variables predicted significant variability in total red cards received, $F(17, 1401) = 20.37$, $p < .001$, $R^2 = .20$, two of the potential covariates did not have significant coefficients: player weight, $\beta = .003$, $t(1401) = 0.08$, $p = .93$, and player height, $\beta = -.044$, $t(1401) = -1.14$, $p = .26$. Further, this model revealed a significant, unique relation between

average player skin tone and total red cards received after accounting for shared variability with the other predictors, $\beta = .060$, $t(1401) = 2.43$, $p = .02$.

Because the potential covariates of player weight and player height did not have significant, unique relations with total red cards received, these variables were removed from the model. Their removal did not result in a significant decrement in model fit, $\Delta F(2, 1401) = 0.95$, $p = .39$. The final model indicated that the variables predicted significant variability in total red cards received, $F(15, 1417) = 23.54$, $p < .001$, $R^2 = .20$. The final model revealed a significant, unique relation between average player skin tone and total red cards received after accounting for shared variability with the other predictors, $\beta = .062$, $t(1417) = 2.54$, $p = .01$. Conversion of this standardized regression coefficient to Cohen's *d* results in an effect size of 0.124 with a 95% confidence interval of 0.028 to 0.220.

Final approach. Average player skin tone was derived by taking the arithmetic average of the two research assistants' skin tone ratings for each player (this was justified because of the high correlation of .924, $p < .001$, between the two research assistants' ratings). Player age was calculated by choosing an arbitrary date (01/01/2013) within the soccer season under consideration and calculating player age based on their birthdate and this arbitrary date. Contrast coding was used to represent player position and player club. For each contrast, players were assigned a 0 if the position under consideration was not their position and a 1 if it was their position. Thus, 11 contrasts were used to represent the 12 possible positions. The player club contrasts were created in the same manner.

To examine the relation between average player skin tone and red cards received, the data was restructured such that each case was restructured into a single player as opposed to the original dataset where each case was a player-referee dyad. Thus, the number of cases was

reduced from 146,028 player-referee dyads to 2,053 players. As part of this restructuring, the variables representing number of games played in a player-referee dyad, number of goals scored in a player-referee dyad, and number of red cards received from a referee were summed across referee-player dyads to create the new variables of total games played, total goals scored, and total red cards received.

Missing data was handled by excluding any cases with missing data on the variables under consideration for the given analysis. Thus, as can be seen below, the sample size, and therefore the degrees of freedom, can be different for different analyses. The final multiple linear regression analysis for the relation between player skin tone and total red cards received consisted of 1,433 of 2,053 (69.8%) players. Those excluded from the final analysis were missing position data (n = 152), skin tone ratings (n = 253), or both position data and skin tone ratings (n = 215).

Multiple linear regression was used to examine the relation between the predictor variable of average player skin tone and the continuous outcome variable of total red cards received. Additional predictor variables were examined for inclusion as potential covariates because it was desired to determine if there was a unique relation between average player skin tone and total red cards received after accounting for other observed variables that might be related to player skin tone or total red cards.

Player club was examined as a covariate because it was significantly related to both average player skin tone, $F(93, 1491) = 3.20$, $p < .001$, and total red cards, $F(114, 1938) = 1.49$, $p < .001$. Player position was examined as a covariate because it was significantly related to both average player skin tone, $F(11, 1421) = 5.33$, $p < .001$, and total red cards, $F(11, 1674) = 8.27$, $p < .001$. Total games played was examined as a potential covariate because it was significantly

related to both average player skin tone, $r(1583) = -.06$, $p = .03$, and total red cards, $r(2051) = .40$, $p < .001$. Player height was examined as a potential covariate because it was significantly related to both average player skin tone, $r(1580) = -.06$, $p = .02$, and total red cards, $r(2031) = .07$, $p = .003$. Total red cards was significantly related to player age, $r(2051) = .35$, $p < .001$, player weight, $r(1971) = .09$, $p < .001$, and total goals scored, $r(2051) = .19$, $p < .001$, so these variables were also examined as potential covariates.

A multiple linear regression analysis was conducted with total red cards received as the outcome variable, average player skin tone as the predictor variable of main interest, and all of the variables listed above included as covariates. A simultaneous regression analysis was used with forced entry of average player skin tone and all potential covariates. Whereas the overall model indicated that the variables predicted significant variability in total red cards received, $F(105, 1313) = 4.45$, $p < .001$, $R^2 = .26$, two of the potential covariates did not have significant coefficients: player weight, $\beta = .006$, $t(1313) = 0.16$, $p = .87$, and player height, $\beta = -.001$, $t(1313) = -0.03$, $p = .98$. Further, this model did not reveal a significant, unique relation between average player skin tone and total red cards received after accounting for shared variability with the other predictors, $\beta = .038$, $t(1313) = 1.41$, $p = .16$.

