# Investigating team penalties: A multi-method analysis of player’s skin tone and referee characteristics

S. A. Sommer[1]*, D. M. Kennedy[2]

**Affiliations**

[1]HEC Paris.

[2]University of Washington, Bothell.

*Correspondence to: S. Amy Sommer, HEC Paris, Management and Human Resources Department, 1 rue de la Libération, 78351 Jouy-en-Josas, France, SOMMERa@hec.fr.

## Abstract

We examine the research question: “do soccer referees give more red cards to dark skin toned players?” Our answer is based on the analysis of two models: (1) investigates the difference in red card penalties awarded to light skin toned players versus dark skin toned players while controlling for other factors, and (2) assesses the moderating impact of the referee’s average country bias, implicit or explicit, on the relationship between soccer player skin tone and red cards awarded while controlling for other factors. For model (1), the ANCOVA results indicate that a disparity exists in the number of red cards awarded to darker skin toned players. For model (2), the results of a negative binomial regression with a log link analysis show no significant moderating effect of referee’s average country bias, implicit or explicit, on the relationship between soccer player skin tone and red cards awarded.

## One Sentence Summary

A multi-method analysis indicates that soccer player skin tone matters for the number of red cards awarded by a referee, but this link is not augmented by the country biases of the soccer referee.

# Results

## Initial Approach

Our initial analysis sought to determine whether (1) there is a difference in red cards awarded to dark skin toned players versus light skin toned players, and (2) if there is a moderating effect of the referee's average country bias, implicit or explicit, on the relationship between soccer player skin tone and red cards awarded while controlling for other factors. The data set used to assess these questions initially contained 146,028 cases with information about player-referee dyads. That is, cases contained information about the games each player participated with a particular referee. We removed cases containing incomplete data in one or more of the variables used in the model to be analyzed via listwise deletion. In particular, since player skin tone, as rated by two independent raters, was of interest, we removed the data with missing values for rater1 and/or rater2. This resulted in 21,407 cases being removed (14.7% of the dataset) leaving 124,621 cases remaining. We then created an average rating variable (RateAve = {1, 1.5,2,2.5,3,3.5,4,4.5,5}), which is the average skin tone score of rater1 and rater2 variables given the high inter-rater agreement. Table 1 lists the variables used in our analyses, definitions, and descriptive statistics. All analyses were performed using SPSS 19.

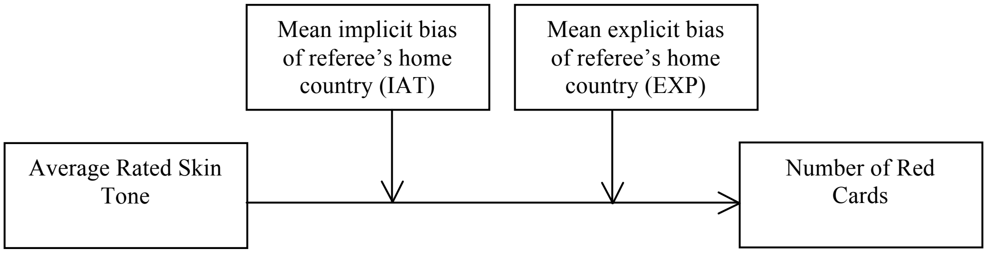
*Insert Table 1 about here*

To assess (1) the difference in red cards awarded to dark skin toned players versus light skin toned players the data was transformed to be at the player level. We collapsed the data across player's names (playerShort), keeping RateAve for the player and summing the number of red cards awarded for the player (i.e., redCards summed across all cases for a player to create RedCards_sum). The resulting dataset consisted of 1585 unique cases. By collapsing the data to the player level we removed redundancy in player representation in the dataset as well as provided clarity about the number of red cards any one player received. We conducted a series of analysis of variance (ANOVA's) to analyze the difference in the average number of red cards earned between different groups (e.g., dark skin toned players and light skin toned players). We grouped players based on their average rated skin tone (RateAve). Since the average rated skin tone was a continuous variable with a range between 1 and 5, we explored different grouping cut-off values to determine "light" skin tone and "dark" skin tone.

