Soccer player's skin tone retrodicts probability of receiving a red card

Authors:

F. D. Schönbrodt[1]*, M. Heene[1]

Affiliations

[1]Ludwig-Maximilians-Universität München, Germany.

*Correspondence to: felix@nicebread.de

Abstract

Using an extensive data set of all soccer players from 4 European leagues we investigated the research question whether soccer referees were more likely to give red cards to dark skin toned players than light skin toned players. For statistical analysis, we employed a generalized linear mixed effects model (GLMM) which modeled the probability for each player of receiving a red card in a game. The analyses revealed that the darkest skin toned players have a 1.5 times higher risk of receiving a red card than the lightest skin toned players (OR = 1.48, 95% CI [1.20; 1.84]). Referees showed a considerable variability in their susceptibility to skin tone. This variability, however, could not be explained by the average implicit or explicit racial biases of their countries of origin.

One Sentence Summary

Dark skin toned players received 1.5 times more red cards than light skin toned players, an effect that could not be explained by the average racial biases of the referee's countries.

Results

All calculations were performed in the R Environment for Statistical Computing (R Core Team, 2014) using the following packages: *lme4* (Bates, Maechler, Bolker, & Walker, 2014), *party* (Hothorn, Hornik, & Zeileis, 2006), and *psych* (Revelle, 2014). The full script of the analyses is in the online appendix at OSF.

Initial Approach

Participants. We removed all participants without any skin tone rating, assuming that this information was missing completely at random. This reduced the data set by 14.7% to 124,621 unique player-referee dyads.

Variable transformations. The skin tone ratings of both raters were averaged after assessing the inter-rater reliability ($ICC_{(2, k)} = .96$). The scale ranged from 0 to 1, where 0 represents the lightest and 1 the darkest skin tone rating. Player positions that were missing in the data set were not removed but defined as "*unknown*". The position variable then was deviation-coded. This coding scheme compares all but one position to the grand mean of all positions, and the intercept represents the average position. From overall 161 referee countries, we selected the four most frequent countries (Germany, England, France, and Spain), representing 61.3% of all player-referee dyads. Referees from all other countries were assigned to the category "*other country"*. This categorical variable also was coded using a deviation coding scheme. Finally, we *z* standardized the explicit and the implicit bias score of the referee countries.

Statistical model (exploration and covariate selection). We used binary classification trees from the *party* package (Hothorn, Hornik, & Zeileis, 2006) to explore the impact of potential covariates and their interactions[1].
Binary classification trees aim to predict an outcome by splitting predictor variables from a potentially large set of predictors into binary parts, which are then combined into a hierarchical tree of decisions. Although many potential variable splits are investigated, the method implemented in the *party* package automatically corrects for multiple testing and ensures that the combined Type-I error rate does not exceed a specified rate. Therefore it is not necessary to cross-validate the results in a fresh data set.
Besides clear main effects of players' positions and the referees' country, one potentially meaningful interaction emerged between the referees country and skin tone, in a way that skin tone is only relevant for German, English, and "other countries", but not for French and Spanish referees. This binary classification tree, however, does not take the multilevel structure of the data set into account (see also below). Therefore we did not treat this result as conclusive, but rather used the classification tree as a tool for generating hypotheses, which were then more formally tested in the following steps.
To summarize, based on this exploratory analysis, we included main effects for position, referee country, and skin tone, as well as an interaction term between skin tone and referee country in our final models, which are described in the next section.

[1]We included variables skintone, position, meanIAT (*z* standardized), meanExp (*z* standardized), referee country, number of goals, yellow cards, yellow-red cards, age, and overall number of games.

Statistical model (hypothesis tests). The data set has a cross-classified nested structure, which potentially introduces a non-independency of the data: Most referees met several players, and most players met several referees. We therefore decided to employ a multilevel logistic regression model, with random intercepts for players and referees. The *lme4* package was used to estimate the models (Bates, Maechler, Bolker, & Walker, 2014). The random intercept of referees models the variability in referee's leniency, and the random intercept of players models variability in their individual propensity to get a red card. The outcome variable was the binary result whether a player received a red card in a unique encounter with a referee or not. The focal predictor for RQ1 was the skin tone rating, and player's position and referee's country were added as covariates. For RQ2, implicit and explicit racial bias, and their interaction with skin tone were added to the model. We built increasingly complex models:

- "*m0*" = Baseline model: Position and referee country were added as fixed effect covariates, along with random intercepts for player and referee.
- "*m1*": Skin tone was added as fixed effect predictor. (In the final analysis, see next section, this slope was allowed to vary between referees). This is the main model for RQ1.
- "*m1b*": Based on the explorative results, we added an interaction between skin tone and referee country.
- "*m2a*": This model extends model *m1* by the main effect of *implicit* racial bias and its interaction with skin tone. The interaction term tests RQ2a: If this interaction is positive, the skin tone has a stronger predictive value on red cards for referees from countries with higher implicit racial bias.
- "*m2b*": This model extends model *m1* by the main effect of *explicit* racial bias and its interaction with skintone. Again, the interaction term tests RQ2b: If this interaction is positive, the skin tone has a stronger predictive value on red cards for referees from countries with higher explicit racial bias.

