# Skin Tone Penalties:

## Bottom-up and Algorithmic Skin Tone Discrimination in Football

Donia Kamel* Guillermo Woo-Mora†

First version: August 2023
This version: December 2025‡

**Abstract**

This paper tests for direct skin–tone discrimination using professional mens football as a natural laboratory. We first develop a computer–vision algorithm that extracts a continuous, perception–consistent measure of skin tone from official player headshots. We then combine these measures with geolocated shot data and exploit quasi–random variation in whether a shot narrowly scores or narrowly misses, implementing a Regression Discontinuity and Difference–in–Discontinuities design to identify how evaluators reward identical on–field actions. Scoring a marginal goal raises post–match ratings by about 0.5 standard deviations, but Light–skinned players receive goal premia up to 0.35 standard deviations larger than Tan or Dark players. No other observable player attribute exhibits comparable heterogeneity. Replicating our design with journalist ratings reveals similarly patterned gaps, indicating that algorithmic systems likely inherit biases present in human evaluations. We then show that these rating disparities accumulate over a season and translate into lower wages and market valuations for darker-skinned players, highlighting how skin tone discrimination can generate meaningful labor market inequalities.

JEL Codes: D63, F22, J15, J71, Z20, Z22
Keywords: Race, Colorism, Discrimination, Algorithms, Football

*Paris School of Economics, donia.kamel@psemail.eu
†Paris School of Economics and EHESS, guillermo.woo-mora@psemail.eu

‡**This is a revised version of the previous draft, with ongoing work. Please do not cite nor distribute this version**. We are especially grateful to Suanna Oh, Hillel Rapoport, Thierry Verdier, and Ekaterina Zhuravskaya for their guidance. We also thank Damien Leprince for outstanding research assistance. We gratefully acknowledge comments and suggestions from Anjali Adukia, Ran Abramitzky, Sebastian Bustos, Enzo Brox, Jacob Conway, Christopher Dawes, Glen Ellison, Paolo Falco, Ricardo Hausmann, Yagan Hazard, Peter Hull, Matt Lowe, Karen Macours, Benjamin Marx, María Montoya-Aguirre, Salma Mousa, Nathan Nunn, Artur Obminski, Tim Pawlowski, Thomas Piketty, Daniel Rubenson, Evan Rose, Baptiste Roux, and Muhammed A. Yildirim, as well as seminar and conference participants at AMIE Workshop, ASSA 2025, Boston University, Central European University, CIDE, Collegio Carlo Alberto, NYUAD, PSE, PhD - Economics Virtual Seminar, University of Exeter, and the University of Tübingen. This work has benefited from the support of the Agence Nationale de la Recherche through the program Investissements dAvenir ANR-17-EURE-0001, SoBigData++, and EHESS.

# 1 Introduction

From boardrooms to playing fields, skin tone influences status, opportunity, and reward. Colorism, or discrimination based on skin tone, systematically advantages individuals with lighter skin while disadvantaging those with darker skin across various domains, including income, education, housing, marriage markets, and representation —both in the short term and intergenerationally (Hirschman, 2004; Monk, 2021; Adukia et al., 2023; Woo-Mora, 2026)

Despite its prevalence, causal evidence on skin tone discrimination remains scarce. First, skin tone is rarely measured in surveys or administrative datasets, even though it is central to how racial hierarchies are constructed and experienced (Sen and Wasow, 2016; Davenport, 2020; Adukia et al., 2025). When recorded, measurements reflect enumerators perceptions and cultural context (Dixon and Telles, 2017; Monk, 2021), leading to inconsistent coding and conflation with racial categories (Adukia et al., 2025; Woo-Mora, 2026). Second, few empirical settings allow researchers to hold productivity and ability constant when comparing groups. If discrimination is the differential treatment of otherwise identical individuals in similar circumstances (Bertrand and Duflo, 2017), most real-world settings provide insufficient observability or variation to identify direct skin tone penalties. This makes it especially difficult to separate taste-based discrimination (Becker, 1957), statistical discrimination (Phelps, 1972; Arrow, 1973), and other channels such as selection or institutional frictions (Arnold et al., 2022).

This paper provides causal evidence of direct skin tone discrimination using professional mens football as a natural laboratory. Football offers a rare combination of controlled conditions, high-frequency performance data, and real-world economic relevance (Palacios-Huerta, 2025). The sports structure has been used to study prejudice toward minorities (Alrababa'H et al., 2021), the formation of national identity (Depetris-Chauvin et al., 2020), and tax-induced mobility of high-skilled workers (Kleven et al., 2013).

We build on this literature by exploiting a feature unique to match play: shots that narrowly score and shots that narrowly miss are mechanically similar events whose outcomes differ due to micro-variation in ball trajectory. This allows us to causally identify how evaluators reward an identical action depending on the players skin tone.

Our empirical strategy proceeds in two steps. First, we develop a computer-vision algorithm that constructs a continuous, pixel-based measure of players skin tone from high-resolution headshots. The algorithm isolates facial skin regions using standard segmentation and masking procedures (Otsu, 1979; Kolkur et al., 2017) and identifies the dominant skin-color composition through $k$-means clustering, following Adukia et al. (2023). We then convert this dominant skin-color centroid into the CIELAB color space, a perceptually uniform system in which differences in color correspond closely to human visual perception.

From this representation, we compute the Individual Typology Angle (ITA)—a dermatological index that jointly captures lightness and undertone (Chardon et al., 1991; Adukia et al., 2023). ITA improves substantially on luminosity-based measures, which can misclassify individuals with warmer or more yellow undertones as lighter or darker than they appear. Because ITA is grounded in human perception, it provides an accurate, reproducible, and perception-consistent metric of skin color. Although continuous, ITA maps naturally into well-established skin-tone categories

(Light, Tan, Dark), allowing us to abstract from subjective or context-dependent racial labels while preserving interpretable distinctions relevant for studying discrimination.

Second, we use these skin-tone measures to study direct discrimination in post-match player ratings in a setting where performance is fully observed and directly comparable. Our empirical strategy exploits the quasi-random nature of shots that narrowly score or narrowly miss the goal frame. Using detailed shot-location data for over 76,000 attempts across seven major European leagues, we focus on shots taken within roughly one ball circumference of the goal posts or crossbar. Within such a narrow window, whether a shot becomes a goal is determined by micro-variation in ball trajectory—such as spin, bounce, or contact angle—that players cannot precisely control.

Comparing post-match ratings for players who barely score with those who barely miss therefore yields a local regression-discontinuity estimate of the causal goal premium, implemented with modern bias-corrected RD methods and data-driven bandwidth selection (Calonico et al., 2014). We then estimate this local effect separately for Light, Tan, and Dark players and compare them using a Difference-in-Discontinuities framework (Grembi et al., 2016; Picchetti et al., 2024). To ensure credible subgroup comparisons, we draw on recent advances in the estimation of heterogeneous treatment effects in RD designs (Calonico et al., 2025*a*).

The results reveal both a sizable overall goal premium and a clear skin tone gradient in how performance is rewarded. In the pooled sample, narrowly scoring rather than narrowly missing increases a players algorithmic rating by approximately 0.50 standard deviations, or 6.8% relative to the sample mean. However, this premium is distributed unequally across skin tone groups. Light-skinned players receive an average increase of 0.78 standard deviations following a marginal goal, compared to 0.53 for Tan players and just 0.39 for Dark players. When we replace categorical skin tone groups with our continuous ITA measure, we find that each one standard deviation increase in skin lightness corresponds to a 0.17 standard deviation larger causal rating reward for scoringa result that holds even after conditioning on fine-grained controls for shot type, player role, and team context.

These disparities are remarkably robust. They persist after adjusting for a rich set of pre-determined covariates selected via Double LASSO (Belloni et al., 2014), and across a battery of identification checks: density and balance tests around the scoring threshold, event-study plots showing parallel trends before the cutoff, narrow-window local randomization inference, placebo thresholds, and two-way clustered bootstrap standard errors. Conditioning on within-match performance metrics (e.g., total passes, tackles, and minutes played) yields nearly identical results. No other observable player attributeincluding nationality, age, height, weight, or preferred footexhibits comparably significant treatment heterogeneity once adjusted for multiple comparisons. These patterns are difficult to reconcile with purely statistical or noise-based explanations, and instead suggest that taste-based or identity-driven discrimination shapes how evaluators convert on-field performance into post-match ratings (Becker, 1957).

A natural question is what mechanism could generate such systematic discrimination in an ostensibly objective, statistics-based algorithm. If algorithmic ratings are designed to neutrally reflect on-field performance, the presence of large skin-tone gaps suggests that the algorithm is absorbing and amplifying biases embedded in the surrounding evaluative ecosystem. To test this

hypothesis, we replicate our regression discontinuity and Difference-in-Discontinuities design using post-match ratings assigned by professional journalists in three major European leagues. These expert evaluations display strikingly similar disparities: marginal goals yield significantly larger rating increases for Light-skinned players than for Tan or Dark-skinned players, with magnitudes that closely match those observed in algorithmic ratings. This alignment makes it unlikely that the bias originates purely from the algorithms internal weighting scheme and instead supports the view that the model inherits human-evaluation biases present in its training, calibration, or usage environment.

We further examine how these biased assessments align with market-based perceptions of player value. Using crowd-sourced valuations from Transfermarkt, we show that algorithmic ratings are strongly predictive of perceived market value across all skin tone groups, with nearly identical slopes. This suggests that depressed ratings for Dark-skinned players translate proportionally into lower market valuationsreflecting not only algorithmic bias but also its uptake and reinforcement by fans and analysts. Importantly, no other observable player characteristic (e.g., nationality, age, height, weight, preferred foot) shows comparable rating disparities, further underscoring the specificity of skin tone as a salient axis of differential recognition.

Finally, we document that these rating disparities have tangible economic consequences. Using player-season data on wages and market valuations, we show that Light-skinned players benefit at two key stages: first, they receive higher season-average ratings than Dark-skinned players for identical quasi-random on-field success; second, they convert those ratings into significantly greater wage growth and market value appreciation in the following season. These compounding advantages establish a direct transmission channel from biased performance evaluations to unequal labor market outcomes, indicating that colorism in football influences not only perception, but also pay, valuation, and career mobility.

We are currently developing a series of online and field experiments to more directly examine the mechanisms behind the algorithmic discrimination we document. In the online component, participants will evaluate player performances based on fixed performance statistics, while headshots will be randomly assigned to depict light-, tan-, or dark-skinned players. Participants will also be randomly assigned to receive additional context: either algorithmic ratings, journalist ratings, or no external benchmark. This design allows us to test (i) whether baseline color biases exist among fans, (ii) whether these biases are amplified or mitigated by exposure to existing rating systems, and (iii) whether providing objective, statistics-based information can reduce discriminatory evaluations. These experimental extensions aim to isolate the sources of bias and inform the design of interventions to mitigate skin tone discrimination in performance assessment.

This paper contributes first to the empirical literature on discrimination. A long tradition in economics highlights how differential treatment can arise from taste-based or statistical motives (Arrow, 1973; Becker, 1957; Phelps, 1972). A large body of evidence documents persistent disparities across racial and ethnic groups in labor, credit, education, and justice markets (Arceo-Gomez and Campos-Vazquez, 2014; Arnold et al., 2022; Bertrand and Mullainathan, 2004; Bohren et al., 2025; Darity, 2022; Derenoncourt et al., 2023; Kline et al., 2022; Lang and Spitzer, 2020). Most empirical strategies rely on observational comparisons or on audit and correspondence experiments that necessarily bundle multiple identity dimensions or focus on coarse categorical classifications

(Bertrand and Duflo, 2017; Guryan and Charles, 2013). By contrast, our setting isolates a single, visually salient attribute—skin tone—under near-experimental conditions: narrowly scored and narrowly missed shots are quasi-random realizations of the same underlying action, enabling us to measure differential evaluation while holding task execution and productivity fixed.

A second contribution is to the growing literature on colorism and the economic salience of phenotypic variation. A large interdisciplinary body of work shows that skin tone predicts wages, wealth, mobility, education, and marriage outcomes even within racial groups (e.g. Adukia et al., 2023, 2025; Abramitzky et al., 2023; Dixon and Telles, 2017; Goldsmith et al., 2006, 2007; Hunter, 2007; Monk, 2021; Rosenblum et al., 2016). Recent work conceptualizes race as a socially constructed bundle of attributes whose salience varies across institutional settings (Davenport, 2020; Rose, 2023; Sen and Wasow, 2016). Yet clean causal evidence on color-based discrimination remains scarce because individual performance is difficult to observe at high frequency and even harder to hold constant across workers. We contribute by developing a pixel-based algorithmic measure of skin tone that removes interviewer subjectivity and by estimating causal color gradients in the rewards for identical on-field performance. These results complement correlational evidence on skin tone penalties in Latin America (Campos-Vázquez, 2020; Woo-Mora, 2026) and extend experimental demonstrations of skin-tone-based discrimination in market settings (Doleac and Stein, 2013) to a global, high-stakes labor market.

Third, we contribute to research using football as a natural empirical laboratory for economics. Football provides quasi-experimental variation in performance, rich micro-data, and standardized evaluation environments (Alrababa'H et al., 2021; Depetris-Chauvin et al., 2020; Kleven et al., 2013; Palacios-Huerta, 2003, 2025). Prior work documents discrimination in refereeing (Gallo et al., 2013; Reilly and Witt, 2011), labor markets (Caselli et al., 2023; Deschamps and De Sousa, 2021; Principe and van Ours, 2022), disciplinary sanctions (Faltings et al., 2023), and public attention (Magistro and Wack, 2023). Two recent papers speak directly to identity in evaluator responses: Alrababah et al. (2024) find little evidence that minorities are punished more harshly for underperformance in the English Premier League, while Colombe et al. (2025) document overcorrection in decisions affecting Black and darker-skinned players under heightened media scrutiny. We differ by exploiting quasi-random variation in narrowly scored versus narrowly missed shots, yielding within-match causal identification of how evaluators translate identical performance into ratings across skin tone groups.

Finally, we contribute to a growing literature on algorithmic fairness and the distributional consequences of predictive systems. Recent research shows that algorithms can reproduce or amplify group disparities when trained on proxies that encode historical inequities—for example, in online ad delivery and health-risk prediction (Lambrecht and Tucker, 2019; Obermeyer et al., 2019). Our setting offers a rare opportunity to study these issues under near-experimental conditions: individual output is observed directly, standardized, and comparable across workers. We show that algorithmic evaluations in football display a clear, monotonic gradient by skin tone, and that this gradient closely mirrors the one emerging from journalist-based ratings. This alignment is consistent with algorithms inheriting and formalizing human-evaluation biases present in the surrounding informational environment, providing quasi-experimental evidence of algorithmic colorism in a real labor market.

## 2 Setting

Professional football offers a rare environment for studying discrimination in performance evaluation. Individual contributions are recorded at high frequency, assessed using standardized metrics, and widely disseminated to clubs, media, and fans (Palacios-Huerta, 2003; Kleven et al., 2013; Depetris-Chauvin et al., 2020; Alrababa'H et al., 2021; Palacios-Huerta, 2025). Each match generates a detailed sequence of on-ball actions, and multiple institutions translate those actions into post-match ratings—numerical scores that, while not financial outcomes themselves, shape perceptions of quality and may influence scouting, contract negotiations, and market values.

We study two primary forms of evaluation. Algorithmic ratings (from FotMob/Opta) use fixed procedures to summarize event-level statistics into a single performance score, aiming to minimize subjective bias. Journalist ratings (from Sky Sports, La Gazzetta dello Sport, and LÉquipe) reflect expert assessments formed through full-match viewing. These systems coexist across Europes top leagues and evaluate identical performances using distinct principles, allowing us to test whether disparities in recognition by skin tone appear in both algorithmic and human judgments.

The setting also reflects well-documented racial disparities. After the 2021 Euro final, Bukayo Saka, Marcus Rashford, and Jadon Sancho—three Black English players—were subjected to a wave of racist abuse after missing penalty kicks in the deciding shootout. By contrast, Harry Kane, who is white, missed a critical penalty in Englands 2022 World Cup loss but faced no comparable backlash. These incidents underscore a broader pattern: performance is not always evaluated equally. Football provides an ideal case to test whether such disparities extend to formal, match-level evaluations and whether they translate into downstream labor market outcomes.

## 3 Data

We exploit multiple data sources to study the prevalence of colorism in men's football.

### 3.1 Algorithm-based Data

We use data from FotMob, an online football analytics platform that provides granular, algorithmically generated ratings for players and matches across major leagues. Player ratings on FotMob are not crowdsourced or based on user input; instead, they are computed by a proprietary algorithm developed by Opta Sports, using a wide set of individual statistics.[1] Despite their non-human origin, FotMob ratings are strongly correlated with fan-based assessments, suggesting that algorithmic systems may absorb or reflect broader audience preferences.[2]

We extract the data using the `worldfootballR` package (Zivkovic, 2023), which enables access to both player-level ratings and detailed event-level data. Our analysis focuses on a subset of players who recorded shots on goal. For each shot, we observe its geolocation relative to the goal frame, the outcome, and the shooters post-match rating. This level of granularity is central to our identification strategy, as it allows us to compare outcomes for players who narrowly succeeded versus narrowly failed.

[1] FotMob confirmed the algorithmic basis of these ratings in a tweet on February 17, 2018. See Figure A.1.

[2] See binned scatterplot in Figure **??**.

Geolocated shot data is available for seven top-tier European leagues: the English Premier League, Spanish La Liga, German Bundesliga, Italian Serie A, French Ligue 1, Portuguese Primeira Liga, and Dutch Eredivisie. The data spans three complete seasons: 2020/21, 2021/22, and 2022/23.[3] The final dataset includes over 76,000 shots on goal, and is enriched with covariates at multiple levels: event-level (e.g., shot coordinates, outcome, minute of the match, etc), player-level (e.g., position, team, minutes played, etc) and match-level (e.g., date, stadium, home and away teams, final score, etc). Figure 1 provides a visual summary of these data. Each point represents a shot on goal, plotted relative to the football goal (7.32 x 2.44 meters). The shape of the point indicates the outcome (goal, miss, hit post), while the color gradient reflects the shooters post-match rating.