Our results indicate significant differences for the number of red cards awarded between different groups (e.g., dark skin toned players and light skin toned players) at different grouping cut-off values including scenario 1 (where light skin tone players were those with RateAve = [1] and dark skin tone players were those players with RateAve = (1, 5]), [$F(1,1583)=6.216$, $p<.05$)] and for scenario 2 (where light skin tone players were those with RateAve [1, 2) and dark skin tone players were those players with RateAve = [2, 5]), [$F(1,1583)=9.542$, $p<.01$)]. No significant differences were found for scenario 3 (where light skin tone players were those with RateAve [1,3) and dark skin tone players were those players with RateAve = [3, 5]), [$F(1,1583)=.765$, $p=.382$], scenario 4 (where light skin tone players were those with RateAve [1,4) and dark skin tone players were those players with RateAve = [4, 5]), [$F(1,1583)=.439$, $p=.508$)], or scenario 5 (where light skin tone players were those with RateAve [1,5) and dark skin tone players were those players with RateAve = [5]), [$F(1,1583)=1.106$, $p=.293$)]. In sum, the results indicate that a disparity in red cards awarded is evident when comparisons are made between groups of very light skin tone players (RateAve = [1,2)) and darker skin toned players (RateAve = [2,5]).

To answer (2), our analysis model, depicted in Figure 1, is a moderated relationship between average rated skin tone and red cards by mean level of implicit prejudice (i.e., meanIAT) and mean level of explicit prejudice (i.e., meanEXP) by the referee's home country. To assess the model we use a hierarchical multiple regression on the un-collapsed data parsed by player-referee dyad. We assumed a normal distribution of the number of red cards awarded in order to use this approach. We removed the data with missing values so every case had full values in the final analysis for the regression sample (*n*=116,014). In addition to the variables of interest we recognized that other factors may explain variation in the dependent variable and therefore, need to be controlled. In particular, we controlled for games played using the continuous variable in the dataset because players that participate in more games have more opportunity to receive red cards than players that participate in fewer games. Additionally, we controlled for the player position because certain soccer positions may be more prone to receiving red cards than other positions. Since there were essentially four player positions (i.e., forward, midfielder, back and goalkeeper) represented in the dataset, we created three dummy variables per convention (Aiken & West, 1991).

**Figure 1. Conceptualization of moderated relationship between implicit and explicit skin-tone prejudice, red card penalties, and soccer player's skin tone.**


Mean implicit bias of referee's home country (IAT)
Mean explicit bias of referee's home country (EXP)
Average Rated Skin Tone
Number of Red Cards

The result of the hierarchical regression analysis with all variables in the equation were significant [$R^2$ = .016 and adjusted $R^2$ = .016 (*F*(2, 116 003)= 191.102, *p*<.001)]. For the regression coefficients, 95% confidence limits were calculated. There was a significant positive relationship between skin rating and the number of red cards [*b*=.001, *t*(10, 116 013)=3.264, *p*<.001, the 95% confidence limits were .0 to .002]. The interactions were not significant for RateAve and meanIAT [*b*=-.011, *t*(10, 116 003)= -.883, *p*=.377, and the 95% confidence limits were from .034 to .013)]. There was a significant positive relationship between redCards and meanIAT [*b*=.033, *t*(10, 116 003)=3.264, *p*<.05, the 95% confidence limits were from .001 to .066], and there was no significant relationship between the number of red cards and meanEXP [*b*=.001, *t*(10, 116 003)= .492, *p*=.623, the 95% confidence limits were -.004 to .006]. The interactions were not significant for RateAve and meanEXP [*b*=.000, *t*(10, 116 003)=.147 *p*=.883, and the 95% confidence limits were from -.003 to .004]. Since the interaction effects were not significant, then the moderating effects of the referee's average country bias, implicit or explicit, on the relationship between soccer player skin tone and red cards awarded while controlling for other factors is not supported.