Because the potential covariates of player weight and player height did not have significant, unique relations with total red cards received, these variables were removed from the model. Their removal did not result in a significant decrement in model fit, $\Delta F(2, 1313) = 0.02$, $p = .983$. The final model indicated that the variables predicted significant variability in total red cards received, $F(104, 1328) = 4.60$, $p < .001$, $R^2 = .27$. The final model did not reveal a significant, unique relation between average player skin tone and total red cards received after accounting for shared variability with the other predictors, $\beta = .039$, $t(1328) = 1.47$, $p = .14$.

Conversion of this standardized regression coefficient to Cohen's *d* results in a (non-significant) effect size of 0.078 with a 95% confidence interval of -0.026 to 0.182. All analyses were also examined using listwise deletion of all cases that had missing data for any of the variables under consideration (n = 1419 players). This resulted in some changes to degrees of freedom, statistical test values, p values, and magnitude of effect sizes. However, the same variables wound up in the final model and the unique relation between average player skin tone and total red cards received after accounting for shared variability with the other predictors was very similar, $\beta = .038$, $t(1315) = 1.32$, $p = .15$.

Regression assumptions were examined through visual inspection of graphical displays. To examine whether or not the form of the relation (linear relation between the outcome and predictor variables) was correctly specified, scatterplots were used to plot regression residuals against values for each predictor. Lowess lines were fit to visually inspect whether or not the mean of the residuals was zero regardless of the value of the predictor. Although visual inspection of the Lowess lines revealed slight departures from a mean of zero in places, these scatterplots seemed to indicate that the form of the relation was generally correctly specified as linear. These scatterplots were also used to visually inspect the assumption of homoscedasticity of the residuals. No large, obvious differences in the variance of the residuals for different values of the predictor variables were observed. Case numbers were randomly assigned to each case and the residuals were plotted against the case number using a scatterplot. Visual inspection did not reveal any obvious departures from independence of residuals (see discussion of non-independence of data in the section below). To examine the assumption of normality of residuals, a histogram of the regression standardized residuals was overlaid with a normal curve and a normal P-P plot of regression standardized residuals was generated to display the observed

cumulative probability versus the expected cumulative probability. Both of these displayed indications that the distribution of the residuals had somewhat heavier tails than a normal distribution. However, we were not overly concerned by this: “In large samples, nonnormality of the residuals does not lead to serious problems with the interpretation of either significance tests or confidence intervals.” (Cohen, Cohen, West, & Aiken, 2003, p. 120).

Initial approach. There were only a few changes between the initial approach and the final approach. In the initial approach, player position contrasts were included or excluded based on the significance of the coefficient associated with the individual contrast. In the final approach, all player position contrasts were included to fully represent standing on the variable. Based on the recommendation of a reviewer, player club was examined, and subsequently included, as a covariate in the final approach (but not the initial approach). Also, based on information sent to all teams by the lead authors, country of player club was removed as a potential covariate in the final approach (it was included in the original analyses), and scale of the player skin tone ratings was transformed from 1, 2, 3, 4, 5 to 0, 0.25, 0.5, 0.75, 1 for the final approach.

This approach resulted in a starting model that was similar to the one above, including player position (certain contrasts), player age, player height, player weight, total games played, total goals scored, country of player club (certain contrasts), and average player skin tone as predictors of total red cards received. The only differences in the starting model were that player club was not included as a potential covariate, only certain player position contrasts were examined as covariates (instead of all of the contrasts as done in the final approach), and certain country of player club contrasts were examined as covariates (country of player club was not examined in the final approach). As with the final approach, the overall model indicated that the

variables predicted significant variability in total red cards received, $F(17, 1401) = 21.84$, $p < .001$, $R^2 = .21$. However, this model did reveal a significant, unique relation between average player skin tone and total red cards received after accounting for shared variability with the other predictors, $\beta = .051$, $t(1401) = 2.00$, $p = .045$.