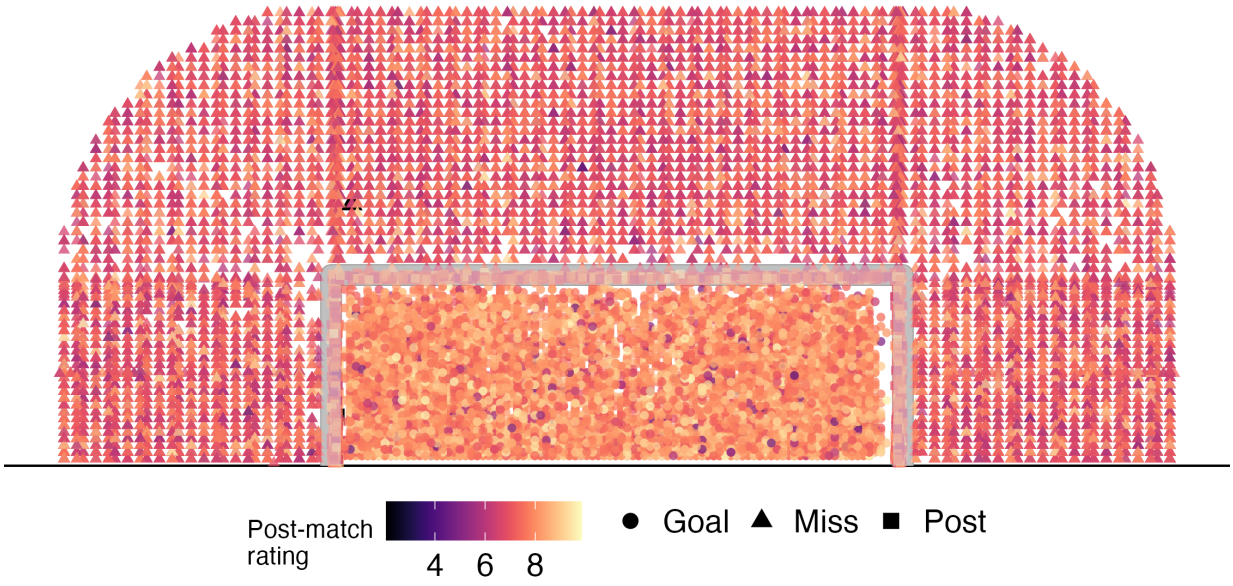



Figure 1: Shots on Goal: Outcome and Player's Post-Match Rating

*Notes*: Each point represents a shot on goal during a match. The grey area indicates the goal frame. Circles denote shots that resulted in goals, triangles indicate missed shots, and rectangles represent shots that hit the post. The color of each point reflects the players post-match rating, with brighter colors indicating higher ratings. The regulation goal frame measures 7.32 meters in width and 2.44 meters in height, with a post width of 12 cm. The standard match ball (Size 5) has a circumference of 68—70 cm.

### 3.2 Journalist-based Data

To complement our algorithmic player post-match ratings, we collect journalist-assigned post-match ratings. We compile data from three major leagues: the English Premier League, Italian Serie A, and the French Ligue 1. These ratings are published shortly after each match and are assigned by sports journalists affiliated with major football media outlets (Sky Sports, 2023; Fantacalcio.it, 2023; L'Équipe, 2023). They serve as one of the most prominent public-facing forms of expert player evaluation. We match these journalist ratings to players in our main

[3]Figure A.2b shows a screenshot of the FotMob website with shot-level information.

events-within-match dataset (such as near-goal shots) using a combination of match identifiers (date, round, season, and teams) and fuzzy string matching on player names using the Jaro—Winkler distance metric, with a maximum allowable distance of 0.10 (corresponding to 90% similarity).

**English Premier League** We obtain ratings from Sky Sports, a leading UK broadcaster and sports news outlet. Ratings are assigned on a 1-10 scale by sport journalists and reflect a players performance across tactical, physical, and technical dimensions. Our dataset covers all matches from the 2020-2021 through 2022-2023 seasons.[4] Using our matching procedure, we successfully assign journalist ratings to over 90% of players involved in relevant match events.

**Italian Serie A** We collect ratings from La Gazzetta dello Sport, an Italian newspaper dedicated to coverage of various sports. We focus exclusively on the journalist-generated ratings, excluding those derived from fan votes or automated statistical systems. These ratings also follow a 1-10 scale and span the 2020-2021 through 2023-2024 seasons. Our matching yields coverage for over 90% of players involved in relevant match events.

**French Ligue 1** We collect ratings for the 2020-2021 and 2021-2022 seasons from the French sports newspaper L'Équipe, one of the most influential sources for French football coverage. Ratings were extracted using optical character recognition (OCR) on digitized match images, which display player ratings in standard pitch diagram format.[5] Figure A.3 shows an example of one of these images displaying post-match journalist ratings in Ligue 1. Despite limited formatting, we achieve approximately 75% coverage of relevant players.

Figure A.4 shows the distribution of normalized scores by league. Furthermore, we are currently expanding our journalist-based dataset to include additional seasons and leagues. That is to include the last season of the Premier League (2023-2024), seasons 2022-2023 and 2023-2024 for Ligue 1, and to collect data for the other leagues: Bundesliga of Germany, Primeira Liga of Portugal, and Eredivise of the Netherlands.

### 3.3 Skin tone detection algorithm

Colorism, racial discrimination concerned with actual skin-color, has consequences on various dimensions such as education, living conditions, economics outcomes, and a wide range of social outcomes (Espino and Franz, 2002; Hunter, 2005; Rondilla and Spickard, 2007; Cole et al., 2014). However, empirical work on the effects of colorism remains limited due to the inherent data and measurement problems. This work abstracts from such limitations. From Sofifa.com and FotMob, we obtain headshots of players, which feed into our algorithm, enabling the construction of our factual measure of skin tone. This is due to the fact that computers and fine-tuned algorithms can directly observe the color of each pixel unlike the categorization of race which can vary depending

[4] Therefore only missing 2023-2024 season from our main dataset

[5] In the images, players are listed only by surname and initial, we refine our fuzzy-matching strategy by conditioning on the first letter of the players first name from the main dataset. This helps mitigate misclassification when players have the same or similar last names.

on the cultural context and the enumerator or observer. We proceed by explaining our algorithm. Figure 2 illustrates descriptively the steps we take and the outcome.

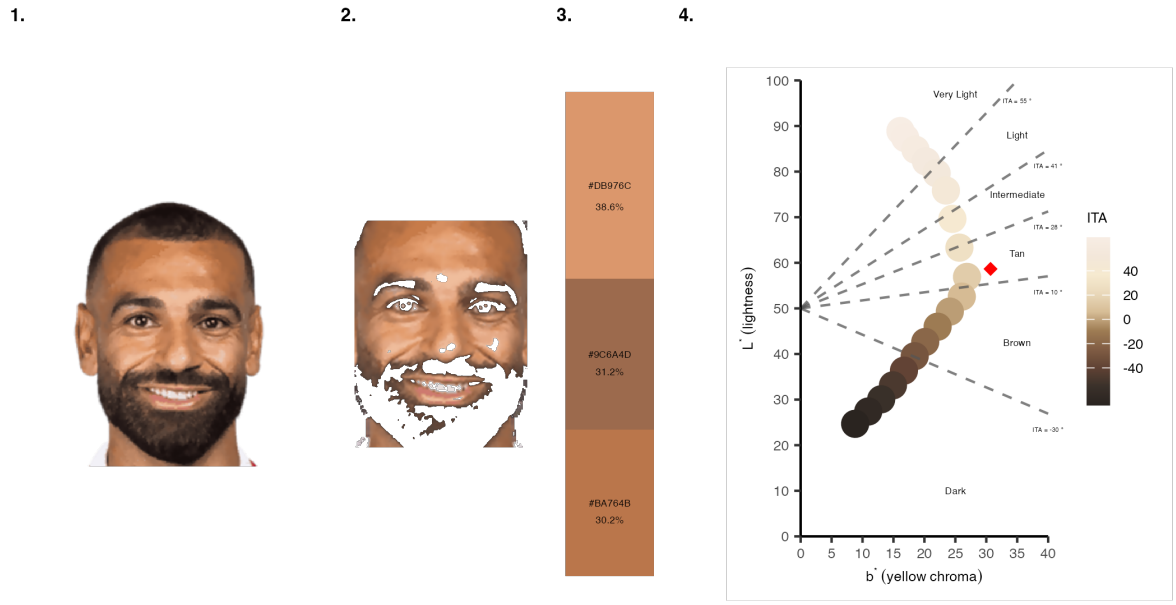



Figure 2: Illustration of the steps of the skin tone detection algorithm.

*Notes*: This figure illustrates the steps of the skin tone classification algorithm applied to Mohamed Salah. The method follows Otsu (1979) and Kolkur et al. (2017) to segment and mask skin regions before converting the image to the CIELab color space (Adukia et al., 2023). The Individual Topology Angle (ITA) is then computed using the $L^*$ (lightness) and $b^*$ (yellow-blue spectrum) components, allowing for an objective and reproducible classification of skin tone.

The high-quality headshots from our data are RGB images stored as 2-by-2-by-3 data arrays that define red, green, and blue color components for each individual pixel. Our approach starts with image segmentation. First, we separate each image into two classes, foreground and background, following (Otsu, 1979), based on the images' pixel values for grayscale intensity. Using this information on foreground and background pixels, we extract a *mask* for the targeted facial region. We then use (Kolkur et al., 2017)'s algorithm for skin detection that considers combinational ranges, such as texture, and edge features, to have a more fine-grained segmentation result and greater accuracy in recognizing the skin area.

Once we isolate the facial skin region, we employ Adukia et al. (2023) and apply $k$-means clustering (with $k = 3$) to identify the dominant color groupings among the remaining skin pixels. This unsupervised machine learning algorithm groups the pixel colors and areas into distinct clusters in the RGB color space. We exclude the smallest cluster, which typically captures non-skin areas such as shadows, highlights, or glare. Among the remaining two clusters, we compute a weighted average of their RGB centroids, with weights corresponding to each cluster's share of of the total skin-region pixels. This results in a single representative RGB value that captures the dominant skin color for each image. We then convert this RGB value into the CIELAB color space, a perceptually uniform space that represents colors using three components: $L^*$ (lightness), $a^*$ (green-red axis), and $b^*$ (blue-yellow axis). This transformation allows us to analyze skin color in a way that aligns more closely with human perception.

To quantify skin tone, we compute the Individual Typology Angle (ITA), a metric widely used in dermatology and cosmetic science (Chardon et al., 1991). The ITA captures the perceived skin tone based on a ratio of lightness ($L^*$) to yellowness ($b^*$)in the CIELAB color space. Specifically, it is defined as the angle formed by the vector between a reference lightness value (50) and the skin's blue—yellow chromaticity component ($b^*$):

$$\text{ITA} = \arctan\left(\frac{L^* - 50}{b^*}\right) \cdot \frac{180}{\pi}$$

The use of this angular measure provides a more nuanced representation of skin tone by accounting for both the overall lightness ($L^*$) and the undertone or warmth ($b^*$). Higher ITA values correspond to lighter, less yellow skin tones, while lower ITA values correspond to darker or more yellow/olive skin tones. We classify individuals into ITA-based skin tone groups: Very Light: ITA $> 55°$; Light: $41° <$ ITA $\leq 55°$; Intermediate: $28° <$ ITA $\leq 41°$; Tan: $10° <$ ITA $\leq 28°$; Brown: $-30° <$ ITA $\leq 10°$ and Dark: ITA $\leq -30°$.

This method enables us to quantify skin tone variation in a reproducible and objective manner, avoiding subjective racial classification or reliance on self-reported categories. Compared to simpler measures that rely only on the lightness component ($L^*$), ITA provides a richer, more perceptually accurate representation of skin tone.

Using ITA offers several key advantages for our setting. First, it better reflects how skin tone is perceived by human observers by combining lightness and undertone. Second, it avoids the pitfalls of using a single-dimension measure like luminosity ($L^*$), which can misclassify reddish or yellowish skin as lighter or darker than it appears. Third, ITA values can be grouped into consistent categories, facilitating comparisons across populations while reducing arbitrariness. Finally, because its based on pixel-level data rather than social categories, it allows us to isolate the role of skin tone itself in shaping outcomes—such as ratings by fans, journalists, or algorithms—without conflating it with broader notions of race or ethnicity.

As shown in Figure 2, Mohamed Salah's ITA value places him in the "Tan" category. Notably, his lightness value ($L^*$) is around 55-comparable to players categorized as "Light" or "Intermediate". However, because his skin tone also includes a higher level of yellow chromaticity ($b^*$), the ITA correctly classifies him as Tan. This underscores the importance of incorporating both lightness and undertone into skin tone measurement: relying on $L^*$ alone would risk misclassifying individuals with warmer skin tones. Figure A.5 demonstrates the classification of skin tone for Harry Kane (Light), Sadio Mané (Dark), Mohamed Salah (Tan), and Lionel Messi (Light). For the analysis, we group players into three broader categories: *Light*, *Tan*, and *Dark*, which correspond to ITA thresholds of: *Light* (ITA $> 41°$), combining Very Light and Light; *Tan* ($10° <$ ITA $\leq 41°$), combining Intermediate and Tan; and *Dark* (ITA $\leq 10°$), combining Brown and Dark.

## 4 Empirical Framework

To evaluate direct skin tone discrimination, we exploit a setting that allow us to look for differential treatment between skin tone groups with otherwise identical characteristic in a similar circumstance.

### 4.1 Causal Framework from Match-Level Ratings

**Regression Discontinuity Design.** We exploit the quasi-random nature of shots that narrowly score or narrowly miss the goal frame to estimate the local causal effect of scoring on post-match player ratings. The identifying assumption is that around this scoring threshold, whether a shot becomes a goal is effectively determined by micro-variations—such as ball spin, bounce, contact angle, or wind—that players cannot precisely control.

Formally, we estimate the following RDD specification:

$$y_{pm(ls)} = \alpha_l + \alpha_s + \delta z_{ipm} + \beta \cdot \text{Goal}_{ipm} + \mathbf{X}'_{ipm}\gamma + \varepsilon_{ipm(ls)}, \tag{1}$$

where $y_{pm}$ is the post-match rating of player $p$ in match $m$. Unless otherwise noted, all specifications use algorithmic ratings provided by the FotMob platform. In robustness checks, we replicate our analysis using journalist-assigned player ratings from reputable sports media outlets to assess whether similar disparities emerge under human evaluation—thus providing a benchmark comparison to the algorithmic results.

The term $\text{Goal}_{ipm}$ is an indicator equal to one if shot $i$ resulted in a goal. The running variable $z_{ipm}$ measures the Euclidean distance (in centimeters) between the shot and the goal frame. All specifications include league ($\alpha_l$) and season ($\alpha_s$) fixed effects.

The vector $\mathbf{X}_{ipm}$ contains pre-determined, shot-level covariates selected using the Double LASSO method of Belloni et al. (2014). These include the shot's coordinates, match minute, shot situation (e.g., open play, penalty, set piece), and shot type (e.g., left foot, right foot, header). Player-level predictors include starter status, captaincy, and shirt number; match-level covariates include home-team status and stadium attendance.

The coefficient $\beta$ captures the goal premium—the causal increase in a players post-match rating resulting from scoring a goal, conditional on the shot occurring near the scoring threshold. It reflects the additional recognition a player receives, on average, for converting a marginal shot relative to narrowly missing, holding constant all observable shot- and player-level characteristics. Unless noted otherwise, we report bias-corrected RDD estimates throughout.

Identification relies on the continuity assumption in RDD: potential post-match ratings must vary smoothly with the running variable in the absence of scoring. Under this assumption, any discontinuous jump at the scoring threshold can be interpreted as the causal effect of scoring rather than differences in underlying ability, shot quality, or match context.

We begin with the universe of over 76,000 shots across seven major European leagues, excluding only shots in which the goalkeeper intervened. In this sample, 21.9% of shots result in goals, 74.2% are off-target, and 3.9% strike the frame. We use this full dataset to estimate the optimal bandwidth. Applying the MSE-optimal selector of Calonico et al. (2014), we obtain a bandwidth of roughly 65 centimeters on each side of the scoring threshold and a bias-correction bandwidth of 114 centimeters.

Crucially, this bandwidth is nearly identical to the circumference of a regulation football (about 70 cm). This fact is central to our identification strategy: within such a narrow window, small differences in ball placement—on the order of a single ball circumference—are largely attributable to idiosyncratic, effectively random physical variation rather than deliberate aiming. Shots that

barely score or barely miss therefore have nearly identical mechanics, creating a natural experiment for isolating the causal effect of scoring on ratings.

Our estimation sample consists of all shots taken within 114 cm of the goal frame, covering both the main and bias-correction bandwidths. Because players may take multiple close-range shots within the same match—each contributing to their overall post-match rating—we weight observations by the inverse of the number of within-bandwidth shots taken by a given player-match pair.

To assess the plausibility of the continuity assumption, we examine the density of the running variable around the cutoff. Figure B.1 presents two density plots. Panel (a), which includes shots that hit the post, shows a spike just below the threshold—reflecting the mechanical nature of post-striking shots rather than behavioral manipulation. Panel (b), which excludes post-hitting shots, displays a smooth and continuous density around the cutoff, with no indication of sorting or manipulation. Importantly, the observations within the MSE-optimal bandwidth—those driving identification—are dominated by random micro-variation. Thus, the presence or absence of post shots does not undermine identification, and the continuity assumption holds in the neighborhood of the scoring threshold.