Models were compared using χ^2-likelihood-ratio tests and ΔAICc (Burnham, Anderson, & Huyvaert, 2011). All reported confidence intervals (CIs) are 95% CIs computed with the Wald method.

Final Approach

Based on a suggestion from the feedback round, we additionally allowed the slope for skin tone to vary between referees because some referees might be more reactive to the skin tone of a player than others. The fixed effect for skin tone is the average effect, and the random effect quantifies the variance of the slopes. Although the inclusion of this random slope did not change the conclusions of the model, it has been recommended to generally include random slopes of the focal predictors in a hypothesis testing scenario if the design allows a random variation across the units of analysis (Barr, Levy, Scheepers, & Tily, 2013). Furthermore, in the initial approach, we split the data set into a training and a validation set. As cross-validation is not necessary with the *party* package, we skipped this step.

Conclusion

Model comparisons revealed that the interaction of referee country and skin tone, which was suggested by the exploratory analysis, did not improve the model ($\Delta\chi^2(4) = 3.48$, $p = .482$; ΔAICc = 4.53). Therefore we discarded this model.
Table 1 reports the parameter estimates from models *m0*, *m1*, *m2a*, and *m2b*. Concerning the covariates, we found that German and English referees give less, and Spanish more red cards than the overall average. Furthermore, the center back position receives more red cards than the average position, and some position receive less red cards (e.g., attacking or right midfielder).

Concerning RQ1 (see model *m1*), the skin tone predictor has a significant positive slope of 0.40 [0.18; 0.61], and the inclusion of this predictor improves the model fit compared to the baseline model *m0* ($\Delta\chi^2(3) = 11.85$, $p = .008$; ΔAICc = -5.85). This indicates that players with darker skin have a higher probability of receiving a red card. The effect size of this predictor, expressed as an odds-ratio (OR) is 1.48 [1.20; 1.84]. Expressed in probabilities, this OR corresponds to an increase from a probability of 0.28% for light-tone players (skin tone = 0) to 0.42% for the darkest players (skin tone = 1), for an average position and average referee's country.
Concerning RQ2, there was no evidence that the average implicit or explicit racial biases of the referee's countries modulated the slope of the skin tone effect.
The variances of the random intercepts indicate that there is more variability between players (0.28) than between referees (0.16). Furthermore, there is variance in the skintone slope of referees (0.03). The slopes of 90% of referees are estimated to be between 0.10 and 0.69. These individual differences in susceptibility to skin tone, however, can neither be explained by the country of origin, nor by the implicit or explicit biases in these countries.

To summarize, our analyses revealed quite strong evidence for a skin tone effect, in a way that darker players received more red cards than lighter players. Reported as raw probabilities, the effect seems very small, due to the fact that red cards simply are very rare events. Expressed as relative probabilities, however, the darkest skin toned players have a 1.5 times higher risk of receiving a red card than the lightest skin toned players. Due to the non-experimental nature of the data, however, we are cautious to draw causal conclusions. This effect might be caused by a racial bias of the referees, but could also be caused by different playing styles of dark skinned players, which themselves could be due to, for example, genetic factors, cultural factors, or group dynamic factors within the soccer teams.

Tables

Table 1. Parameter estimates of the GLMM models.