As with the initial approach, player height and weight were removed from the model and their removal did not result in significant decrement in model fit. In addition, a few player position contrasts were removed from the model and their removal did not result in significant decrement in model fit. This overall model indicated that the variables predicted significant variability in total red cards received after removal of those covariates that did not contribute significantly to the model, $F(11, 1421) = 34.5$, $p < .001$, $R^2 = .21$. This model also revealed a significant, unique relation between average player skin tone and total red cards received after accounting for shared variability with the remaining predictors, $\beta = .051$, $t(1421) = 2.09$, $p = .037$.

Both the initial and final approaches fail to account for non-independence in the data. The authors were aware of this when conducting the initial analyses and received feedback to this effect from multiple reviewers. The authors were aware that multi-level modeling would be one method for addressing this non-independence in the data and would have pursued such analyses had time permitted (and had they been analyzing this data for a manuscript where they had control over submission timing). However, given the relatively short time-frame for the present project, the authors did not feel that they had adequate time to learn multi-level modeling to the extent needed in order to knowledgably conduct analyses using this approach.

Implicit Bias of Referee Country as a Moderator of the Relation Between Player Skin Tone and Red Cards Received

Final approach with player club variable removed. The approach was exactly the same as described below in the Final Approach section except that the variable of player club was not considered as a covariate as requested by the project leaders. Please see the Final Approach section below for additional details of the analyses conducted.

The overall model indicated that the variables predicted significant variability in red card received, $\chi^2(17) = 1{,}055.82$, $p < .001$, Nagelkerke $R^2 = .07$. The interaction of implicit bias score of referee country and average player skin tone did not significantly predict red card received after accounting for shared variability with the other predictors, $e^{\beta} = 0.057$, Wald $\chi^2(1) = 1.061$, $p = .30$. Thus, implicit bias of referee country was not found to moderate the relation between player skin tone and red card received.

Final approach. Average player skin tone, player age, player position, and player club were derived as described above. For moderation analyses, the data was not restructured into single players: each case was a player-referee dyad. Thus, the original variables of number of games played in a player-referee dyad and number of goals scored in a player-referee dyad were used for the moderation analyses (not the variables of total games played and total goals scored that were used for the analyses above). The number of red cards received in a player-referee dyad was transformed to create a dichotomous red card received variable. Thus, if no red card was received, a value of zero was assigned as in the original variable. If one or more red cards were received, a value of one was assigned to indicate that the specific referee had given at least one red card to the specific player. This was done because only 0.017120% (25 of 146,028) of the player-referee dyads with red card data received more than one red card.

For the moderation analyses, both the mean implicit bias score of referee country and the average player skin tone were transformed by mean centering them. An interaction term was

then created by multiplying the mean centered implicit bias score of referee country and the mean centered average player skin tone.

As done above, missing data was handled by excluding any cases with missing data on the variables under consideration. The final multiple linear regression analysis consisted of 116,014 of 146,028 (79.5%) player-referee dyads. Those excluded from analysis were missing player position data (n = 8,454), player skin tone ratings (n = 12,134), implicit bias scores (n = 146), both player position data and player skin tone ratings (n = 9,263), both player skin tone ratings and implicit bias scores (n = 8), both player position data and implicit bias scores (n = 7), and missing player position data, player skin tone ratings, and implicit bias scores (n = 2).

Multiple binary logistic regression was used to examine implicit bias score of referee country as a moderator of the relation between the average player skin tone and the dichotomous outcome variable of red card received. To do this, mean centered implicit bias score of referee country, mean centered average player skin tone, and the interaction of these two variables were all included as predictors of red card received. The covariates identified above (player position, player club, games played, goals scored, and player age) were included as covariates. A simultaneous regression analysis was used with forced entry of these variables.

The overall model indicated that the variables predicted significant variability in red card received, $\chi^2(106) = 1{,}282.17$, $p < .001$, Nagelkerke $R^2 = .09$. The interaction of implicit bias score of referee country and average player skin tone did not significantly predict red card received after accounting for shared variability with the other predictors, $e^{\beta} = 0.329$, Wald $\chi^2(1) = 0.149$, $p = .70$. Thus, implicit bias of referee country was not found to moderate the relation between player skin tone and red card received.

Initial approach. As with the first question above, there were only a few changes between the initial approach and the final approach. In the initial approach, player position contrasts were included or excluded based on the significance of the coefficient associated with the individual contrast. In the final approach, all player position contrasts were included to fully represent standing on the variable. Based on the recommendation of a reviewer, player club was included as a covariate in the final approach (but not the initial approach). Also, based on information sent to all teams by the lead authors, country of player club was removed as a potential covariate in the final approach (it was included in the original analyses), and scale of the player skin tone ratings was transformed from 1, 2, 3, 4, 5 to 0, 0.25, 0.5, 0.75, 1 in the final approach.