**Difference-in-Discontinuities (DiDC).** To estimate whether the causal reward for scoring a goal varies by skin tone, we adopt a Difference-in-Discontinuities (DiDC) strategy.

Originally proposed by Grembi et al. (2016), this framework combines discontinuity-based identification (as in RDD) with comparisons across groups or over time, similar in spirit to a Difference-in-Differences design. Picchetti et al. (2024) provide a formal econometric foundation for this approach, establishing conditions under which the difference in RDD coefficients across groups identifies heterogeneity in treatment effects.

In our case, we do not rely on time variation. Instead, we estimate separate RDDs by mutually exclusive skin tone groups and then compare the size of the treatment effect (i.e., the goal premium) across these groups.

Specifically, we estimate Equation 1 for each group $ST \in \{\text{Light}, \text{Tan}, \text{Dark}\}$:

$$y^{ST}_{pm(ls)} = \alpha_l + \alpha_s + \delta z_{ipm} + \beta^{ST} \cdot \text{Goal}_{ipm} + \mathbf{X}'_{ipm}\gamma + \varepsilon^{ST}_{ipm(ls)}$$

where the variables are exactly the same as those in Equation 1.

From this, we obtain group-specific estimates $\hat{\beta}^{ST}$, which reflect the causal effect of scoring a goal on ratings for each skin tone group. We then compute pairwise differences in these effects:

$$\Delta = \hat{\beta}^{ST_1} - \hat{\beta}^{ST_2} \tag{2}$$

where $\Delta$ captures the differential causal reward for scoring between groups $ST_1$ and $ST_2$. We report all three pairwise comparisons: Dark vs. Light, Brown vs. Light, and Dark vs. Brown.

Conceptually, Equation 2 resembles a Mincer-type earnings regression, where $\Delta$ plays a role analogous to the "returns" or "discrimination" coefficient in an Oaxaca-Blinder decomposition (Jann, 2008). Crucially, given the quasi-random assignment induced by our RDD design, these differences are less likely to be driven by unobserved confounders and instead reflect heterogeneity in the causal recognition (or bias) associated with identical performance across groups.

For estimation, we implement Calonico et al. (2025*a*), which develops robust methods for estimating and conducting inference on treatment effect heterogeneity in regression discontinuity designs. Their framework allows for fully nonparametric estimation of subgroup-specific effects while preserving valid inference under standard continuity-based RD assumptions.

For inference, we use a set of complementary approaches to account for the structure of our DiDC setting. Our baseline estimates rely on the robust, bias-corrected standard errors proposed by Calonico et al. (2025*a*), which provide valid inference for group-specific treatment effects in regression discontinuity designs. However, their current implementation does not allow for two-way clustering, which is crucial in our context: goal outcomes vary at the match level (requiring match clustering), while skin tone is fixed at the player level (requiring player clustering). To address this, we turn to bootstrap-based methods to conduct inference on the difference-in-RD coefficients $\Delta$ across groups.

Yet standard bootstrap procedures are known to perform poorly in RD settings with MSE-optimal bandwidths, due to persistent finite-sample bias (Bartalotti et al., 2017). To overcome this, we rely on the insight from Picchetti et al. (2024) that under a parallel-trends-type assumption—specifically, that untreated potential outcomes (i.e., post-match ratings for non-scoring players) evolve similarly across groups—the bias in the DiDC estimator asymptotically cancels. In the following section, we show evidence that this assumption likely holds in our setting. This makes two-way clustered bootstrap inference on $\Delta$ valid. We implement a two-way Bayesian bootstrap (Rubin, 1981) that respects both clustering dimensions and retains all close-shot observations within the optimal bandwidth, which is essential for credible estimation of heterogeneous causal effects near the scoring threshold.

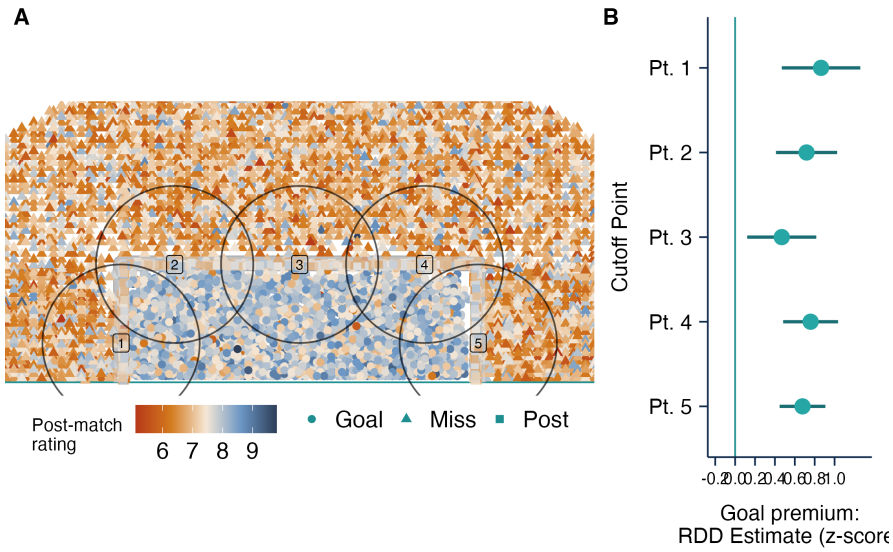
# 5 Results

We now present our main empirical findings on whether players of different skin tones receive differential recognition for equivalent on-field performance. To isolate causal effects, we implement a Differences-in-Discontinuities (DiDC) design that compares the size of the post-match rating premium for narrowly scoring versus narrowly missing a goal, across mutually exclusive skin tone groups.

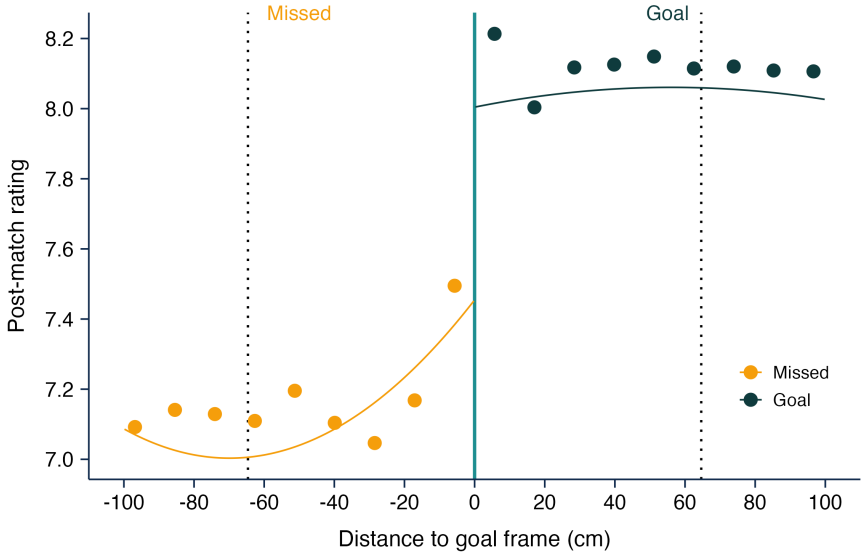
## 5.1 Regression Discontinuity

We begin by validating the first stage of our identification strategy: do narrowly scored goals result in higher post-match ratings than narrowly missed attempts? Before introducing heterogeneity by skin tone, we estimate the average causal effect of scoring on player ratings using a pooled RDD.

Figure 3a illustrates our empirical strategy. Panel A visualizes shot locations relative to the goal frame, using color to represent the shooters post-match rating by FotMob. Near each of five spatial cutoffs (e.g., posts, crossbar), we implement local two-dimensional RDDs comparing outcomes for shots that barely scored versus those that barely missed, following Cattaneo et al. (2025). Panel B shows that in all five locations, scoring leads to statistically significant increases in player ratings. This supports the central identifying assumption that near-threshold scoring




(a) Two-dimensional RDD Estimates




(b) One-dimensional RDD Plot

Figure 3: Goal Scoring Premium

*Notes:* Panel (a) shows the spatial distribution of shots relative to the goal frame, with color indicating the shooter's post-match rating (from algorithmic sources). Each vertical segment denotes a different scoring threshold (e.g., post or crossbar). Using local two-dimensional RDDs using Cattaneo et al. (2025), we compare outcomes for shots just inside versus just outside each threshold. All five cutoffs show a statistically significant increase in ratings for goals. Panel (b) presents the pooled, one-dimensional RDD using Calonico et al. (2014) with an Euclidean distance to the goal frame as the running variable. The solid vertical line denotes the scoring cutoff. The figure shows a discontinuous jump of approximately 0.5 points (or 0.55 standard deviations) in post-match ratings for goals relative to near misses, using a 65 cm optimal bandwidth. All ratings are algorithm-based.

outcomes are quasi-random and suitable for causal inference.

Figure 3b presents the pooled, unidimensional RDD. The outcome variable is the players algorithmic post-match rating, and the running variable is the shots Euclidean distance to the goal frame. The vertical line at zero marks the scoring threshold. The plot shows a clear discontinuity at the cutoff: scoring a goal increases the average rating by approximately 0.5 points, or half standard deviations—a 6.8% increase relative to the mean. The estimate is significant at the 1% level and robust to a range of alternative specifications.

Appendix Table B.1 summarizes the RDD estimates across model specifications. The estimated goal premium remains stable when controlling for a rich set of shot- and player-level covariates (clean controls and Double LASSO), as well as when including match fixed effects. The estimate is also robust to different kernels, polynomials, and bandwidths. Across all specifications, the goal premium ranges from 0.45 to 0.58 standard deviations.

We interpret this discontinuity as a goal premium: a causal reward in player ratings for scoring, conditional on comparable shot characteristics. Identification is driven by a narrow bandwidth of 65 cm on each side of the goal frame, closely aligned with the circumference of a regulation football. This ensures that variation in outcomes is due to random physical noise—ball spin, bounce, or deflection—rather than deliberate player intent.

Importantly, the observed increase in ratings reflects a reward for scoring, not a penalty for missing. Appendix Figure B.2 displays the distribution of post-match ratings for three groups: (i) the full universe of player ratings, including all players regardless of whether they took a shot; (ii) players who missed a shot on goal; and (iii) players who scored a goal. The distribution for missed shots closely resembles the full rating distribution, suggesting that narrowly missing a goal does not reduce a players rating. In contrast, the distribution for goal-scoring players is clearly shifted to the right, with a higher mean. This pattern supports the interpretation that the post-match rating premium reflects positive recognition for scoring, rather than a penalty for missing.

### 5.2 Difference-in-Discontinuities by Skin Tone

Do goal premia differ by skin tone? To examine this, we follow Calonico et al. (2025*a*) and estimate the RDD specification in Equation 1 separately for each skin tone group, defined as $\text{ST} \in \{\text{Dark}, \text{Tan}, \text{Light}\}$. We then compare the estimated effects across groups using the Difference-in-Discontinuities framework described in Equation 2.

For each group, we estimate the group-specific goal premium $\hat{\beta}^{ST}$ using the MSE-optimal bandwidth of 65 centimeters and an Epanechnikov kernel. All specifications include Double LASSO-selected covariates and are weighted by the inverse of the number of shots taken by each player-match pair within the bandwidth.

To formally test for discrimination, we conduct one-sided hypothesis tests of the form $H_0 : \hat{\beta}^{ST_{\text{lighter}}} - \hat{\beta}^{ST_{\text{darker}}} \leq 0$ against $H_1 : \hat{\beta}^{ST_{\text{lighter}}} - \hat{\beta}^{ST_{\text{darker}}} > 0$ for each pairwise comparison. These tests assess whether lighter-skinned players receive greater recognition—measured as a higher rating premium—for scoring an identical goal. Inference is based on linear combinations of the group-specific estimates, with heteroskedasticity-robust standard errors clustered at the player

level.

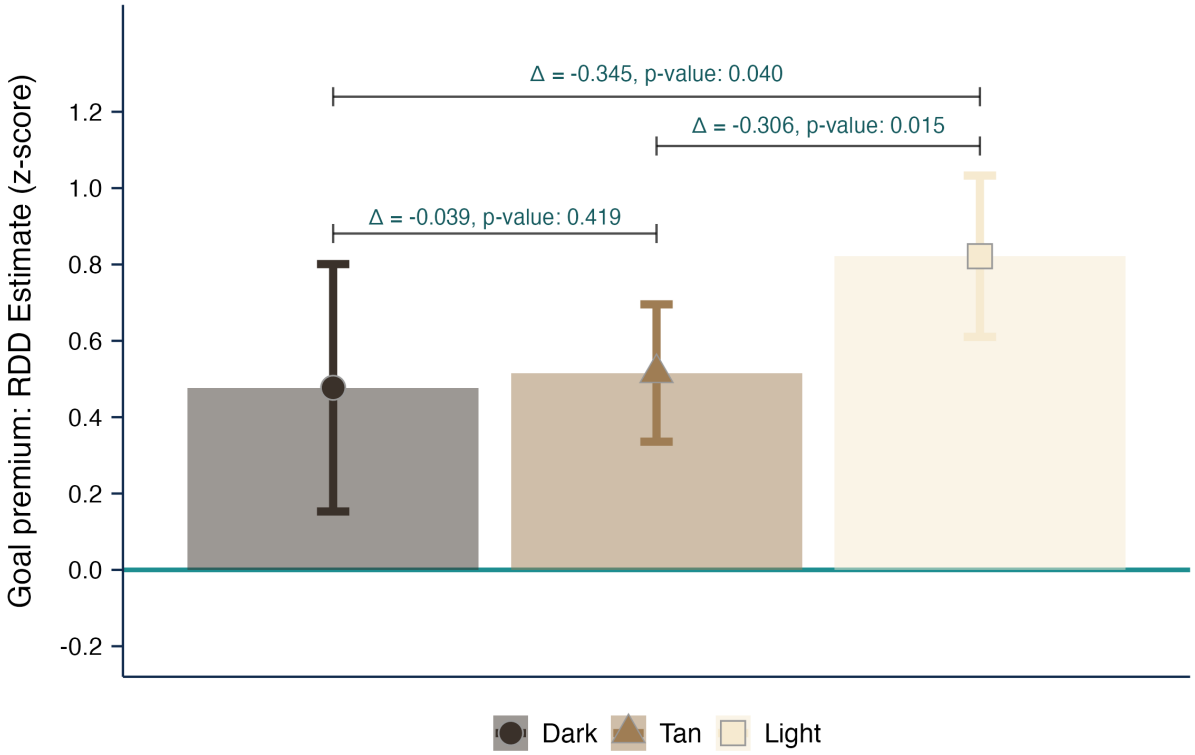



Figure 4: Difference-in-Discontinuities: Goal Premium by Skin Tone

*Notes:* This figure displays heterogeneous bias-corrected RDD estimates of the goal premium, the causal effect of scoring a goal on algorithmic post-match player ratings, for each skin tone category using `rdhte` package Calonico et al. (2025*b*). The y-axis shows the goal premium RDD estimates in z-scores. Point estimates correspond to $\hat{\beta}$ from Equation 1, with 95% confidence intervals (robust to heteroskedasticity and clustered at the player level) shown as error bars. Brackets denote pairwise differences $\Delta = \hat{\beta}^{ST_{\text{lighter}}} - \hat{\beta}^{ST_{\text{darker}}}$, with $p$-values from one-sided tests of $H_0 : \Delta \leq 0$ against $H_1 : \Delta > 0$. All estimates use the MSE-optimal bandwidth of $h = 114$ cm and an Epanechnikov kernel. The specification includes Double LASSO-selected covariates and are weighted by the inverse number of within-bandwidth shots per player-match.

Figure 4 presents the main results. Light-skinned players receive a goal premium of 0.79 standard deviations; Tan-skinned players receive 0.53 standard deviations; and Dark-skinned players receive 0.44 standard deviations. The one-sided tests reveal significant disparities: Light-skinned players receive 0.343 standard deviations more recognition than Dark-skinned players ($p = 0.029$), and 0.246 standard deviations more than Tan-skinned players ($p = 0.030$). While Tan-skinned players also receive more recognition than Dark-skinned players ($\Delta = 0.097$), the difference is not statistically significant ($p = 0.286$).

A natural question is whether the goal premium varies smoothly with skin tone rather than only across discrete groups. To examine this, we exploit the continuous ITA skin tone measure and estimate heterogeneous treatment effects using the `rdhte` procedure, with the standardized ITA score included as the heterogeneity covariate. This approach allows us to trace how the RDD treatment effect evolves across the full distribution of skin tone values. Appendix Figure B.4 plots the estimated marginal effect of scoring as a linear function of the ITA skin tone z-score, with accompanying 95% confidence intervals. The results show a clear monotonic gradient: a one–standard deviation increase toward darker skin tone leads to a 0.12 standard deviation lower

goal premium (two-sided $p = 0.072$).

To illustrate the magnitude, moving from the Very Light ITA category to the Dark category represents roughly a four–standard deviation shift. Over this range, the estimated goal premium rises from about 0.25 standard deviations (for Dark players)—a value not statistically different from zero—to nearly 0.75 standard deviations (for Very Light players). This represents a difference of roughly half a standard deviation in recognition for the identical performance event.

This continuous dose–response pattern complements our categorical estimates and reinforces the conclusion that algorithmic evaluations exhibit a persistent and monotonic colorism gradient.

### 5.2.1 Robustness

**Selection into the Shooters Sample.** A potential concern is that players who take marginal shots may differ systematically by skin tone, biasing our estimates. To assess this, we compare the skin tone distribution of our shooters sample to all players who received post-match ratings during the same period.

Appendix Figure B.5 shows substantial overlap between the two distributions. A Kolmogorov-Smirnov test fails to reject equality ($D = 0.020$, $p = 0.349$), indicating no significant compositional differences. This supports the validity of our RDD and suggests that observed disparities reflect differential treatment, not differential selection.