**Final Approach**

Our analysis sought to determine whether (1) there is a difference in red cards awarded to dark skin toned players versus light skin toned players while controlling for other factors, and (2) there is a moderating effect of the referee's average country bias, implicit or explicit, on the relationship between soccer player skin tone and red cards awarded while controlling for other factors. Based on reviewer feedback we amended our initial analytical approach to address spurious variables in the examination of question (1) and the lack of normality in the distribution of red cards awarded in the examination of question (2). For our final approach we created RaterAve, which is the average skin tone score of the scaled rater1 and scaled rater2 variables given the high inter-rater agreement (RateAve = {0,0.25,0.5,0.75,1}). Also, we created dummy variables for the player's position, referee country, and league country. There were 124, 468 usable cases (85.2%) after using listwise deletion (14.8% of the cases were deleted for incomplete data in one or more of the variables in the model, *n*=21,560). Table 1 reports the variables included in our final approach. All analyses were performed in SPSS 19.

To answer (1) we used an Analysis of Covariance (ANCOVA) in order to test differences in the number of red cards earned between different groups (e.g., dark skin toned players and light skin toned players) and control for a number of variables. Specifically, we control for referee country, league country, player's position, games, victories, and referee country prejudices, both implicit and explicit. ANCOVA is robust to violations of normality assumptions (e.g., Rutherford, 2001). Using an ANCOVA, we found a significant main effect for rater average (RaterAve) [*F*(8,124 468)=3.778, *p*<.001], suggesting the number of red cards received varies significantly with skin color. There were 8 levels of rater average. The significant beta coefficients were: -.003 for forward position (Dumi1_forward *t*=-2.557, with a 95% confidence interval from -.005 to -.001), -.002 for midfielder position (Dumi2_mid *t*=-2.478, with a 95% confidence interval from -.004 to -.001), .004 for back position (Dumi3_back *t*=3.698, with a 95% confidence interval from .002 to .006), .034 mean implicit bias (meanIAT *t*=2.162, with a 95% confidence interval from .003 to .065), -.007 mean explicit bias (meanExp *t*=-3.000, with a 95% confidence interval from -.011 to -.002), -.004 [RaterAve =.00] (*t*=-2.910 with a 95% confidence interval from -.007 to -.001), -.004 [RaterAve =.13] (*t*=-2.574, with a 95% confidence interval from -.007 to -.001), and -.009 [RaterAve =.63] (*t*=-3.464, with a 95% confidence interval from -.014 to -.004). There was a marginally significant beta coefficient for .001 league country dummy code (leagueCountry_NUM *t*=1.929, with a 95% confidence interval from -9.226E-6 to .001, or p=.054); and the remaining coefficients were not significant for [RaterAve =.25], [RaterAve =.38], [RaterAve =.50], [RaterAve =.25], [RaterAve =.38] and [RaterAve =.50]. Overall, the results indicate that a disparity in red cards awarded is evident when comparisons are made between groups of light skin tone players and darker skin toned players.

To address (2), we apply negative binomial regression with a log link analysis to examined the relationship between skin tone, referee country prejudices (meanIAT and meanEXP) and number of red cards because we were looking for predictive relationships among variables. Since the number of red cards is a count variable with a negative distribution, we used a negative binomial regression. More specifically, to test this research question we employed the following four step procedure: (i) the independent variables were mean centered in preparation for regression analysis. The variables included RaterAve, meanIAT and meanEXP. The control variables included: games, victories and dummy codes for position (Dumi1_forward, Dumi2_mid, and Dumi3_back), league country (leagueCountry_NUM) and referee country (refCountry); (ii) the interaction terms were created using the mean centered data. The term IATxRatrA was created from multiplying the variables of meanIAT and RaterAve. The term ExpxRatrA was created from multiplying the variables of meanEXP and RaterAve; (iii) the negative binomial regression model with a log link analysis was initiated using the variable redCards (i.e., number of red cards player received from referee) as the dependent variable, since 0=no red card and 1=red card; (iv) the predictor variables were added to the model including: controls(games, D1, D2, D3, victories, leaguecountry and refCountry), independent variable RaterAve, moderator variables meanIAT and meanEXP, and interaction terms IATxRatrA and ExpxRatrA.