The initial approach resulted in a model that was similar to the final approach above, including player position (certain contrasts), player age, games played, goals scored, country of player club (certain contrasts), and average player skin tone as predictors of red card received. The only differences between the models were that player club was not included as a covariate (it was included in the final approach), only certain player position contrasts were included as covariates (instead of all of the contrasts as done in the final approach), and certain country of player club contrasts were included as covariates (country of player club was not examined in the final approach).

As with the final approach, the overall model indicated that the variables predicted significant variability in red card received, $\chi^2(13) = 1{,}132.01$, $p < .001$, Nagelkerke $R^2 = .08$. The interaction of implicit bias score of referee country and average player skin tone did not significantly predict red card received after accounting for shared variability with the other

predictors, $e^{\beta} = 0.937$, Wald $\chi^2(1) = 0.008$, $p = .93$. Thus, implicit bias of referee country was not found to moderate the relation between player skin tone and red card received.

Explicit Bias of Referee Country as a Moderator of the Relation Between Player Skin Tone and Red Cards Received

Final approach with player club variable removed. The approach was exactly the same as described below in the Final Approach section except that the variable of player club was not considered as a covariate as requested by the project leaders. Please see the Final Approach section below for additional details of the analyses conducted.

The overall model indicated that the variables predicted significant variability in red card received, $\chi^2(17) = 1{,}042.55$, $p < .001$, Nagelkerke $R^2 = .07$. The interaction of explicit bias score of referee country and average player skin tone did not significantly predict red card received after accounting for shared variability with the other predictors, $e^{\beta} = 0.981$, Wald $\chi^2(1) = 0.002$, $p = .96$. Thus, explicit bias of referee country was not found to moderate the relation between player skin tone and red card received.

Final approach. Average player skin tone, player age, player position, and player club were derived as described above. For moderation analyses, the data was not restructured into single players: each case was a player-referee dyad. Thus, the original variables of number of games played in a player-referee dyad and number of goals scored in a player-referee dyad were used for the moderation analyses (not the variables of total games played and total goals scored). The number of red cards received in a player-referee dyad was transformed to create a dichotomous red card received variable. Thus, if no red card was received, a value of zero was assigned as in the original variable. If one or more red cards were received, a value of one was assigned to indicate that the specific referee had given at least one red card to the specific player.

This was done because only 0.017120% (25 of 146,028) of the player-referee dyads with red card data received more than one red card.

For the moderation analyses, both the mean explicit bias score of referee country and the average player skin tone were transformed by mean centering them. An interaction term was then created by multiplying the mean centered explicit bias score of referee country and the mean centered average player skin tone.

As done above, missing data was handled by excluding any cases with missing data on the variables under consideration. The final multiple linear regression analysis consisted of 116,014 of 146,028 (79.5%) player-referee dyads. Those excluded from analysis were missing player position data (n = 8,454), player skin tone ratings (n = 12,134), explicit bias scores (n = 146), both player position data and player skin tone ratings (n = 9,263), both player skin tone ratings and explicit bias scores (n = 8), both player position data and explicit bias scores (n = 7), and missing player position data, player skin tone ratings, and explicit bias scores (n = 2).

Multiple binary logistic regression was used to examine explicit bias score of referee country as a moderator of the relation between the average player skin tone and the dichotomous outcome variable of red card received. To do this, mean centered explicit bias score of referee country, mean centered average player skin tone, and the interaction of these two variables were all included as predictors of red card received. The covariates identified above (player position, player club, games played, goals scored, and player age) were included as covariates. A simultaneous regression analysis was used with forced entry of these variables.

The overall model indicated that the variables predicted significant variability in red card received, $\chi^2(106) = 1{,}277.83$, $p < .001$, Nagelkerke $R^2 = .09$. The interaction of explicit bias score of referee country and average player skin tone did not significantly predict red card received

after accounting for shared variability with the other predictors, $e^{\beta} = 1.255$, Wald $\chi^2(1) = 0.285$, $p = .59$. Thus, explicit bias of referee country was not found to moderate the relation between player skin tone and red card received.

