Darker skin-toned soccer players receive more red cards than lighter skin-toned players: A mixed effects logistic regression analysis

Authors: Frederik Aust[1*] & Fabia Högden[1]

Affiliations

[1]Department for Research Methods and Experimental Psychology, University of Cologne, Germany.

*Correspondence to: frederik.aust@uni-koeln.de.

Abstract

To test if soccer referees give more red cards to dark skin-toned players we conducted a mixed effects logistic regression analysis with crossed random effects for referees and players. Specification of crossed random effects allows generalization of our results beyond the referees and players of the sample at hand and accounts for the correlational structure of the data (e.g., clustering due to individual differences). Our analysis revealed that soccer players with darker as opposed to lighter skin tone receive more red cards throughout their career controlling for players' position and league country, $OR_{lightest,darkest}$ = 1.382 [1.120, 1.705]. The effect of players' skin tone was not moderated by explicit or implicit racial prejudice in the referees' home countries. Thus, the reasons for the increased number of red cards given to players with darker skin tone remain ambiguous.

One Sentence Summary

A mixed effects logistic regression analysis with crossed random effects for referees and players revealed that soccer players with darker as opposed to lighter skin tones receive more red cards ($OR_{lightest,darkest}$ = 1.382 [1.120, 1.705]) regardless of explicit or implicit racial prejudice in the referees' home countries.

Results

The analyzed dataset consisted of player–referee dyads representing all matches in which players and referees encountered each other. Data of 2,052 soccer players playing in the first male divisions of England, Germany, France and Spain in the 2012-2013 season and all 3,147 referees that these players played under in their professional career were included yielding a total of 146,028 dyads. Thus, multiple dyads could relate to the same players or referees. Due to this correlational structure within the dataset, we employed a generalized linear mixed effects modeling approach (e.g., Bolker et al., 2009; Zuur, Ieno, Walker, Saveliev, & Smith, 2009) with crossed random effects for players and referees (e.g., Baayen, Davidson, & Bates, 2008; Carson & Beeson, 2013; Quené & van den Bergh, 2008). Specification of crossed random effects allows generalization of our results beyond the referees and players of the sample at hand and accounts for the correlational structure of the data (clustering due to individual differences). We used R (3.0.1; R Core Team, 2013) and the package lme4 (1.1-7; Bates, Maechler, Bolker, & Walker, 2014) for all analyses. Data and analysis scripts are available at http://osf.io/w6pi5. The α-level was .05; 95%-confidence intervals are reported in brackets.

Initial Approach

Data transformation. Skin tone was assessed by two independent raters on a 5-point Likert scale. As the rater agreement was high (ICC_{avg} = .955 [.951, .959], $F_{(1584,1585)}$ = 22.30, p < .001), we used the mean of the two ratings as predictor in our analyses. We, furthermore, reduced the number of levels of the variable *position* by collapsing "Left Fullback", "Right Fullback", and "Center Back" to "Back", "Left Midfielder", "Right Midfielder", "Center Midfielder", "Attacking Midfielder", and "Defensive Midfielder" to "Middle" and "Left Winger", "Right Winger", and "Center Forward" to "Front". As the position information for each player was given for the time of data extraction only, collapsing to more general positions was justified: Throughout their career a player may, e.g., switch from the left to the right wing; but it is less likely that they will change to a defensive position[1]. We used dummy coding for player positions and league countries with "Back" and "England" as reference categories, respectively.

Exclusion. Skin tone ratings and measures of racial prejudice in referees' home countries were the predictor variables of interest and we wanted to include players' positions as covariate in our model. We, thus, excluded all cases missing skin tone ratings, information on racial prejudice in referees' home country (*meanIAT* / *meanExp*), or information on players' positions (n = 30,014, 20.55%) assuming data was missing completely at random. Finally, we excluded all dyads with referees who in total encountered only one player (n = 513, 0.35%) because we used lme4's default Nelder-Mead optimization method for parameter estimation, which did not converge on the full dataset. The sample analyzed for our initial approach, thus, consisted of 115,501 dyads.

[1] Using the non-collapsed positions in our final approach yielded similar estimates of the effect size, $OR_{lightest,darkest}$ = 1.371 [1.101, 1.709].

Distribution. The dependent variable was given as count data (number of red cards given in a dyad). Count data are commonly modeled using the Poisson distribution. However, in the majority of dyads no red cards were given and there was no dyad in which more than two red cards were given. Count data with these characteristics may be difficult to model using the Poisson distribution. As the number of games in each dyad was known, we considered the binomial distribution as an alternative approach to the analysis of the dataset. For research question 1, we fitted a Poisson model with the number of games as offset and a binomial logistic model (s. sections on covariates and research questions for exact model specification). We compared model fits using the Akaike Information Criterion (*AIC*) and the Bayesian Information Criterion (*BIC*). The binomial distribution provided a better fit to the data than the Poisson distribution, $\Delta AIC = 3.35$, $\Delta BIC = 3.35$. Therefore, we adopted a mixed effects logistic regression approach for all subsequent analyses.