**Validation: Binned Event-Study** To complement our main RDD analysis and provide visual evidence for the validity of the DiDC design and the presence of heterogeneous treatment effects, we implement a binned "event-study" approach. We divide the MSE-optimal bias bandwidth ($\pm 114$ cm) into 14 equal-width bins—seven on each side of the scoring threshold—and estimate bin fixed effects interacted with skin tone indicators, controlling for shot- and match-level covariates. This flexible specification allows us to assess (1) whether pre-threshold ratings evolve in parallel across skin tone groups, supporting the credibility of the quasi-experimental design, and (2) whether post-threshold effects diverge, indicating heterogeneous recognition for identical performance.

Figure 5 presents the results. The left panel shows that pre-threshold rating profiles for players who narrowly miss the goal are flat and nearly identical across skin tone groups. A joint F-test of the bin-by-skin-tone interactions fails to reject the null of parallel trends ($F = 0.29$, $p = 0.863$), consistent with the identifying assumption that near-threshold shot outcomes are quasi-random and players would receive similar ratings absent scoring.

The right panel reveals clear divergence immediately after the scoring threshold. All groups experience an upward jump in ratings, but the magnitude varies systematically: ratings rise to roughly 8.2 for light-skinned players, 8.0 for tan-skinned players, and 7.8 for dark-skinned players. These differences persist within the MSE-optimal bandwidth and are strongly rejected as equal by a joint F-test ($F = 4.76$, $p = 0.002$).

Importantly, the gradient attenuates beyond 65 cm from the threshold, where all groups begin to converge. This pattern suggests that algorithmic discrimination is most pronounced for marginal goals—those barely crossing the line—where evaluator discretion is greatest and the distinction

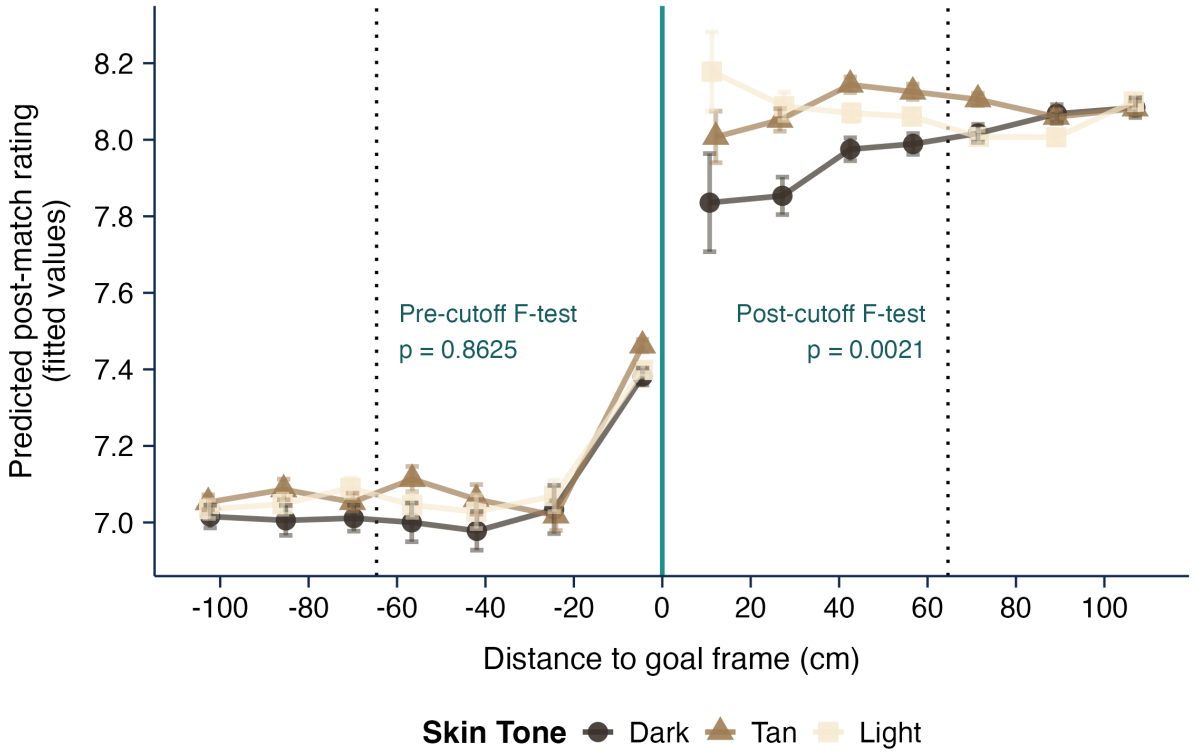



Figure 5: Binned Event-Study: Parallel Trends and Heterogeneous Treatment Effects

*Notes:* This figure shows a binned event-study analysis around the scoring threshold. The bias-correction bandwidth (±114 cm) is divided into 14 equal-width bins (7 on each side), with the solid vertical line marking the scoring cutoff. Each point represents predicted post-match ratings from regressions of player ratings on bin fixed effects interacted with skin tone, controlling for shot and match characteristics. Error bars represent 95% confidence intervals, two-way clustered by player and match. Dashed lines denote the main MSE-optimal bandwidth (±65 cm). The left panel (pre-threshold), tests for 'parallel trends' across skin tone groups prior to the threshold. The right panel (post-threshold), tests for heterogeneous treatment effects. All specifications include controls for shot coordinates, home status, starter status, and fixed effects for match minute, shot situation, player role, shirt number, goal segment, shot type, league, and season.

between skill and luck is least clear. The localization of the gradient within the optimal bandwidth further strengthens the causal interpretation of our findings.

**Inference: Bayesian Bootstrap with Two-Way Clustering** To address dependence in our data—from players taking multiple shots and shots occurring within the same match—we implement a two-way clustered Bayesian bootstrap following Rubin (1981) and Cameron et al. (2011). We generate 5,000 resamples by drawing Dirichlet weights over player and match clusters, preserving both sources of correlation. For each resample, we re-estimate the heterogeneous RDD following Calonico et al. (2025*b*) using the MSE-optimal bandwidth and extract bias-corrected treatment effects.

Appendix Figure B.3 shows the pairwise differences. Light-skinned players receive 0.34 standard deviations more recognition than Dark-skinned players ($p = 0.077$) and 0.39 more than Tan-skinned players ($p = 0.018$). The difference between Tan and Dark players is small and insignificant (0.05, $p = 0.592$). These results confirm a robust colorism gradient, even under more conservative two-way clustered inference.

Appendix Figure B.6 displays the bootstrap distributions. Panel A shows kernel densities of the estimated goal premia by group: Light-skinned players average 0.85, compared to 0.50 (Dark) and 0.46 (Tan). Panel B shows ECDFs: Light-skinned players dominate at all quantiles, indicating first-order stochastic dominance. Kolmogorov-Smirnov tests confirm these differences are statistically significant between Light vs. Dark and Light vs. Tan ($p < 0.001$), with smaller differences between Dark and Tan ($p < 0.001$). These findings reinforce our main result: algorithmic recognition for goals is systematically biased in favor of lighter skin tones.

**Balance Tests for Pre-Determined Covariates** To validate the RDD design, we test for discontinuities in 13 pre-determined covariates across the scoring threshold. Under quasi-random assignment, characteristics determined prior to the shot—such as coordinates, match minute, or home status—should evolve smoothly. We estimate RDDs separately by skin tone group (Calonico et al., 2025*b*), applying FDR-adjusted $q$-values within group (Anderson, 2008). Continuous variables are standardized; binary variables are left unchanged.

Appendix Table B.2 reports the results. Across all groups, covariates are balanced at the threshold. Only 1 of 39 comparisons is statistically significant after correction. This strongly supports the validity of our DiDC framework: marginal scoring outcomes are not systematically correlated with observable shot, player, or match characteristics. Thus, estimated differences in goal premia cannot be attributed to confounding imbalances.

**Within-Match Performance Metrics and Post-Treatment Controls** While pre-determined characteristics should not jump at the scoring threshold, scoring a goal may mechanically influence other within-match performance metrics—such as passes received, shots taken, or defensive pressure—that in turn affect post-match ratings. If these variables differ systematically by skin tone, omitting them could bias our estimates. Yet controlling for them raises concerns about post-treatment conditioning, as they are outcomes of scoring itself (i.e., "bad controls").

To address this tension, we examine whether within-match metrics shift discontinuously at the scoring threshold and test whether our discrimination results are robust to including them. Using the same group-specific RDD framework, we test 35 performance variables spanning offense, passing, defense, and discipline. Group-specific FDR correction at 10% controls for multiple testing.

Appendix Table B.3 shows that six metrics exhibit significant jumps post-scoring: the dependent variable (rating), goals scored, big chances missed, missed penalties, accurate long balls, and dispossessions. The latter three are plausible non-mechanical post-treatment outcomes that could mediate the relationship between scoring and ratings.

To test robustness, we re-estimate the DiDC model controlling for these three non-mechanical outcomes. Appendix Figure B.7 shows that the colorism gradient persists. Compared to our main estimates, controlling for post-treatment outcomes slightly increases the magnitude of all group-level effects—consistent with scoring improving subsequent play—but leaves the skin tone gaps largely unchanged. This robustness check demonstrates that differential recognition by skin tone is not driven by downstream differences in on-field performance following a goal.

**Robustness: Local Randomization Inference** Our main RDD approach relies on continuity assumptions—namely, that potential outcomes evolve smoothly around the scoring threshold. This permits the use of local polynomials within a relatively wide bandwidth (112 cm). As a more conservative alternative, we implement a Local Randomization RDD (Cattaneo et al., 2015), which treats treatment assignment as quasi-random within a narrow window around the threshold.

We define a ±15 cm window around the scoring threshold (roughly 20% of a football's radius) and assume as-if random assignment within this range. This method imposes stronger assumptions (randomization rather than continuity) but avoids reliance on functional form and bandwidth choices.

We estimate group-specific treatment effects using permutation-based $p$-values. Figure B.8 shows the results: Dark-skinned players receive a goal premium of 0.41 standard deviations (95% CI: $[-0.15, 0.95]$), Tan-skinned players 0.64 [0.30, 0.93], and Light-skinned players 0.98 [0.57, 1.37].

The pairwise differences mirror our main findings. Light-skinned players receive significantly more recognition than Dark-skinned players ($\Delta = 0.57$, $p = 0.051$) and Tan-skinned players ($\Delta = 0.34$, $p = 0.095$); the Tan–Dark difference remains small and insignificant. Despite the narrower window and stricter assumptions, the colorism gradient persists, reinforcing the robustness of our conclusions.

**Placebo Test: Artificial Threshold Among Goals** To verify that our results reflect discrimination at the true scoring threshold—and not arbitrary differences in rating trajectories—we conduct a placebo test within the set of scored goals. We re-center the RDD at an artificial threshold: the average distance of goals from the goal frame (approximately 70 cm inside). This point does not represent any meaningful performance change, so no discontinuity in ratings should appear if our design is valid.

Using the same MSE-optimal bandwidth, covariates, and DiDC specification, we estimate group-specific placebo effects. Appendix Figure B.9 shows that all estimates are near zero and statistically insignificant. Pairwise differences are similarly negligible. The absence of a rating gap at this artificial threshold reinforces the credibility of our main findings: the colorism gradient arises precisely at the true scoring cutoff, where evaluations respond to goal outcomes—not at arbitrary points along the shot distance distribution.

**Heterogeneity in Other Player Characteristics** A potential concern with our skin tone findings is that players with different skin tones may also differ systematically in other observable characteristics—such as nationality, physical attributes, or playing style—that could independently affect algorithmic ratings. If the algorithm exhibits preferential treatment along multiple correlated dimensions, our results might partly reflect those alternative channels rather than skin tone itself. To address this, we examine whether similar discriminatory patterns arise across five alternative player dimensions: shooting foot, nationality (native vs. non-native to the league country), age, height, and weight. We also include a two-group skin tone comparison (Light vs.

Tan & Dark) for benchmarking.[6]

For each dimension, we estimate a heterogeneous RDD specification identical to our main model, testing whether the goal premium differs across groups. Appendix Figure B.10 displays the results. Skin tone shows the clearest pattern: Light-skinned players receive a goal premium of 0.84 standard deviations, compared to 0.51 for Tan & Dark players—a difference of 0.303 ($p = 0.011$, $q = 0.071$). In contrast, the other five dimensions show no statistically significant differences after FDR correction.

These results reinforce the conclusion that the observed disparities reflect algorithmic bias based specifically on skin tone. The absence of significant differentials by nativity, age, height, weight, or footedness suggests that the algorithm does not systematically favor other player characteristics correlated with performance. The fact that discrimination appears along racialized appearance—but not nationality—further supports the interpretation that colorism, rather than xenophobia or stylistic bias, drives the unequal recognition we document.

# 6 Mechanisms

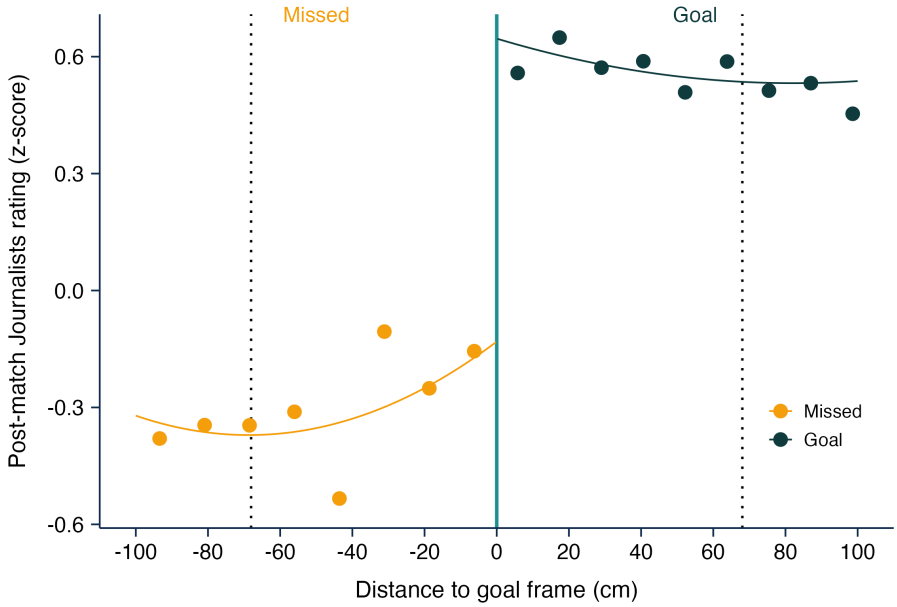
What explains the skin tone penalties in algorithmic ratings documented in Section 5? If algorithmic evaluations are intended to provide neutral, statistics-based assessments of player performance, the presence of a sizeable color gradient suggests that the algorithm must be incorporating information beyond on-field actions. A natural possibility is that the algorithm implicitly reflects human evaluation biases embedded in the data used for training, weighting, or calibration. To assess this mechanism, we examine whether *human* post-match ratings—assigned by professional sports journalists—exhibit similar disparities by skin tone.

## 6.1 Journalist-Based RDD Analysis

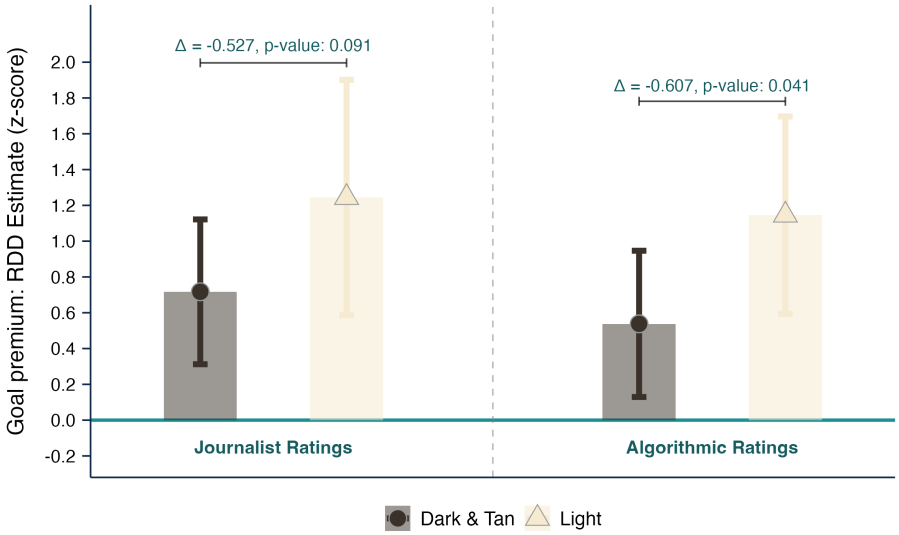
To test whether human evaluators display analogous color-based differentials, we replicate our RDD and DiDC strategy using journalist-assigned post-match ratings. These ratings originate from three major European leagues: the English Premier League, Italian Serie A, and French Ligue 1. Journalists affiliated with leading football media outlets publish player grades shortly after each match, making these assessments a prominent and public-facing benchmark for expert human evaluation (Sky Sports, 2023; Fantacalcio.it, 2023; L'Équipe, 2023).

Because journalist data are available only for a subset of leagues and seasons, we re-estimate the RDD bandwidth using this reduced sample. The empirical design otherwise mirrors Section 4.1: the running variable remains the shot's Euclidean distance to the goal frame, and we apply the same predetermined covariates and weighting scheme. We standardize journalist ratings into $z$-scores and purge match-level variation by demeaning within matches to accont for potential different journalist rating each match.

[6]For age, height, and weight, we merge our shot-level data with the SoFIFA player database. We extract unique players, harmonize units (e.g., convert height to cm, weight to kg), and use fuzzy string matching to align names and nationalities. After aggregating multiple entries per player and selecting the best match, we merge the demographic variables into our main dataset and create binary indicators for above/below-mean age, height, and weight.