Initial approach. As with the first question above, there were only a few changes between the initial approach and the final approach. In the initial approach, player position contrasts were included or excluded based on the significance of the coefficient associated with the individual contrast. In the final approach, all player position contrasts were included to fully represent standing on the variable. Based on the recommendation of a reviewer, player club was included as a covariate in the final approach (but not the initial approach). Also, based on information sent to all teams by the lead authors, country of player club was removed as a potential covariate in the final approach (it was included in the original analyses), and scale of the player skin tone ratings was transformed from 1, 2, 3, 4, 5 to 0, 0.25, 0.5, 0.75, 1 in the final approach.

The initial approach resulted in a model that was similar to the final approach above, including player position (certain contrasts), player age, games played, goals scored, country of player club (certain contrasts), and average player skin tone as predictors of red card received. The only differences between the models were that player club was not included as a covariate (it was included in the final approach), only certain player position contrasts were included as covariates (instead of all of the contrasts as done in the final approach), and certain country of player club contrasts were included as covariates (country of player club was not examined in the final approach).

As with the final approach, the overall model indicated that the variables predicted significant variability in red card received, $\chi^2(13) = 1{,}129.28$, $p < .001$, Nagelkerke $R^2 = .08$. The

interaction of explicit bias score of referee country and average player skin tone did not significantly predict red card received after accounting for shared variability with the other predictors, $e^{\beta} = 1.092$, Wald $\chi^2(1) = 0.707$, $p = .40$. Thus, explicit bias of referee country was not found to moderate the relation between player skin tone and red card received.

Conclusion

To examine the relation between average player skin tone and red cards received, the data was restructured such that each case was restructured into a single player as opposed to the original dataset where each case was a player-referee dyad. Thus, the number of cases was reduced from 146,028 player-referee dyads to 2,053 players. Potential covariates were identified and were examined as covariates if they correlated with player skin tone and/or number of red cards received. Covariates included in the final model included player position, total games played, player age, and total goals scored. A multiple linear regression analysis was conducted with total red cards received as the outcome variable, average player skin tone as the main predictor variable, and all of the variables listed above included as covariates. A simultaneous regression analysis was used with forced entry of average player skin tone and all potential covariates. The overall model indicated that the variables, taken together, predicted significant variability in total red cards received. This model also revealed a significant, unique relation between average player skin tone and total red cards received after accounting for shared variability with the covariates, such that darker player skin tone was associated with receiving a greater number of red cards.

In order to determine if implicit and explicit bias scores of referee countries acted as moderator variables of the relation between average player skin tone and red cards received, the original variables of a player-referee dyad were used for the unit of analyses (as opposed to the

first analysis where each player was represented only once in the data set regardless of the number of player-referee dyads). The number of red cards received in a player-referee dyad was transformed to create a dichotomous red card received variable (where 0 red card = 0, and 1 or more red cards = 1). Both the mean implicit and explicit bias scores of referee country and the average player skin tone were transformed by mean centering prior to creating the cross-product for the interaction terms. Multiple, simultaneous binary logistic regressions were used to examine the potential moderation effects. Player position, games played, goals scored, and player age were included as covariates. The overall model focusing on the moderation effect of implicit bias scores of referee countries indicated that the variables, taken together, predicted significant variability in red cards received. However, the interaction of implicit bias score of referee country and average player skin tone did not significantly predict red card received after accounting for shared variability with the other predictors. Thus, implicit bias of referee country was not found to moderate the relation between player skin tone and red card received. In addition, the overall model focusing on the moderation effect of explicit bias scores of referee countries indicated that the variables, taken together, predicted significant variability in red cards received. However, the interaction of explicit bias score of referee country and average player skin tone did not significantly predict red cards received after accounting for shared variability with the other predictors. Hence, neither implicit nor explicit biases of referee countries were found to moderate the relation between player skin tone and red cards received. In conclusion, the current analyses demonstrate a significant, unique relation between average player skin tone and total red cards received, which may indicate that referees in the European soccer leagues use player skin tone in determining the issuance of red cards. However, the implicit and explicit biases of countries of the referees do not appear to moderate the relation between player skin

tone and red cards received. A limitation of the present analyses is that they did not account for potential non-independence in the data; it is possible that use of an alternate statistical technique that would account for such non-independence, such as multi-level modeling, would provide a more accurate depiction of the relation among these variables.

Data and Output

SPSS data, syntax, and output files for both the initial analyses before the feedback round and final analyses after the feedback round can be found at the Open Science Framework (OSF): https://osf.io/bqi6d/files/.