Using negative binomial regression, the goodness of fit of the final model was acceptable [Pearson Chi-Square (123,355)=107,268]. The omnibus test was significant (Likelihood Ratio Chi-Square (1112)=2135, *p*<.001). The test of model effects was significant for the Wald Chi-Square for games (32)=587.677, *p*<.001, victories (22)=79.726, *p*<.001, Dumi1_forward (1)=11.754, *p*<.001, Dumi2_mid(1)=12.423, *p*<.001, Dumi3_back (1)=5.992, *p*<.014, and RaterAve(3)=35.996, *p*<.001. The remaining variables were not significant including leagueCountry_NUM , refCountry, meanIAT_C, meanExp_C, and the interaction terms (RaterAve_C * meanIAT_C, and RaterAve_C * meanExp_C). Therefore, referee country implicit skin tone prejudice was not significantly related to the number of red cards received by dark skin-toned players.

## Conclusion

This research was driven by the research question "do soccer referees give more red cards to dark skin toned players?"  Our initial approach decomposed the question to examine two models that tested (1) the fundamental difference in red cards awarded to light skin toned players versus dark skin toned players, and (2) the way referee's average country bias, implicit or explicit might compound any relationship between soccer player skin tone and red cards. In our initial approach we sought simplicity as the tradeoff to a number of assumptions regarding the need for control variables (to assess model 1) and the data distribution of the dependent variable (to assess model 2). Given the constructive feedback of reviewers we were motivated to change our final approach to methods that allowed us to address the concerns about spurious variables and the non-normality of the data distributions. Essentially, however, we obtained very similar results across our initial and final approaches suggesting robustness in the findings. Specifically, based on our results we suggest that dark skin toned soccer players receive a higher number of red cards than light skin toned soccer players, even after controlling for other variables. However, we did not find evidence that the referee's average country bias, implicit or explicit, significantly compounded the relationship between soccer player skin tone and red cards awarded. Overall, this multi-method analysis indicated that soccer player skin tone matters for the number of red cards awarded by a referee, but this link is not augmented by the soccer referee's country biases.

## Table

Table 1. Variable Names and Definitions in the Initial and Final Approach

| ***Initial Approach*** | | | |
|---|---|---|---|
| Variable Name | Definition | Initial Approach | Final Approach |
| Dumi1_forward | Center Forward, Right Winger, and Left Winger | X | X |
| Dumi2_mid | Center Midfielder, Right Midfielder, and Left Midfielder, Attacking Midfielder, Defensive Midfielder | X | X |
| Dumi3_back | Center Back, Left Fullback, and Right Fullback | X | X |
| Games | Number of games in the player-referee dyad | X | X |
| LeagueCountry_NUM | Country of player club (England, Germany, France, and Spain) recoded numerically | | X |
| MeanEXP | The mean explicit bias score (using a racial thermometer task) for referee country, higher values correspond to greater feelings of warmth toward whites versus blacks | X | X |
| MeanIAT | The mean implicit bias score (using the race IAT) for referee country, higher values correspond to faster white \| good, black \| bad associations | X | X |
| RateAve | Average of skin rating of photos by rater 1 and rater 2 | X | X |
| RedCards | Number of red cards player received from the referee | X | X |
| RefCountry | Unique referee country ID number | | X |
| Victories | Number of games won | X | X |

## References and Notes

Aiken, L. S., & West, S. G. (1991). *Multiple regression: Testing and interpreting interactions.* Newbury Park, CA: Sage.

Rutherford, A. (2001). *Introducing ANOVA and ANCOVA: A GLM approach.* London, UK: Sage.