A Poisson GLM approach predicts general bias in the distribution of red cards in soccer

Authors: A.J. Hofelich Mohr[1]*, T.A. Lindsay[1]

Affiliations

[1]Research Support Services, College of Liberal Arts Office of Information Technology, University of Minnesota.

*Correspondence to: hofelich@umn.edu

Abstract

Looking at first division male soccer players from Spain, Germany, England, and France (*N* = 2,053) and the referees they have encountered in their lifetime of play (*N* = 3,147), we tested two research questions: 1) Are referees more likely to give red cards to dark skin toned players than light skin toned players? and 2) Are referees from countries high in skin-tone prejudice more likely to award red cards to dark skin toned players? We used Poisson multi-level modeling to test whether the number of red cards (offset by number games) would be predicted by skintone and whether this effect would be influenced by prejudice measures, controlling for a players' position, and including referee and player as random effects. We found that referees were 41% more likely to award red cards to dark versus light skin toned players, but this effect was not influenced by implicit or explicit prejudice.

One Sentence Summary

Our analysis supports the hypothesis that referees are more likely to give red cards to players with darker, versus lighter, skin, but this effect was not influenced by implicit or explicit measures of racial bias collected from the referees' home country.

Results

The dependent variable for all analyses was the number of red cards given within a player-referee dyad. Because of the count nature of this data (obtained over a variable number of games played within each dyad), we decided a Poisson distribution was the best fit for the data. We used the log number of games for each dyad as the offset variable in each analysis to account for the fact that the overall number of red cards is directly related to the number of interactions a dyad has had. Because the mean and variance of the Poisson distribution is characterized by a single parameter, λ, we confirmed that the mean ($M = 0.0125$) and the variance ($\sigma^2 = 0.0127$) of the sample were indeed equal (at least approximately), and that overdispersion was not an issue. We also noted that there were many zeros in the dataset, reflecting the rare (i.e., low probability) nature of getting a red card in a game. However, the Poisson distribution is particularly well suited for modeling rare events that occur across a very large number of observations (Nussbaum, Elsadat, & Khago, 2008), and we did not see any indication that the large number of zero events was due to anything other than the rare nature of red cards (e.g., we did not believe them to have been generated by a different distribution, as the zero-inflated Poisson assumes; this would be the case, for example, if some players were not on the field and therefore had no chance of getting a red card during a game).

We then checked the inter-rater reliability between the two raters of skintone, which was high ($r = .92$). These ratings were averaged to create a single skintone variable. No complete cases were excluded from any analysis, and missing variables were excluded pair-wise. All analyses were conducted using R 3.0.2 on OSX 10.9.3 and an alpha of .05 was used for all tests.

Initial Approach

Because the rating of skintone was done on a five point Likert scale by two raters, we were concerned about interpreting this ordinal scale as if it were a continuous, interval measure of skintone, as many such ratings often do not conform to these assumptions. Instead, we dichotomized it to create a new variable that treated average ratings of less than 3 as "light skin" and average ratings greater than 3 as "dark skin". We deliberately left out players from this dichotomy who were rated by both raters to be "neutral". We felt this would capture the relevant differences in light versus dark skintone, without making assumptions about the linearity of the rating measure.

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

To test this hypothesis, we ran a glm with a Poisson link investigating whether the number of red cards (offset by *log*(games)) differed by skintone. We included position in the model, as we felt that some positions (e.g., goal keeper) would be inherently less likely to receive red cards during a game than other positions. The results indicated that while dark skin toned players were 12.3% more likely to get a red card than light skin players, this this was not significantly different from zero, $\exp(\beta) = 1.123$, 0.116, $z(107774) = -1.6$, $p = .11$. Our conclusion from this test was that soccer referees were not more likely to give red cards to dark skin toned players compared to light skin toned players.

Hypothesis 2: Are soccer referees from countries high in skintone prejudice more likely to award red cards to dark skin toned players?
To test this hypothesis, we first aggregated the data by referee country and skin color of player, creating a new dataset with separate variables for the number of games played and red cards to given dark and light skin toned players for each referee country. Initially, we tried creating a "relative risk" ratio variable comparing the probability of giving a red card to a dark versus light skin toned players ([*#red cards/#games* dark] / [*#red cards/#games*] light). However, because many probabilities were zero, this left a very small number of data points in the analysis.

Instead, we decided to model the number of red cards as a Poisson distribution as in Hypothesis 1, with position included in the model. However, because we specifically were testing whether the probability of awarding cards to dark skin players changed with the measures of skintone prejudice from the referees' countries, the number of red cards to dark skin toned players was used as the dependent variable (with *log*(games with dark skin toned players) as the offset). As the number of red cards given to dark skin toned players is likely a function of the general propensity to give red cards to any player, we included the probability of giving a light skin toned player a red card as a covariate in the model. Position was included as above, and the implicit (mean IAT) and explicit (mean Exp) scores were included as predictors separately in each model. Our results indicated neither measure of skintone significantly predicted increased red cards to dark skin toned players: IAT estimate: $\exp(\beta) = 1.68$ (increase in probability of a red card for dark skin toned players with one unit increase in IAT score), $z(123) = 0.24$, $p = .81$; Exp estimate: $\exp(\beta) = 1.04$ (increase in probability with one unit increase in Exp score), $z(123) = 0.11$, $p = .92$.