(a) Regression Discontinuity: Journalist Ratings




(b) Difference-in-Discontinuities: Journalist vs. Algorithmic Ratings

Figure 6: Journalist-Based RDD and DiDC Estimates

*Notes:* The top panel shows the RDD estimate of the journalist-based goal premium, using an MSE-optimal bandwidth re-estimated on the journalist dataset. The bottom panel reports heterogeneous RDD estimates by skin tone (Light vs. Dark & Tan) for both journalist and algorithmic ratings. Bars display point estimates; lines denote 95% confidence intervals. $\Delta$ values represent pairwise differences in the estimated goal premium across skin tone groups, with associated $p$-values from one-sided tests of $H_0 : \Delta \leq 0$. All estimates include predetermined covariates and are weighted by the inverse number of within-bandwidth shots per player-match.

**RDD Results.** Figure 6 (top panel) shows the RDD plot for journalist ratings. We find a clear and statistically significant jump in ratings when a marginal shot becomes a goal. The estimated goal premium is comparable in magnitude to that obtained from algorithmic ratings, indicating that journalists reward narrowly scored goals to a similar extent.

**Difference-in-Discontinuities by Skin Tone.** To evaluate whether journalists exhibit heterogeneous recognition by skin tone, we estimate group-specific RDDs and compute pairwise differences in the goal premium between lighter- and darker-skinned players. Because the journalist dataset is smaller and includes fewer observations per group, we partition players into two mutually exclusive categories: Light, and Dark & Tan. Figure 6 (bottom panel) reports the resulting DiDC estimates for both journalist and algorithmic ratings, using the identical restricted sample for comparability.

Journalist ratings reveal a clear skin tone gradient: Light-skinned players receive a substantially larger goal premium than Dark and Tan players. The estimated gap (0.53 standard deviations) is economically meaningful and statistically significant at conventional levels. When we repeat the exercise with algorithmic ratings on the same subsample, we obtain a strikingly similar differential.

Taken together, these results indicate that algorithmic disparities are unlikely to stem solely from internal model mechanics. Instead, the close alignment between algorithmic and journalist-based gaps is consistent with a mechanism in which algorithmic evaluations partially inherit human-evaluation biases embedded in the data environment used to train, calibrate, or benchmark the algorithm.

### 6.2 Algorithmic Ratings and Fan Market Valuations

f algorithmic ratings reflect learned human preferences, they should align with market-based valuations that capture collective assessments of player worth. To test this, we examine the relationship between FotMobs algorithmic season ratings and SoFIFA market values—a platform that aggregates crowd-sourced estimates of player transfer values from Transfermarkt and player ratings from EA Sports FIFA.[7]

Transfermarkt, a German-based site that gained prominence after the 2006 World Cup, has become the de facto standard for player market valuations. While its estimates are derived from fan discussions and staff adjustments, they are widely cited by media outlets, referenced by club managers in negotiations, and even used in legal and financial contexts (Smith, 2021; Coates and Parshakov, 2022). SoFIFA closely mirrors these valuations, confirming that both platforms reflect a unified, crowd-driven consensus.[8]

These values approximate revealed preference or willingness-to-pay in the football labor market, aggregating thousands of individual judgments into influential price signals. Their widespread use in transfer negotiations suggests they not only reflect but also help shape real economic outcomes.

Figure 7 plots binscatter estimates of the relationship between algorithmic ratings and log market values, stratified by skin tone. We find a strong positive association across all groups: higher FotMob ratings predict higher market valuations. Crucially, the slopes are statistically indistinguishable by skin tone, implying that darker-skinned players—who receive lower algorithmic ratings (as shown in our RDD analysis)—see those depressed ratings translate proportionally into lower perceived market worth.

[7] See Appendix Figure B.12a for an example of a FIFA player rating card and Figure B.12b for the SoFIFA interface.

[8] See Appendix Figure B.13 for a visual comparison showing a correlation between SoFIFA and Transfermarkt market values.

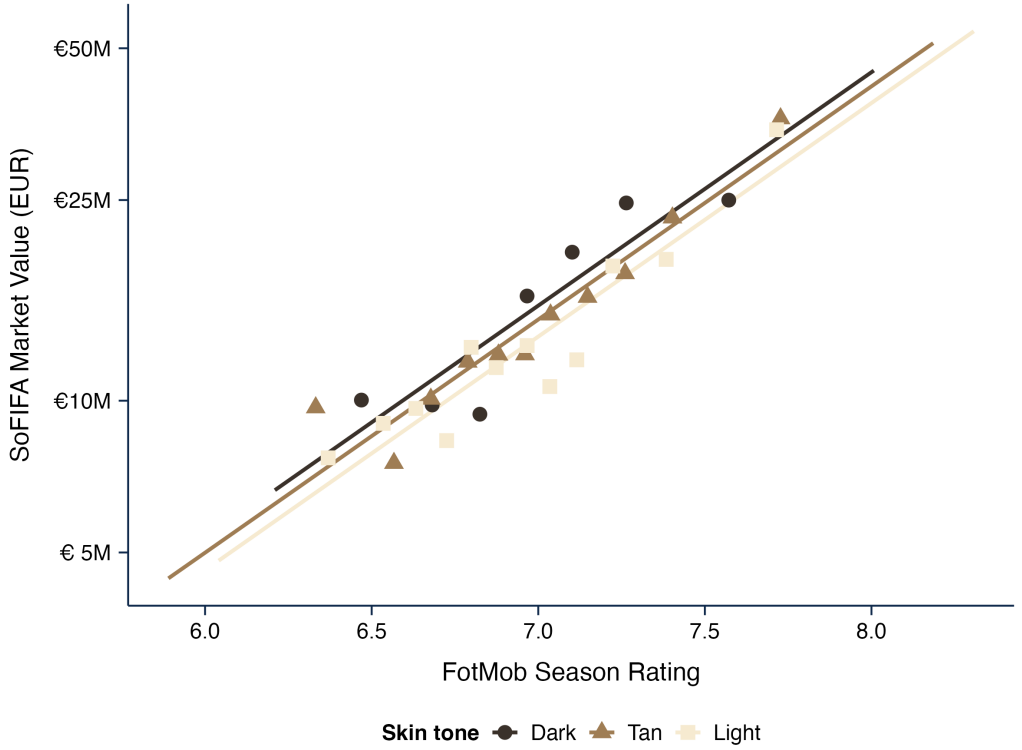



Figure 7: Algorithmic Ratings and Market Valuations by Skin Tone

*Notes:* This figure shows the relationship between FotMob algorithmic season ratings (x-axis) and SoFIFA crowd-sourced market values in EUR (y-axis, log scale), estimated using binscatter with linear polynomial fits. Points represent binned means; lines show fitted values by skin tone group. SoFIFA market values aggregate Transfermarkt's crowd-sourced valuations (see Appendix Figure B.13 for validation of this correspondence). Data are matched across platforms using fuzzy string matching (Jaro-Winkler distance) on player names and clubs, with maximum distance threshold of 0.10 (90% similarity or higher). Sample includes players from seasons 2020/21–2022/23 with non-missing values in both datasets. The parallel slopes across skin tone groups indicate that algorithmic ratings predict market values equally within each group, but darker-skinned players systematically occupy lower positions on both dimensions—consistent with correlated discrimination across evaluation systems.

This alignment between algorithmic and crowd-based evaluations supports our interpretation that algorithmic bias is inherited rather than generated: algorithms reflect broader human biases embedded in the data ecosystem. Alongside our findings from journalist ratings, the evidence suggests that colorism in football operates systemically across algorithmic, expert, and crowd-sourced evaluations—each reinforcing disparities in recognition and economic value.

### 6.3 Heterogeneity

Since algorithmic ratings may reflect context-dependent evaluation patterns, we explore whether the colorism gradient in goal-scoring premiums varies across match circumstances. We conduct heterogeneity analyses using the Local Randomization Regression Discontinuity (RDD) framework with a narrow window of ±15 cm around the scoring threshold. Given that we stratify by both skin tone and contextual dimensions, sample sizes within each subgroup become limited. The local randomization approach is particularly well suited for these settings, as it provides robust inference via randomization-based tests that remain valid in finite samples and do not rely on large-sample approximations.

**Home vs. Away.** We first assess whether the discriminatory rating premium varies depending on whether a player is playing at home or away. Appendix Figure B.14 presents the estimated goal-scoring premiums by skin tone group, separately for away matches (left panel) and home matches (right panel), along with pairwise differences. In away matches, we observe a modest colorism gradient and not statistically significant. However, the skine tone gaps are substantially more pronounced in home matches. Light-skinned players receive 0.361 standard deviations more recognition than dark-skinned players for identical goals ($p = 0.003$), and 0.216 standard deviations more than tan-skinned players ($p = 0.013$). These results suggest that home-crowd presence amplifies bias, potentially reflecting the algorithm's greater responsiveness to fan reactions, media narratives, or other home-context factors that favor lighter-skinned players.

**Match Result.** The outcome of a match may shape how much players are rewarded for scoring a goal, and whether this reward varies by skin tone. We estimate the goal-scoring premium separately for each skin tone group depending on whether their team wins or experiences a loss or tie. Appendix Figure B.15 reveals that the colorism gradient persists across both match outcomes but is more pronounced in losses or ties. When teams lose or draw, light-skinned players receive 0.212 standard deviations more recognition than dark-skinned players for scoring ($p = 0.061$). In contrast, when teams win, the light-dark difference is smaller at 0.187 standard deviations ($p = 0.088$), though still statistically meaningful. The tan-light differences follow a similar pattern: 0.095 standard deviations in losses/ties ($p = 0.212$) versus 0.089 standard deviations in wins ($p = 0.181$). These results suggest that algorithmic discrimination in goal recognition is relatively stable across match contexts, though it may be slightly amplified when teams underperform. Notably, darker-skinned players face unequal recognition regardless of team success, indicating that the bias is systematic rather than confined to particular emotional or narrative contexts.

### 6.4 Downstream Labour Market Consequences

Our main results demonstrate that algorithmic player ratings exhibit systematic bias favoring lighter-skinned players for identical on-field performance. A natural question follows: do these biased evaluations translate into tangible economic consequences? If algorithmic ratings influence scouts, journalists, fans, and other stakeholders in the football labor market, discrimination in performance measurement could compound into discrimination in wages and market valuations. We investigate this hypothesis by examining whether the seasonal accumulation of biased match-level ratings predicts differential labor market outcomes across skin tone groups.

We proceed in three steps. First, we aggregate distance-weighted near-goals to the player-season level, generating a measure of quasi-random scoring success that should be equally valuable across skin tone groups. Second, we test whether this measure predicts higher season-average ratings, and whether this relationship varies by skin tone—replicating our match-level findings at the seasonal scale. Third, we link our data to external sources on player wages (Capology/FBref) and market valuations (Transfermarkt) using fuzzy string matching on standardized player names within league-season cells. We successfully match 99.7% of our 3,191 player-season observations.[9]

[9] We achieve 94.7% exact matches, 3.2% fuzzy matches on full names (Jaro-Winkler distance $\leq 0.15$), and 2.0%

We then estimate whether season-average ratings predict subsequent (t+1) economic outcomes differentially by skin tone. All specifications include team×season, nationality, and position fixed effects, and control for other performance metrics (non-random goals, assists, minutes played) as well as demographic characteristics (age).

Table 1 presents the results. Specifications (1) and (2) confirm that distance-weighted near-goals—our measure of quasi-random scoring success—significantly increase season-average ratings. Crucially, Specification (2) shows that this effect is 49% larger for light-skinned players compared to dark-skinned players (interaction coefficient = 0.049, $p < 0.10$), indicating that algorithmic bias accumulates over the course of a season. Specifications (3) and (4) demonstrate that higher season-average ratings strongly predict wage growth in the subsequent season. However, Specification (4) reveals that light-skinned players convert ratings into 12.5 percentage points higher wage growth than dark-skinned players (interaction coefficient = 0.141, $p = 0.06$). Similarly, Specifications (5) and (6) show that ratings predict market value appreciation, with light-skinned players experiencing 10.1 percentage points higher returns (interaction coefficient = 0.101, $p = 0.08$).

These findings indicate that algorithmic bias in match ratings distorts seasonal performance metrics, which in turn lead to differential economic rewards. Even after controlling for actual performance (goals, assists, minutes played), skin tone shapes both how quasi-random success is evaluated and how those evaluations are translated into labor market outcomes. A dark-skinned and a light-skinned player who score an identical number of quasi-random goals receive different seasonal ratings and, conditional on those ratings, experience divergent wage and market value trajectories. This pattern suggests that colorism in football operates through multiple channels—affecting both the assessment of performance and the monetization of that assessment.

# 7 Conclusion

This paper provides causal evidence that players with darker skin receive systematically lower rewards for identical on-field performance. Using near-miss variation in professional football and a perception-consistent measure of skin tone, we show that the causal premium for scoring a marginal goal is substantially larger for Light players than for Tan or Dark players. These disparities emerge in both algorithmic and journalist evaluations, suggesting that algorithmic systems inherit and formalize biases present in the broader evaluation environment rather than generating them independently.

We further demonstrate that these biased evaluations have tangible economic consequences. Season-level algorithmic ratings predict subsequent wages and market values, and these returns differ significantly by skin tone: light-skinned players convert identical seasonal ratings into higher wage growth and greater increases in market value than darker-skinned peers. This establishes a troubling causal chain in which quasi-random on-field success leads to differential ratings, which in turn result in unequal labor market outcomes. The alignment between algorithmic and crowd-sourced valuations reinforces our interpretation that algorithms replicate existing human biases—colorism embedded in expert judgment, fan perception, and institutional

fuzzy matches validated by team membership.

Table 1: Algorithmic Bias and Labor Market Outcomes

| | Season avg. rating | | Log weekly wage$_{t+1}$ | | Log market value$_{t+1}$ | |
|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) |
| Distance-weighted near-goals (z) | 0.107 | 0.074 | | | | |
| | (0.018) | (0.023) | | | | |
| Tan skin | | 0.094 | | -0.012 | | -0.072 |
| | | (0.038) | | (0.091) | | (0.069) |
| Light skin | | 0.085 | | -0.031 | | -0.126 |
| | | (0.040) | | (0.096) | | (0.071) |
| | | (0.040) | | | | |
| Near-goals (z) × Tan | | 0.042 | | | | |
| | | (0.026) | | | | |
| Near-goals (z) × Light | | 0.049 | | | | |
| | | (0.028) | | | | |
| Season avg. rating (z) | | | 0.226 | 0.113 | 0.402 | 0.360 |
| | | | (0.031) | (0.065) | (0.023) | (0.053) |
| Rating (z) × Tan | | | | 0.122 | | 0.011 |
| | | | | (0.072) | | (0.058) |
| Rating (z) × Light | | | | 0.141 | | 0.101 |
| | | | | (0.075) | | (0.058) |
| Observations | 2,905 | 2,905 | 1,097 | 1,097 | 1,067 | 1,067 |
| $R^2$ | 0.760 | 0.761 | 0.772 | 0.773 | 0.854 | 0.856 |
| Within $R^2$ | 0.566 | 0.568 | 0.182 | 0.187 | 0.621 | 0.626 |
| Team×Season FE | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Nationality FE | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Position FE | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Performance controls | ✓ | ✓ | – | – | – | – |
| Age controls | – | – | ✓ | ✓ | ✓ | ✓ |

*Notes:* This table examines whether algorithmic bias in match ratings translates into differential labor market outcomes. Models (1)-(2) aggregate distance-weighted near-goals (quasi-random scoring success within the RDD bandwidth) to the player-season level and test whether they predict season-average FotMob ratings. Models (3)-(4) test whether season-average ratings predict log weekly wages in the following season ($t + 1$), measured from Capology/FBref. Models (5)-(6) test whether ratings predict log market values in the following season, measured from Transfermarkt. All models include team×season, nationality, and position fixed effects. Performance controls (Columns 1-2) include standardized measures of non-random goals, assists, minutes played, and the count of near-shots. Age controls (Columns 3-6) include age and age squared. The reference category for skin tone is Dark. Standard errors (in parentheses) are clustered at the player level.

behavior—across multiple channels of evaluation.

Our findings contribute to the broader literature on discrimination by showing that colorism persists even in contexts where productivity is transparent and performance is evaluated at high frequency. They also underscore how modern labor markets—mediated by algorithmic, expert, and crowd-based systems—can replicate or amplify inequality under the guise of objectivity. Future work will examine the underlying mechanisms through online and field experiments and assess whether exposure to biased ratings influences evaluator behavior. Together, these results highlight the importance of designing evaluation systems—human or algorithmic—that resist rather than reproduce the biases embedded in their data environments.

## References


Abramitzky, R., Conway, J., Mill, R. and Stein, L. (2023), The gendered impacts of perceived skin tone: Evidence from african-american siblings in 1870–1940.

Adukia, A., Eble, A., Harrison, E., Runesha, H. B. and Szasz, T. (2023), 'What we teach about race and gender: Representation in images and text of childrens books', *The Quarterly Journal of Economics* **138**, 2225–2285.
**URL:** *https://academic.oup.com/qje/article/138/4/2225/7247000*

Adukia, A., Hornbeck, R., Keniston, D. and Lualdi, B. (2025), 'The social construction of race during reconstruction'.
**URL:** *https://www.nber.org/papers/w33502*

Alrababa'H, A., Marble, W., Mousa, S. A. and Siegel, A. A. (2021), 'Can exposure to celebrities reduce prejudice? the effect of mohamed salah on islamophobic behaviors and attitudes', *American Political Science Review* **115**, 1111–1128.
**URL:** *https://www.cambridge.org/core/journals/american-political-science-review/article/can-exposure-to-celebrities-reduce-prejudice-the-effect-of-mohamed-salah-on-islamophobic-behaviors-and-attitudes/A1DA34F9F5BCE905850AC8FBAC78BE58*

Alrababah, A., Marble, W., Mousa, S. and Siegel, A. A. (2024), Are minorities punished more harshly for underperformance? evidence from premier league soccer. OSF Preprints 7d2cu, Center for Open Science.