Covariates. We assumed the number of red cards a player receives throughout his career to be related to his position on the field. We reasoned that, e.g., defending players may foul strategically to stop the attacking team from scoring and thus be more likely to receive a red card than other players. Furthermore, we found an association between player position and skin tone ($\varphi_c = .122$, X^2 [24, $n = 1433$] = 63.46, $p < .001$) and an association between player positions and number of red cards received, $\varphi_c = .119$, X^2 (33, $n = 1433$) = 61.33, $p = .002$. Unsurprisingly, we also found skin tone to be related to league country, $\varphi_c = .228$, X^2 (24, $n = 1433$) = 224.30, $p < .001$. We assumed that the frequency of red cards might differ between league countries, due to different player styles or referee norms. Indeed, we found an association between the frequency of red cards and league country, $\varphi_c = .124$, X^2 (33, $n = 1433$) = 66.11, $p < .001$. Thus, in all our analyses we controlled for player position and league country[2].

Research questions. For our initial approach, we specified crossed random intercept effects for players and referees without nesting structure. Because of the extensive computation time needed for Nelder-Mead optimization (see final approach) and the time constraints of the project, we were unable to perform model comparisons with more complex random effects structures. For research questions 2a and 2b we performed moderation analyses to test if the effect of skin tone on the number of red cards varied with the racial prejudice in referees' home countries. A summary of the results for research question 1 is given in Table 1. The analyses revealed that players with darker skin tone as opposed to lighter skin tone received more red cards, *OR* = 1.087 [1.027, 1.151], $OR_{lightest,darkest} = 1.399$[3]. This effect was, however, not moderated by explicit (*OR* = 1.060 [0.816, 1.377]) or implicit racial prejudice in referees' home countries, *OR* = 0.790 [0.134, 4.674].

[2] We saw no reason to assume player weight or height to be associated with both skin tone and the frequency of red cards. Testing of these associations confirmed our assumptions. Thus, neither player weight nor height was added to the models. The given statistics replace those in a previous version of the report, which inappropriately applied to associations at the dyad level.

[3] Skin tone ratings were rescaled to range from 0 to 1 for the second round of analysis. We report here the odds ratio for lightest compared to darkest skin tone as $OR_{lightest,darkest}$ to facilitate comparison of the results.

Final Approach

For our final approach, we used the Bound Optimization by Quadratic Approximation (BOBYQA) method for parameter estimation. BOBYQA optimization was faster by a factor of four and overcame non-convergence issues we experienced in our initial approach. We were, thus, able to analyze all 116,014 complete dyads.

Random effects. We first compared random effect structures for referees. The crossed random intercept model of our initial approach served as baseline model. We compared models with random slopes for the skin tone effect, nesting referees in their home countries, and a combination of both. Nesting referees in their home countries without a random slope for the skin tone effect produced the best fit to the data ($\Delta AIC = 13.85$, $\Delta BIC = 4.20$, X^2 [1, $n = 116{,}014$] = 15.86, $p < .001$) compared to the baseline model. We then used this model as baseline for comparisons with more complex random effects for players (nesting players in clubs, league countries, or in clubs, which in turn were nested in league countries). Neither of these models improved fit to the data; all fit statistics lead to the same conclusion. Therefore we specified a non-nested random intercept term for players.

Research questions. A summary of the results for research question 1 is given in Table 2. The analyses corroborated the results of our initial approach. We found that players with darker skin tone as opposed to lighter skin tone received more red cards, $OR_{lightest,darkest} = 1.382$ [1.120, 1.705]. Again, this effect was not moderated by explicit ($OR = 1.220$ [0.445, 3.345]) or implicit racial prejudice in referees' home countries, $OR = 0.474$ [0.001, 385.189].

Additional analysis. In the project discussion phase, Morey and Wagenmakers (http://osf.io/57ku9/) reported that in their analysis the skin tone effect was contingent on the inclusion of seven outliers (0.3% of the dyads). Six of these seven cases miss data on either skin tone or player position and were not included in our analyses. Removing the remaining outlier did not affect our conclusions, $OR_{lightest,darkest} = 1.416$ [1.149, 1.745].