Final Approach
After receiving feedback from other groups and responding to other analyses, our approach changed in several ways. First and most importantly, we realized we had neglected to account for the repeated instances of referees and players in the dataset (combinations were unique, but non-independence remained as each were repeated across dyads). We decided to address this critique in our final analysis by including referee and player as random effects in our models. We also changed our approach to Hypothesis 2 in order to accommodate this added level. Second, we decided to use the measure of skintone as a continuous variable instead of dichtomizing it. Many reviewers questioned why we had dichotomized the variable, and although we believe our concerns about the linearity of the measure are still valid, we based our decision on literature suggesting a dichotomized representation of skintone is inappropriate for non-US cultures, such as Brazil (e.g., Lovell, 2000), and that such a grouping of light versus dark may not be appropriate for everyone in this sample.

Hypothesis 1: Are soccer referees more likely to give red cards to dark skin toned players than light skin toned players?
To test this hypothesis, we ran a Poisson distributed GLM using number of red cards as the dependent variable, with *log*(games) as the offset variable. Position was included as a covariate, and referee and player were included as random effects. The continuous measure of skin tone was used as the predictor of interest. The results indicated that skin tone significantly predicted the number of red cards given, with the probability of a red card increasing 41% for dark versus

light skin tone (one unit increase in skin tone), exp(β) = 1.41 [95% CI 1.13 – 1.75] , z = 3.08, p = .002.

Hypothesis 2: Are soccer referees from countries high in skintone prejudice more likely to award red cards to dark skin toned players?
To determine whether the effect in Hypothesis 1 would be stronger for referees from countries with high skintone prejudice, two additional models were run with the implicit (mean IAT) and explicit (mean Exp) measures included separately. Because we were interested in whether the positive relationship between skintone and red cards would change as the referees' country-level prejudice increased (as measured by the same IAT/Exp scores given to each referee within the same country), the interaction between IAT or Exp and skintone was modeled and was the predictor of interest. Neither interaction was significant, IAT: exp(β) = 0.13 [95% CI .0002 – 68.06], z = -0.65, p = .52; Exp: exp(β) = 1.09 [95% CI 0.42 – 2.84], z = 0.18 , p = .86, suggesting that the general skintone bias found in the distribution of red cards was not influenced by whether the referee was from a country high in skintone prejudice.

Conclusion

Our analysis confirms the hypothesis that soccer referees are more likely to give red cards to dark skin toned players than to light skin toned players, but we found no evidence that this effect was related to whether a referee came from a country with high or low skintone prejudice (based on implicit and explicit measures).

While we ultimately decided to use the skintone variable continuously to capture potential effects from more granular changes in skintone (rather than a binary light/dark), we remain concerned about the quality and linearity of the skintone ratings since we do not fully know the details about how these ratings were established; this could affect the validity of our conclusions. However, there is compelling research that skintone as a range is a strong predictor of bias, and despite its flaws, we feel the data were well suited to test the first research question. That being said, we do feel the results are restricted in generalizability by nature of the sample. Because the data was from top tier European leagues, there is a high prevalence of European referees, which is likely not representative of the entire population of soccer referees. Further, dyads that include referees from non-Eurpean countries are likely from exceptional or unusual situations (e.g., exceptional players playing in lower-tier leagues early in their careers, international friendlies, etc.), which could also systematically influence both players' likelihood of engaging in behavior that may lead to a red card and referees' likelihood of giving a red card.

We felt the dataset was far less well suited to test the second research question. While the measures of explicit and implicit bias were potentially promising sources of data to measure skintone prejudice, we do not know the quality or representativeness of these measures, and expect both would vary significantly from country to country. Even if these measures perfectly represented a country's attitude, we do not know the extent to which a given referee identifies with their country (for example, did they live there only a short time?) or whether they share the "popular" opinion. All these factors make it very difficult to draw meaningful conclusions about this hypothesis, and we expect that a different (although understandably more difficult, if not impossible, to obtain) dataset would be needed to answer this question with reasonable certainty.

References

Nussbaum, E., Elsadat, S., & Khago, A. (2008). 21 Best Practices in Analyzing Count Data Poisson Regression. In J. Osborne (Ed.), Best practices in quantitative methods. (pp. 306-324). Thousand Oaks, CA: SAGE Publications, Inc. doi: http://dx.doi.org.ezp1.lib.umn.edu/10.4135/9781412995627.d26

Lovell, P.A. (2000) Gender, race, and the struggle for social justice in Brazil. *Latin American Perspectives, 27*(6), 85-102.

Data and Output

Can be found at https://osf.io/ritk7/