Fixed effects	**m0**	**m1**	**m2a**	**m2b**
(Intercept)	-5.74 [-5.83; -5.66]*	-5.87 [-5.97; -5.76]*	-5.86 [-5.97; -5.76]*	-5.86 [-5.97; -5.75]*
position - Attacking Midfielder	-0.41 [-0.66; -0.17]*	-0.39 [-0.63; -0.14]*	-0.39 [-0.63; -0.14]*	-0.39 [-0.63; -0.14]*
position - Center Back	0.59 [0.46; 0.73]*	0.60 [0.46; 0.73]*	0.60 [0.46; 0.73]*	0.60 [0.46; 0.73]*
position - Center Forward	-0.08 [-0.26; 0.10]	-0.10 [-0.28; 0.08]	-0.10 [-0.28; 0.07]	-0.10 [-0.28; 0.08]
position - Center Midfielder	0.04 [-0.26; 0.34]	0.07 [-0.23; 0.37]	0.07 [-0.23; 0.37]	0.07 [-0.23; 0.37]
position - Defensive Midfielder	0.06 [-0.12; 0.24]	0.05 [-0.13; 0.23]	0.05 [-0.13; 0.23]	0.05 [-0.13; 0.23]
position - Goalkeeper	0.16 [-0.04; 0.35]	0.19 [-0.01; 0.39]	0.19 [-0.01; 0.39]	0.19 [-0.01; 0.39]
position - Left Fullback	0.15 [-0.06; 0.36]	0.15 [-0.06; 0.36]	0.15 [-0.06; 0.36]	0.15 [-0.07; 0.36]
position - Left Midfielder	-0.03 [-0.29; 0.23]	-0.03 [-0.29; 0.23]	-0.03 [-0.29; 0.23]	-0.03 [-0.29; 0.23]
position - Left Winger	-0.36 [-0.71; -0.02]*	-0.39 [-0.73; -0.04]*	-0.39 [-0.74; -0.04]*	-0.39 [-0.74; -0.04]*
position - Right Fullback	0.00 [-0.23; 0.22]	-0.01 [-0.23; 0.22]	-0.01 [-0.23; 0.22]	-0.01 [-0.23; 0.22]
position - Right Midfielder	-0.43 [-0.78; -0.08]*	-0.41 [-0.76; -0.07]*	-0.41 [-0.76; -0.07]*	-0.42 [-0.76; -0.07]*
position - Right Winger	0.03 [-0.26; 0.31]	-0.01 [-0.30; 0.27]	-0.01 [-0.30; 0.27]	-0.01 [-0.30; 0.27]
countryCode - Germany	-0.16 [-0.30; -0.01]*	-0.12 [-0.26; 0.03]	-0.10 [-0.25; 0.05]	-0.10 [-0.26; 0.05]
countryCode - England	-0.21 [-0.36; -0.06]*	-0.23 [-0.38; -0.08]*	-0.19 [-0.36; -0.03]*	-0.21 [-0.37; -0.06]*
countryCode - Spain	0.24 [0.10; 0.37]*	0.25 [0.11; 0.39]*	0.21 [0.05; 0.37]*	0.23 [0.07; 0.38]*
countryCode - France	0.15 [0.00; 0.30]	0.11 [-0.04; 0.26]	0.13 [-0.02; 0.29]	0.14 [-0.02; 0.29]
skin tone		**0.40 [0.18; 0.61]***	0.40 [0.18; 0.61]*	0.41 [0.20; 0.63]*
Implicit racial bias			0.09 [-0.14; 0.32]	
skin tone X Implicit racial bias			**0.02 [-0.35; 0.38]**	
Explicit racial bias				0.03 [-0.19; 0.24]
skin tone X Explicit racial bias				**0.13 [-0.22; 0.49]**
AIC	15462.60	15456.75	15457.51	15457.38
BIC	15647.53	15670.88	15691.07	15690.95
Log Likelihood	-7712.30	-7706.38	-7704.75	-7704.69
Deviance	15424.60	15412.75	15409.51	15409.38
Num. obs.	124621	124621	124468	124468
Num. groups: referee ID	2978	2978	2967	2967
Num. groups: player ID	1585	1585	1585	1585
Random effect variances				
Variance: referee ID (Intercept)	0.12	0.16	0.16	0.16
Variance: player ID (Intercept)	0.28	0.28	0.28	0.28
Variance: referee ID X skin tone		0.03	0.03	0.03
Variance: Residual	1.00	1.00	1.00	1.00

Note. * Zero outside the 95% confidence interval. Coefficients relevant for research questions are printed in boldface.

References and Notes

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*, 255–278. doi:10.1016/j.jml.2012.11.001

Bates, D., Maechler, M., Bolker, B. M., & Walker, S. (2014). *lme4: Linear mixed-effects models using Eigen and S4.* Retrieved from http://arxiv.org/abs/1406.5823

Burnham, K. P., Anderson, D. R., & Huyvaert, K. P. (2011). AIC model selection and multimodel inference in behavioral ecology: Some background, observations, and comparisons. *Behavioral Ecology and Sociobiology*, *65*, 23–35. doi:10.1007/s00265-010-1029-6

Hothorn, T., Hornik, K., & Zeileis, A. (2006). Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistics*, *15*, 651–674.

R Core Team. (2014). *R: A language and environment for statistical computing*. Vienna, Austria. Retrieved from http://www.R-project.org/

Revelle, W. (2014). *psych: Procedures for psychological, psychometric, and personality research.* Evanston, Illinois: Northwestern University. Retrieved from http://CRAN.R-project.org/package=psych