While drafting the manuscript, there was a discussion among analysts concerning the validity of analytical approaches that used information on players' club and league country because these variables were only available at the time of data extraction, not for each game. Molden argued that these variables are not appropriate for analyses of the longitudinal dataset and that their use may be responsible for some variance in the skin tone effect sizes reported by teams. As is apparent from our approach of collapsing player positions–also a static variable–we were among the teams that considered the longitudinal variation as noise. Reanalysis of the data showed that removing league country ($OR_{lightest,darkest} = 1.405$ [1.141, 1.731]) and additionally player position ($OR_{lightest,darkest} = 1.346$ [1.082, 1.674]) from the model produced only minor changes in effect size estimates.

Conclusion

Our analysis revealed that soccer players with darker as opposed to lighter skin tone receive more red cards throughout their career regardless of their position and league country, $OR_{lightest,darkest}$ = 1.382 [1.120, 1.705]. The effect of players' skin tone was not moderated by explicit or implicit racial prejudice in the referees' home countries. The absence of the moderation may be interpreted as evidence against a racial prejudice explanation for the skin tone effect. One of several other possible explanations is that country-level measures of racial prejudice are inadequate approximations of individual referees racial prejudice. Thus, the reasons for the increased number of red cards given to players with darker skin tone remain ambiguous.

Tables

Table 1

Results of the mixed effects logistic regression analysis of our initial approach.

Fixed effects	***OR***	***OR 95%-CI***	**β**	***SE***	*z*	*p*
Skin tone	1.087	[1.027, 1.151]	.084	.029	2.86	.004
Front position	0.603	[0.505, 0.720]	-.506	.090	-5.60	< .001
Middle position	0.612	[0.525, 0.714]	-.491	.078	-6.27	< .001
Goalkeeper position	0.834	[0.663, 1.048]	-.182	.117	-1.56	.120
France	1.310	[1.058, 1.621]	.270	.109	2.48	.013
Germany	0.950	[0.782, 1.156]	-.051	.100	-0.51	.611
Spain	1.421	[1.178, 1.713]	.351	.095	3.68	< .001
Random effects	***Var***	***SD***	***n***			
Referee (intercept)	0.103	0.322	2393			
Player (intercept)	0.330	0.575	1433			

Note. The model was fit to ***N*** = 115,501 player-referee dyads, *AIC* = 14,301.66, *BIC* = 14,398.23.

Table 2

Results of the mixed effects logistic regression analysis of our final approach.

Fixed effects	***OR***	***OR 95%-CI***	**β**	***SE***	*z*	*p*
Skin tone	1.382	[1.120, 1.705]	.324	.107	3.02	.003
Front position	0.600	[0.511, 0.705]	-.511	.082	-6.22	< .001
Middle position	0.612	[0.532, 0.703]	-.491	.071	-6.92	< .001
Goalkeeper position	0.831	[0.675, 1.023]	-.185	.106	-1.75	.081
France	1.186	[0.939, 1.498]	.171	.119	1.43	.151
Germany	0.882	[0.705, 1.103]	-.126	.114	-1.10	.271
Spain	1.151	[0.933, 1.421]	.141	.107	1.31	.189
Random effects	***Var***	***SD***	***n***			
Referee country (intercept)	0.281	0.530	152			
Referee in referee country (intercept)	0.064	0.254	2906			
Player (intercept)	0.259	0.509	1433			

Note. The model was fit to *N* = 116,014 player-referee dyads, *AIC* = 14,357.20, *BIC* = 14,463.47.

References

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*, 390–412. doi:10.1016/j.jml.2007.12.005

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2014). lme4: Linear mixed-effects models using Eigen and S4. URL: http://CRAN.R-project.org/package=lme4.

Bolker, B. M., Brooks, M. E., Clark, C. J., Geange, S. W., Poulsen, J. R., Stevens, M. H. H., & White, J.-S. S. (2009). Generalized linear mixed models: A practical guide for ecology and evolution. *Trends in Ecology & Evolution*, *24*, 127–135. doi:10.1016/j.tree.2008.10.008

Carson, R. J., & Beeson, C. M. L. (2013). Crossing Language Barriers : Using Crossed Random Effects Modelling in Psycholinguistics Research. *Tutorials in Quantitative Methods for Psychology*, *9*, 25–41.

Quené, H., & van den Bergh, H. (2008). Examples of mixed-effects modeling with crossed random effects and with binomial data. *Journal of Memory and Language*, *59*, 413–425. doi:10.1016/j.jml.2008.02.002

R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL: http://www.R-project.org/.

Zuur, A., Ieno, E. N., Walker, N., Saveliev, A. A., & Smith, G. M. (2009). *Mixed Effects Models and Extensions in Ecology with R*. New York, NY: Springer. doi:10.1007/978-0-387-87458-6