Anderson, M. L. (2008), 'Multiple inference and gender differences in the effects of early intervention: A reevaluation of the abecedarian, perry preschool, and early training projects', *Journal of the American Statistical Association* **103**, 1481–1495.
**URL:** *https://www.tandfonline.com/doi/full/10.1198/016214508000000841*

Arceo-Gomez, E. O. and Campos-Vazquez, R. M. (2014), 'Race and marriage in the labor market: A discrimination correspondence study in a developing country', *American Economic Review* **104**(5), 376–80.

Arnold, D., Dobbie, W. and Hull, P. (2022), 'Measuring racial discrimination in bail decisions', *American Economic Review* **112**(9), 2992–3038.
**URL:** *https://www.aeaweb.org/articles?id=10.1257/aer.20201653*

Arrow, K. (1973), 'The theory of discrimination', *O Ashenfelter and A Rees (eds), Discrimination in Labor Markets* .

Bartalotti, O., Calhoun, G. and He, Y. (2017), 'Bootstrap confidence intervals for sharp regression discontinuity designs', *Advances in Econometrics* **38**, 421–453.
**URL:** *https://www.emerald.com/books/edited-volume/13936/chapter/84780022/Bootstrap-Confidence-Intervals-for-Sharp*

Becker, G. S. (1957), *The economics of discrimination*, University of Chicago press.

Belloni, A., Chernozhukov, V. and Hansen, C. (2014), 'Inference on treatment effects after selection among high-dimensional controls', *The Review of Economic Studies* **81**, 608–650.
**URL:** *https://dx.doi.org/10.1093/restud/rdt044*

Bertrand, M. and Duflo, E. (2017), 'Field experiments on discrimination', *Handbook of economic field experiments* **1**, 309–393.

Bertrand, M. and Mullainathan, S. (2004), 'Are emily and greg more employable than lakisha and jamal? a field experiment on labor market discrimination', *American Economic Review* **94**(4), 991–1013.

Bohren, J. A., Hull, P. and Imas, A. (2025), 'Systemic discrimination: Theory and measurement', *The Quarterly Journal of Economics* .
**URL:** *https://dx.doi.org/10.1093/qje/qjaf022*

Calonico, S., Cattaneo, M. D., Farrell, M. H., Palomba, F. and Titiunik, R. (2025*a*), 'Treatment effect heterogeneity in regression discontinuity designs'.
**URL:** *http://arxiv.org/abs/2503.13696*

Calonico, S., Cattaneo, M. D., Farrell, M. H., Palomba, F. and Titiunik, R. (2025*b*), Treatment effect heterogeneity in regression discontinuity designs. Working paper; methods implemented in the `rdhte` package.

Calonico, S., Cattaneo, M. D. and Titiunik, R. (2014), 'Robust nonparametric confidence intervals for regression-discontinuity designs', *Econometrica* **82**(6), 2295–2326.

Cameron, A. C., Gelbach, J. B. and Miller, D. L. (2011), 'Robust inference with multiway clustering', *Journal of Business and Economic Statistics* **29**, 238–249.
**URL:** *https://www.tandfonline.com/doi/abs/10.1198/jbes.2010.07136*

Campos-Vázquez, R. M. (2020), 'The higher price of whiter skin: an analysis of escort services', *Applied Economics Letters* .

Caselli, M., Falco, P. and Mattera, G. (2023), 'When the stadium goes silent: How crowds affect the performance of discriminated groups', *Journal of Labor Economics* **41**(2), 431–451.

Cattaneo, M. D., Frandsen, B. R. and Titiunik, R. (2015), 'Randomization inference in the regression discontinuity design: An application to party advantages in the US Senate', *Journal of Causal Inference* **3**(1), 1–24.

Cattaneo, M. D., Titiunik, R., Yu, R., Gautier, E., Hanin, B., Klusowski, J., Londoño-Vélez, J., Ma, X., Shigida, B., Shkolnikov, M. and Wooldridge, J. (2025), 'Estimation and inference in boundary discontinuity designs', *arXiv Working Paper* .
**URL:** *https://arxiv.org/pdf/2505.05670*

Chardon, A., Cretois, I. and Hourseau, C. (1991), 'Skin colour typology and suntanning pathways', *International journal of cosmetic science* **13**(4), 191–208.

Coates, D. and Parshakov, P. (2022), 'The wisdom of crowds and transfer market values', *European Journal of Operational Research* **301**, 523–534.

Cole, K., Wilson, M. and Hall, R. (2014), 'The color complex: The politics of skin color in a new millennium'.

Colombe, K., Krumer, A., Lavelle-Hill, R. and Pawlowski, T. (2025), Racial bias, colorism, and overcorrection: Evidence from WNBA refereeing. Working paper.

Darity, William A., J. (2022), 'Position and possessions: Stratification economics and intergroup inequality', *Journal of Economic Literature* **60**(2), 400–426.

Davenport, L. (2020), 'The fluidity of racial classifications', *Annual Review of Political Science* **23**(1), 221–240.
**URL:** *https://doi.org/10.1146/annurev-polisci-060418-042801*

Depetris-Chauvin, E., Durante, R. and Campante, F. (2020), 'Building nations through shared experiences: Evidence from african football', *American Economic Review* **110**(5), 1572–1602.

Derenoncourt, E., Kim, C. H., Kuhn, M. and Schularick, M. (2023), 'Wealth of two nations: The us racial wealth gap, 1860-2020', *The Quarterly Journal of Economics* .

Deschamps, P. and De Sousa, J. (2021), 'Labor mobility and racial discrimination', *European Economic Review* **135**, 103738.

Dixon, A. R. and Telles, E. E. (2017), 'Skin color and colorism: Global research, concepts, and measurement', *Annual Review of Sociology* .

Doleac, J. L. and Stein, L. C. (2013), 'The visible hand: Race and online market outcomes', *The Economic Journal* **123**(572), F469–F492.
**URL:** *http://www.jstor.org/stable/42919259*

Espino, R. and Franz, M. M. (2002), 'Latino phenotypic discrimination revisited: The impact of skin color on occupational status', *Social Science Quarterly* **83**(2), 612–623.

Faltings, R., Krumer, A. and Lechner, M. (2023), 'RotJauneVerde: On linguistic bias of referees in Swiss soccer', *Kyklos* **76**(3), 380–406.

Fantacalcio.it (2023), 'Voti fantacalcio: Le pagelle dei giornalisti'. Accessed July 2025.
**URL:** *https://www.fantacalcio.it*

Gallo, E., Grund, T. and James Reade, J. (2013), 'Punishing the foreigner: implicit discrimination in the premier league based on oppositional identity', *Oxford Bulletin of Economics and Statistics* **75**(1), 136–156.

Goldsmith, A. H., Hamilton, D. and Darity, W. (2007), 'From dark to light: Skin color and wages among african-americans', *Journal of Human Resources* **42**(4), 701–738.

Goldsmith, A. H., Hamilton, D. and Darity, William, J. (2006), 'Shades of discrimination: Skin tone and wages', *American Economic Review* **96**(2), 242–245.

Grembi, V., Nannicini, T. and Troiano, U. (2016), 'Do fiscal rules matter?', *American Economic Journal: Applied Economics* **8**(3), 1–30.

Guryan, J. and Charles, K. K. (2013), 'Tastebased or statistical discrimination: The economics of discrimination returns to its roots', *The Economic Journal* **123**, F417–F432.
**URL:** *https://dx.doi.org/10.1111/ecoj.12080*

Hirschman, C. (2004), 'The origins and demise of the concept of race', *Population and Development Review* **30**(3), 385–415.

Hunter, M. (2007), 'The persistent problem of colorism: Skin tone, status, and inequality', *Sociology compass* **1**(1), 237–254.

Hunter, M. L. (2005), 'Gender, race, and the politics of skin tone'.

Jann, B. (2008), 'The blinder–oaxaca decomposition for linear regression models', *The Stata Journal* **8**(4), 453–479.

Kleven, H. J., Landais, C. and Saez, E. (2013), 'Taxation and international migration of superstars: Evidence from the european football market', *American economic review* **103**(5), 1892–1924.

Kline, P., Rose, E. K. and Walters, C. R. (2022), 'Systemic discrimination among large us employers', *The Quarterly Journal of Economics* **137**(4), 1963–2036.

Kolkur, S., Kalbande, D., Shimpi, P., Bapat, C. and Jatakia, J. (2017), 'Human skin detection using RGB, HSV and YCbCr color models', *arXiv preprint arXiv:1708.02694* .

Lambrecht, A. and Tucker, C. E. (2019), 'Algorithmic bias? an empirical study of apparent gender-based discrimination in the display of stem career ads', *Management Science* **65**(7), 2966–2981.

Lang, K. and Spitzer, A. K.-L. (2020), 'Race discrimination: An economic perspective', *Journal of Economic Perspectives* **34**(2), 68–89.

L'Équipe (2023), 'Notes des joueurs: Ligue 1 match ratings'. Accessed July 2025.
**URL:** *https://www.lequipe.fr*

Magistro, B. and Wack, M. (2023), 'Racial bias in fans and officials: Evidence from the italian serie a', *Sociology* **0**(0), 00380385221138332.

Monk, E. P. (2021), 'The unceasing significance of colorism: Skin tone stratification in the united states', *Daedalus* **150**, 76–90.
**URL:** *https://direct.mit.edu/daed/article/150/2/76/98313/The-Unceasing-Significance-of-Colorism-Skin-Tone*

Obermeyer, Z., Powers, B., Vogeli, C. and Mullainathan, S. (2019), 'Dissecting racial bias in an algorithm used to manage the health of populations', *Science* **366**(6464), 447–453.

Otsu, N. (1979), 'A threshold selection method from gray-level histograms', *IEEE transactions on systems, man, and cybernetics* **9**(1), 62–66.

Palacios-Huerta, I. (2003), 'Professionals play minimax', *The Review of Economic Studies* **70**(2), 395–415.

Palacios-Huerta, I. (2025), 'The beautiful dataset', *Journal of Economic Literature* **63**, 1363–1423.
**URL:** *https://pubs.aeaweb.org/doi/10.1257/jel.20241616*

Phelps, E. S. (1972), 'The statistical theory of racism and sexism', *The American Economic Review* **62**(4), 659–661.

Picchetti, P., de Xavier Pinto, C. C. and Shinoki, S. T. (2024), 'Difference-in-discontinuities: Estimation, inference and validity tests'.
**URL:** *https://arxiv.org/pdf/2405.18531*

Principe, F. and van Ours, J. C. (2022), 'Racial bias in newspaper ratings of professional football players', *European Economic Review* **141**, 103980.

Reilly, B. and Witt, R. (2011), 'Disciplinary sanctions in english premiership football: Is there a racial dimension?', *Labour Economics* **18**(3), 360–370.

Rondilla, J. L. and Spickard, P. (2007), *Is lighter better?: Skin-tone discrimination among Asian Americans*, Rowman & Littlefield Publishers.

Rose, E. K. (2023), 'A constructivist perspective on empirical discrimination research', *Journal of Economic Literature* **61**, 906–23.

Rosenblum, A., Darity Jr, W., Harris, A. L. and Hamilton, T. G. (2016), 'Looking through the shades: the effect of skin color on earnings by region of birth and race for immigrants to the united states', *Sociology of Race and Ethnicity* **2**(1), 87–105.

Rubin, D. B. (1981), 'The bayesian bootstrap', *The Annals of Statistics* pp. 130–134.

Sen, M. and Wasow, O. (2016), 'Race as a bundle of sticks: Designs that estimate effects of seemingly immutable characteristics', *Annual Review of Political Science* **19**(1), 499–522.
**URL:** *https://doi.org/10.1146/annurev-polisci-032015-010015*

Sky Sports (2023), 'Football player ratings'. Accessed July 2025.
**URL:** *https://www.skysports.com/football*

Smith, R. (2021), 'The wisdom of the crowd: How transfermrkt helps determine the value of soccer players', *The New York Times* . Accessed: 2025-02-28.
**URL:** *https://www.nytimes.com/2021/08/12/sports/soccer/soccer-football-transfermarkt.html*

Woo-Mora, L. G. (2026), 'Unveiling the cosmic race: Skin tone and intergenerational economic disparities in latin america and the caribbean', *Journal of Development Economics* **179**, 103594.
**URL:** *https://linkinghub.elsevier.com/retrieve/pii/S0304387825001452*

Zivkovic, J. (2023), *worldfootballR: Extract and Clean World Football (Soccer) Data.* R package version 0.6.3.0013.
**URL:** *https://github.com/JaseZiv/worldfootballR*

# A Setting and Data

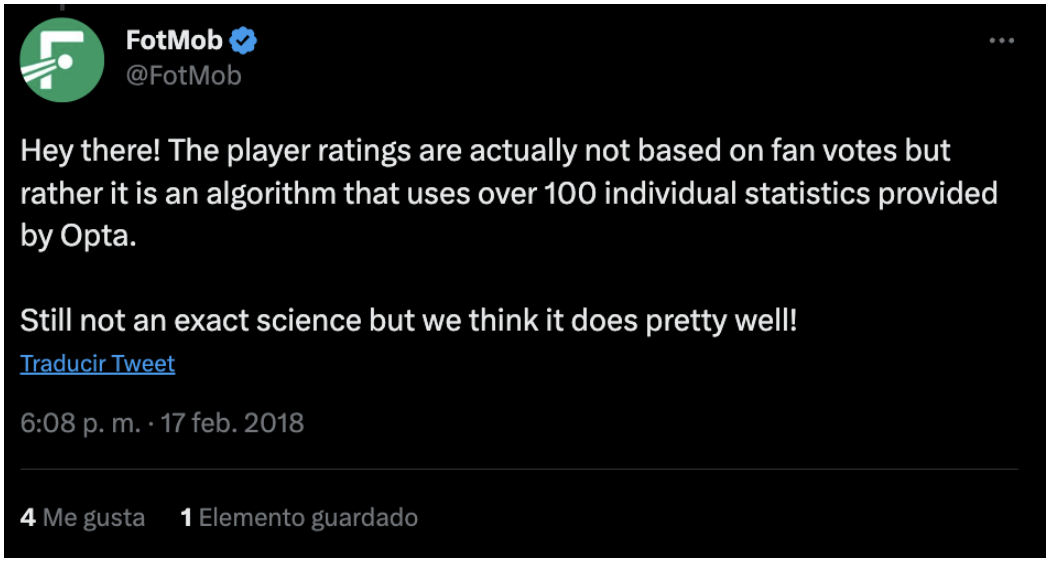
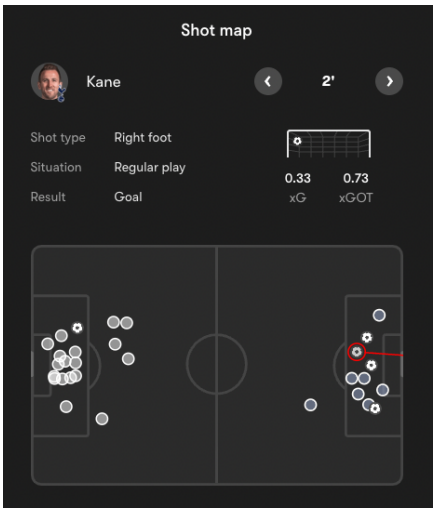



Figure A.1: FotMob Statement: Algorithm-based Player Ratings

*Notes*: Tweet from FotMob verified account on February 17, 2018, stating the algorithmic nature of their player ratings. No additional information about the algorithm's construction and implementation were given by FotMob. See tweet.




(a) FotMob Ratings




(b) FotMob Shot Map

Figure A.2: FotMob Data

*Notes*: FotMob ratings are algorithm based and provided post match. Events *within* a match are also reported. Panel (b) shows the shot map for a goal by Hary Kane. Geolocalized data on penalty shots with coordinates with respect to the goal provided by FotMob are used in our RD design.

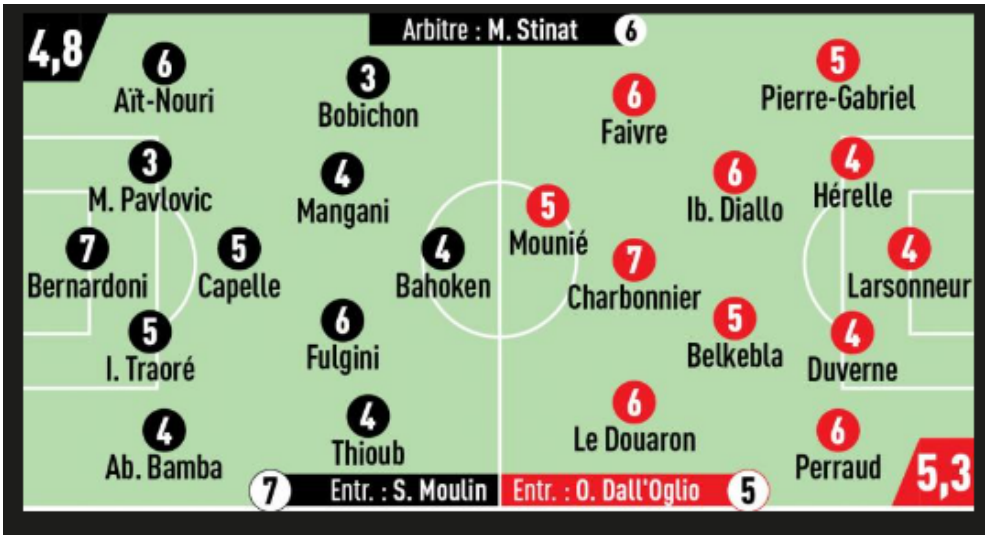



Figure A.3: Example of Post-Match Journalist Ratings published by L'Équipe

*Notes*: Figure shows a Ligue 1 match (Angers vs. Brest). Player ratings are displayed in a pitch diagram, with journalists assigning each player a score on a 1-10 scale based on in-match performance.

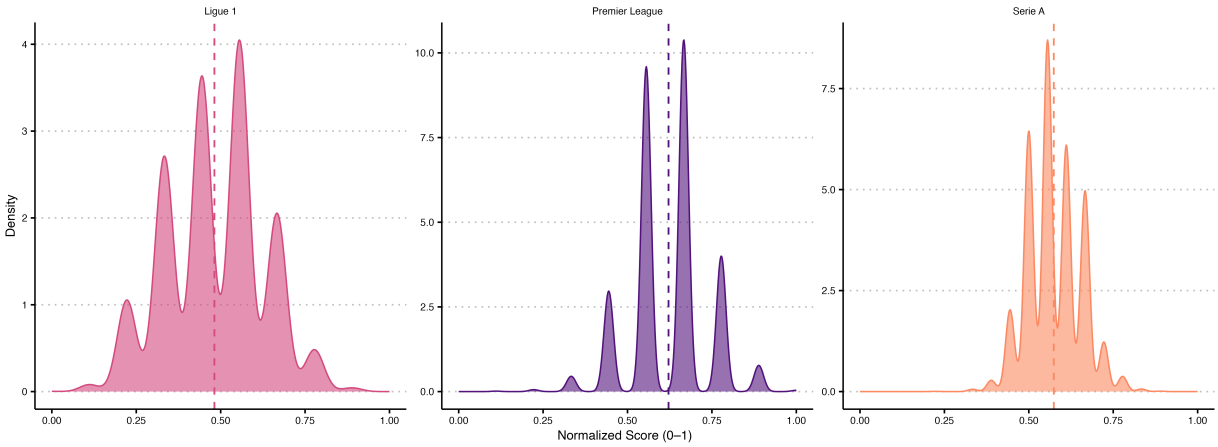



Figure A.4: Distributions of Journalist Ratings by League

*Notes*: Distributions of normalized journalist ratings for Ligue 1, the Premier League, and Serie A. Dashed lines denote the mean of ratings for each league.

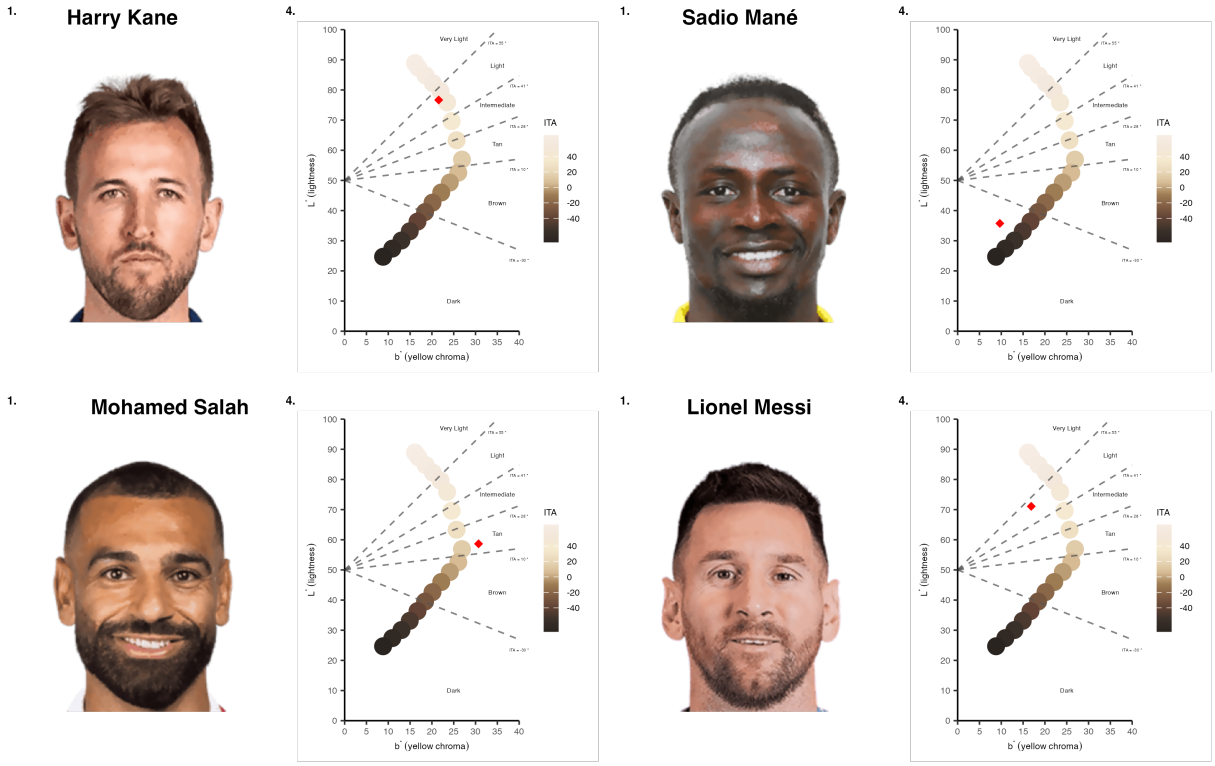
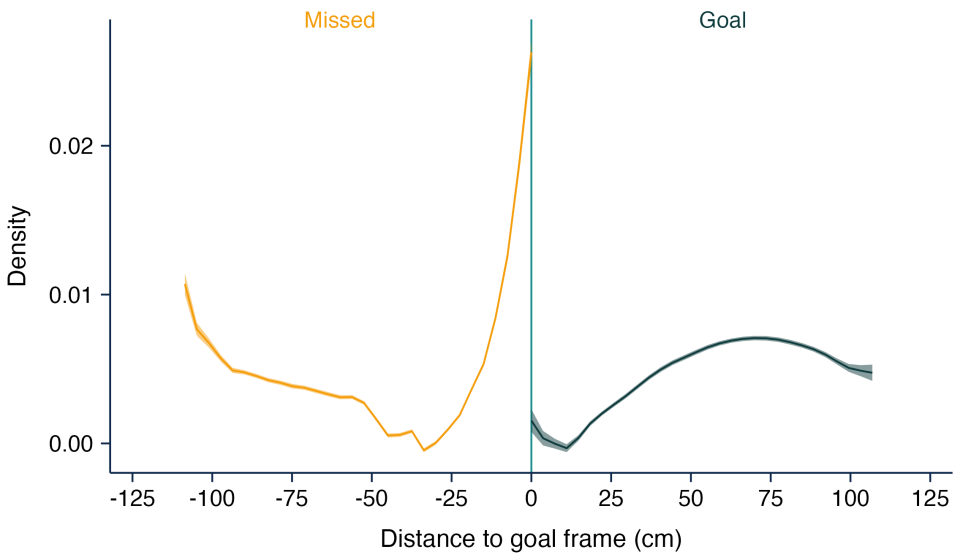


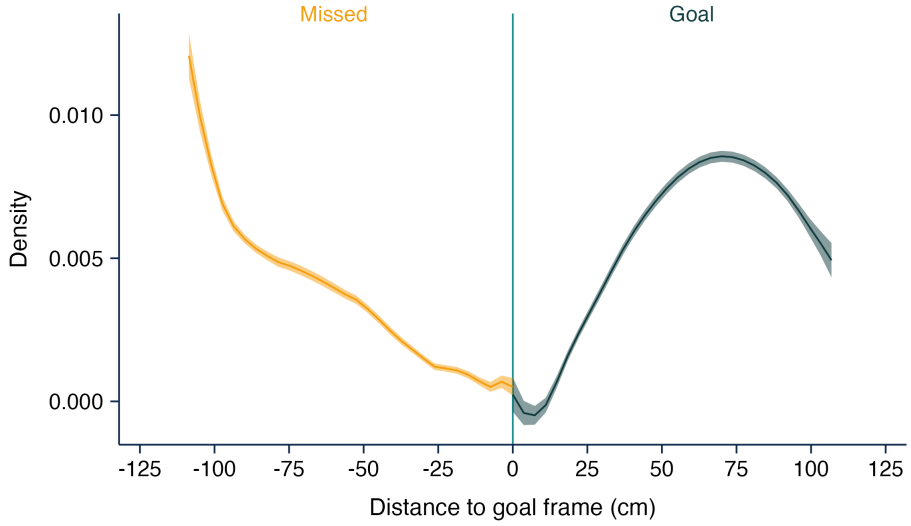

Figure A.5: Skin tone classification through ITA groupings for selected players

*Notes*: This figure displays the resulting skin tone classifications for four players based on their ITA values: Harry Kane and Lionel Messi are classified as Light, Mohamed Salah as Tan, and Sadio Mané as Dark. The ITA is computed from the CIELAB color space using the $L^*$ and $b^*$ values extracted from the players segmented facial regions. Higher ITA values correspond to lighter skin tones.

# B Results and Robustness




(a) Including shots on goal frame (post)




(b) Excluding shots on goal frame (post)

Figure B.1: Density of Shot Distances Around Goal Threshold

*Notes:* This figure shows the distribution of the running variable: shot distance to the goal frame. Negative values represent missed shots; positive values represent goals. Panel (a) includes all shots, including those that hit the post, which introduces visible bunching before the threshold. Panel (b) excludes post-hitting shots, yielding a smooth distribution and supporting the continuity assumption required for valid RDD estimation.

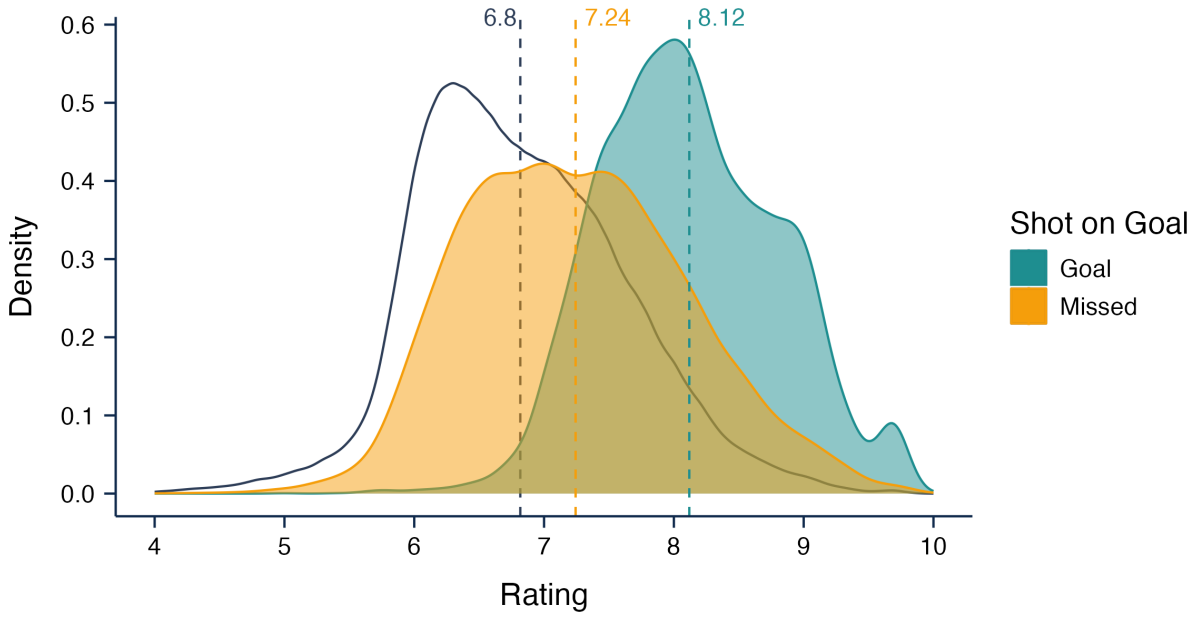



Figure B.2: Post-match Ratings Distribution: Scoring Goal Premium

*Notes:* This figure shows the distribution of algorithmic post-match ratings (FotMob) for three groups: the full universe of rated players (black line), players who missed a shot on goal (orange), and players who scored a goal (teal). The full sample includes all players, regardless of whether they took a shot or appeared in the close-shot sample. The distribution for players who missed closely resembles the overall distribution, while ratings for goal-scoring players are shifted to the right, with a noticeably higher mean (8.12 vs. 7.24). This visual evidence supports the interpretation that scoring a goal generates a reward in post-match ratings, rather than missing incurring a penalty.

Table B.1: Regression Discontinuity Estimates of Goal Premium on Player Ratings

| Specification | Estimate | Std. Error | CI Lower | CI Upper | p-value | Degree | Bwd h | Bias Bwd b | N (h) | N (b) | Kernel | VCE |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Levels | 0.499 | 0.058 | 0.385 | 0.613 | 0 | 1 | 65.045 | 111.955 | 8378 | 17922 | Epanechnikov | HC3 |
| Log points | 0.067 | 0.007 | 0.053 | 0.082 | 0 | 1 | 65.045 | 111.955 | 8378 | 17922 | Epanechnikov | HC3 |
| Z-score | 0.562 | 0.065 | 0.433 | 0.690 | 0 | 1 | 65.045 | 111.955 | 8378 | 17922 | Epanechnikov | HC3 |
| Z-score + Clean controls | 0.556 | 0.061 | 0.436 | 0.675 | 0 | 1 | 65.045 | 111.955 | 8378 | 17922 | Epanechnikov | HC3 |
| Z-score + DLASSO controls | 0.580 | 0.062 | 0.459 | 0.701 | 0 | 1 | 65.045 | 111.955 | 8378 | 17922 | Epanechnikov | HC3 |
| Z-score + DLASSO controls | 0.449 | 0.107 | 0.239 | 0.658 | 0 | 2 | 65.045 | 111.955 | 8378 | 17922 | Epanechnikov | HC3 |
| Z-score + DLASSO controls | 0.572 | 0.063 | 0.449 | 0.696 | 0 | 1 | 65.045 | 111.955 | 8378 | 17922 | Triangular | HC3 |
| Z-score + DLASSO controls | 0.657 | 0.056 | 0.547 | 0.767 | 0 | 1 | 65.045 | 111.955 | 8378 | 17922 | Uniform | HC3 |
| Z-score + DLASSO controls | 0.512 | 0.121 | 0.275 | 0.749 | 0 | 1 | 32.522 | 55.978 | 4212 | 6941 | Epanechnikov | HC3 |
| Z-score + DLASSO controls | 0.584 | 0.058 | 0.470 | 0.699 | 0 | 1 | 130.090 | 111.955 | 17922 | 17922 | Epanechnikov | HC3 |
| Z-score + DLASSO + Match FE | 0.485 | 0.079 | 0.331 | 0.639 | 0 | 1 | 65.045 | 111.955 | 8378 | 17922 | Epanechnikov | HC3 |

*Notes:* This table reports RDD estimates using the robust bias-corrected approach of Calonico et al. (2014), estimating the causal effect of narrowly scoring (versus narrowly missing) a goal on players post-match ratings. The outcome is the algorithmic rating from FotMob. Each row corresponds to a different specification—using alternative outcome scales (levels, logs, Z-scores), control sets (clean or DLASSO-selected), bandwiths, kernels, polynomials, and match fixed effects.

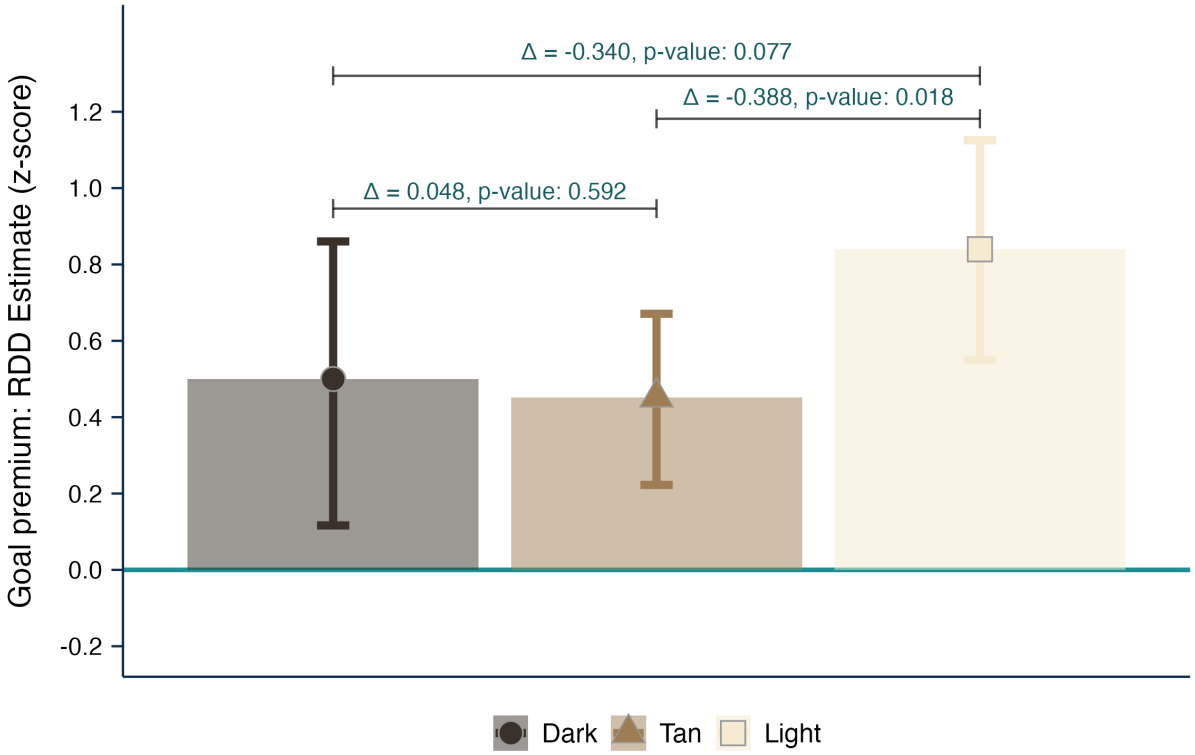



Figure B.3: Bayesian Bootstrap Estimates: Goal Premium by Skin Tone

*Notes:* This figure displays bias-corrected RDD estimates of the goal premium from a Bayesian bootstrap procedure with two-way cluster resampling. We generate 5,000 bootstrap samples by independently drawing Dirichlet weights for player clusters ($n$ = 2,361) and match clusters ($n$ = 5,735), then combining these perturbations to create observation-level weights that preserve correlation within both players (who may take multiple shots) and matches (where shots share common shocks). For each bootstrap sample, we re-estimate the heterogeneous RDD using `rdhte` with the MSE-optimal bandwidth ($h$ = 114 cm), extracting bias-corrected treatment effects $\hat{\tau}_{bc}^{ST}$ for each skin tone group. Point estimates show the bootstrap mean; error bars represent bootstrap-based 95% confidence intervals (2.5th and 97.5th percentiles). Brackets display pairwise differences $\Delta = \hat{\tau}_{bc}^{ST_{\text{darker}}} - \hat{\tau}_{bc}^{ST_{\text{lighter}}}$ with one-sided $p$-values computed as the proportion of bootstrap samples where the difference is greater than or equal to zero (testing $H_0 : \Delta \geq 0$ vs. $H_1 : \Delta < 0$). The bootstrap procedure accounts for complex dependence structures and provides robust inference under two-way clustering, which is more conservative than the one-way player-clustered standard errors reported in Figure 4. All bootstrap samples use the same bandwidth, kernel (Epanechnikov), controls (Double LASSO-selected covariates), and weighting scheme (inverse number of within-bandwidth shots per player-match) as the main specification.

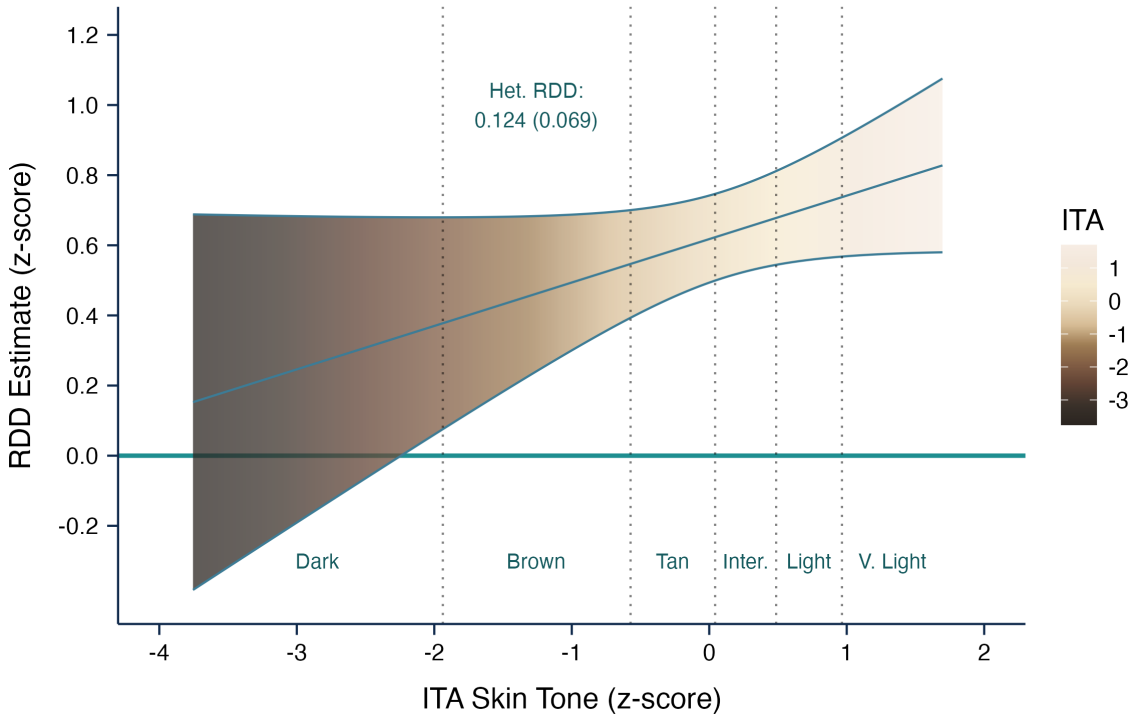



Figure B.4: Continuous Heterogeneous RDD: Goal Premium by ITA Skin Tone (z-score)

*Notes:* This figure displays heterogeneous bias-corrected RDD estimates of the goal premium, the causal effect of scoring a goal on algorithmic post-match player ratings, including the standardized ITA skin tone score as a continuous heterogeneity covariate using the `rdhte` package Calonico et al. (2025*b*). The solid line shows the estimated goal premium as a linear function of ITA skin tone; the shaded region displays 95% confidence intervals based on the bias-corrected variance-covariance matrix. Vertical dotted lines indicate the empirical boundaries of ITA skin tone classes (Very Light, Light, Intermediate, Tan, Brown, Dark). All estimates use the MSE-optimal bandwidth of $h = 114$ cm and an Epanechnikov kernel. The specification includes Double LASSO-selected covariates and are weighted by the inverse number of within-bandwidth shots per player-match.

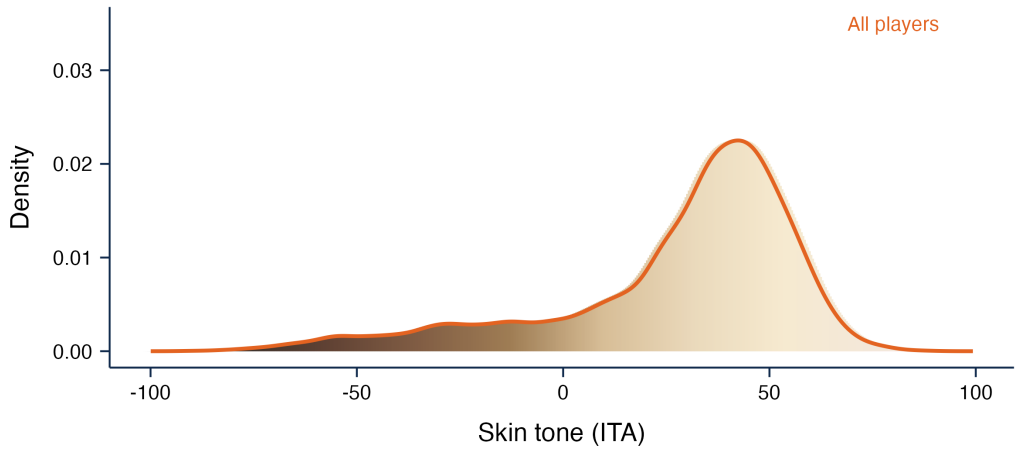


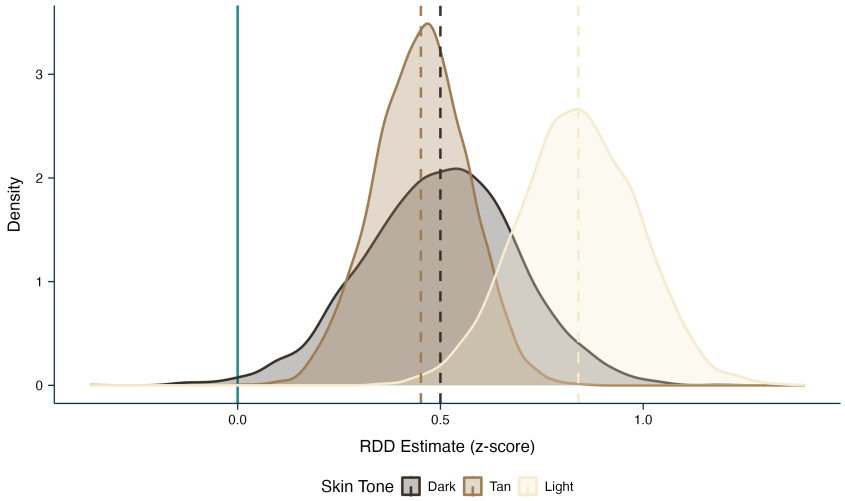

Figure B.5: Skin Tone Distribution: All Players vs. Shooters Sample

*Notes:* This figure compares the distribution of skin tone (measured by the Individual Typology Angle, ITA) between two samples: the universe of all players who received post-match ratings in our dataset (orange line) and the subsample of players who took at least one shot within the RDD bandwidth used in our main analysis (teal line). Skin tone is measured continuously, with lower values indicating darker skin and higher values indicating lighter skin. The distributions are estimated using kernel density estimation with an Epanechnikov kernel. The Kolmogorov-Smirnov test statistic (D) and $p$-value assess whether the two distributions differ significantly. The null hypothesis of distributional equality cannot be rejected, indicating no systematic selection into the shooters sample based on skin tone.

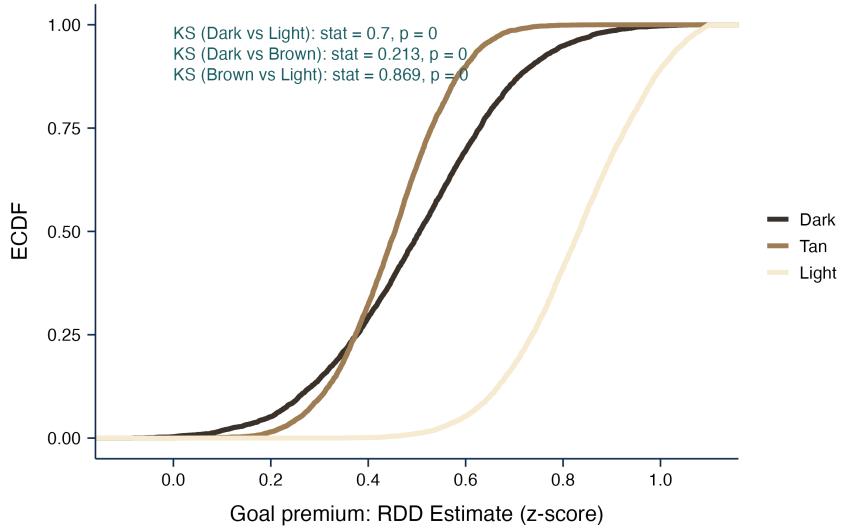
Table B.2: Balance Tests: Pre-Determined Covariates by Skin Tone

| | Dark | | Tan | | Light | |
|---|---|---|---|---|---|---|
| Covariate | Coef. | $q$-value | Coef. | $q$-value | Coef. | $q$-value |
| Match minute | 0.082 | [1.000] | 0.096 | [0.611] | 0.365 | [0.302] |
| Shot Y-coordinate | -0.529 | [0.094] | -0.466 | [0.004] | -0.266 | [1.000] |
| Shot X-coordinate | 0.014 | [1.000] | 0.190 | [0.493] | 0.210 | [1.000] |
| Starting XI | -0.040 | [1.000] | 0.020 | [0.654] | -0.025 | [1.000] |
| Home team | -0.010 | [1.000] | -0.026 | [0.698] | 0.001 | [1.000] |
| Team captain | -0.043 | [1.000] | -0.026 | [0.654] | -0.032 | [1.000] |
| Stadium attendance | 0.285 | [1.000] | -0.191 | [0.493] | 0.084 | [1.000] |
| Penalty kick | 0.017 | [1.000] | 0.057 | [0.493] | 0.010 | [1.000] |
| Defender | 0.014 | [1.000] | -0.052 | [0.556] | -0.059 | [1.000] |
| Header | -0.043 | [1.000] | -0.003 | [1.000] | -0.065 | [1.000] |
| Shirt number 10 | -0.031 | [1.000] | 0.068 | [0.493] | -0.046 | [1.000] |
| Shirt number 16 | -0.008 | [1.000] | 0.015 | [0.600] | -0.013 | [1.000] |
| Free kick | 0.006 | [1.000] | -0.040 | [0.493] | -0.030 | [1.000] |

*Notes:* Each coefficient reports the RDD estimate of the jump at the scoring threshold, estimated separately by skin tone group using `rdhte` (Calonico et al., 2025*b*). Continuous covariates are in z-scores; binary covariates are unmodified. $q$-values apply FDR correction at 5% within group across 13 covariates (Anderson, 2008). The lack of significant discontinuities supports the quasi-random assignment assumption underlying our Difference-in-Discontinuities strategy. Standard errors are robust and clustered at the player level. Bandwidth: 114 cm.




(a) Density Functions




(b) Empirical Cumulative Distribution Functions

Figure B.6: Bootstrap Distributions of Goal Premia by Skin Tone

*Notes:* This figure displays the bootstrap distributions of bias-corrected RDD estimates from 5,000 Bayesian bootstrap samples with two-way cluster resampling. The bootstrap procedure generates observation-level weights by independently drawing Dirichlet weights for player clusters and match clusters, then combining these centered perturbations to preserve correlation both within players (who may take multiple shots) and within matches (where shots share common shocks). Each bootstrap sample re-estimates the heterogeneous RDD using the MSE-optimal bandwidth ($h = 114$ cm), Epanechnikov kernel, Double LASSO-selected controls, and inverse shot-count weighting per player-match pair. Panel (a): Kernel density estimates of the goal premium for each skin tone group. Dashed vertical lines indicate bootstrap means: Dark = 0.50, Tan = 0.46, Light = 0.85 (z-scores). The solid vertical line at 0 marks the null hypothesis of no goal premium. Dark and Tan distributions exhibit substantial overlap, with similar modes around 0.5, while the Light distribution is clearly separated and shifted rightward with mode near 0.8. Panel (b): Empirical cumulative distribution functions (ECDFs) showing the proportion of bootstrap samples below each value. The vertical separation between curves indicates first-order stochastic dominance: at every quantile, Light-skinned players' distribution yields higher goal premia than the other groups. Dark and Tan ECDFs track closely until approximately the 40th percentile. Kolmogorov-Smirnov tests strongly reject equality of distributions: Dark vs. Light (KS = 0.70, $p < 0.001$), Tan vs. Light (KS = 0.87, $p < 0.001$), and Dark vs. Tan (KS = 0.21, $p < 0.001$). The distributions provide visual confirmation of systematic colorism in algorithmic evaluations, with the strongest discrimination between Light-skinned players and darker-skinned groups.

Table B.3: Within-Match Performance Metrics: RDD Estimates by Skin Tone

| | Dark | | Tan | | Light | |
|---|---|---|---|---|---|---|
| Performance Metric | Coef. | $q$-value | Coef. | $q$-value | Coef. | $q$-value |
| FotMob rating | 0.479 | [0.047] | 0.516 | [0.000] | 0.822 | [0.000] |
| Minutes played | 0.075 | [1.000] | 0.013 | [1.000] | -0.057 | [1.000] |
| Accurate passes | 0.008 | [1.000] | -0.038 | [1.000] | 0.104 | [1.000] |
| Accurate long balls | -0.506 | [0.066] | 0.114 | [0.776] | -0.199 | [0.892] |
| Recoveries | -0.035 | [1.000] | 0.084 | [0.810] | -0.018 | [1.000] |
| Touches | 0.129 | [1.000] | -0.029 | [1.000] | 0.044 | [1.000] |
| Goals | 1.282 | [0.000] | 1.276 | [0.000] | 1.316 | [0.000] |
| Assists | -0.105 | [1.000] | -0.091 | [0.781] | 0.059 | [1.000] |
| Total shots | 0.059 | [1.000] | -0.265 | [0.120] | -0.205 | [0.657] |
| Chances created | -0.065 | [1.000] | -0.134 | [0.660] | 0.197 | [0.892] |
| Passes into final third | -0.027 | [1.000] | -0.125 | [0.660] | 0.125 | [1.000] |
| Accurate crosses | -0.042 | [1.000] | 0.075 | [0.891] | -0.025 | [1.000] |
| Offsides | -0.055 | [1.000] | 0.094 | [0.781] | -0.141 | [0.892] |
| Dispossessed | -0.248 | [1.000] | 0.081 | [0.810] | -0.333 | [0.079] |
| Tackles won | 0.148 | [1.000] | 0.179 | [0.432] | 0.170 | [0.892] |
| Clearances | 0.051 | [1.000] | 0.197 | [0.385] | 0.157 | [0.892] |
| Headed clearances | 0.046 | [1.000] | 0.257 | [0.247] | 0.168 | [0.892] |
| Interceptions | -0.279 | [0.608] | 0.123 | [0.776] | 0.049 | [1.000] |
| Ground duels won | -0.072 | [1.000] | -0.111 | [0.776] | 0.064 | [1.000] |
| Aerial duels won | 0.504 | [0.147] | 0.212 | [0.314] | -0.067 | [1.000] |
| Was fouled | 0.008 | [1.000] | 0.039 | [1.000] | -0.172 | [0.892] |
| Fouls committed | 0.124 | [1.000] | 0.253 | [0.247] | 0.030 | [1.000] |
| Clearances off line | 0.008 | [1.000] | -0.001 | [1.000] | 0.003 | [1.000] |
| Blocks | 0.173 | [1.000] | 0.213 | [0.660] | -0.040 | [1.000] |
| Dribbled past | -0.030 | [1.000] | -0.078 | [0.883] | 0.264 | [0.631] |
| Corners | -0.111 | [1.000] | 0.044 | [1.000] | 0.055 | [1.000] |
| Blocked shots | 0.216 | [1.000] | -0.049 | [1.000] | -0.160 | [0.892] |
| Successful dribbles | -0.059 | [1.000] | -0.246 | [0.299] | 0.158 | [0.892] |
| Big chances missed | -0.804 | [0.000] | -0.697 | [0.000] | -0.411 | [0.020] |
| Error led to goal | 0.033 | [1.000] | -0.005 | [0.441] | -0.008 | [0.528] |
| Conceded penalty | -0.011 | [1.000] | -0.012 | [0.314] | 0.004 | [1.000] |
| Penalties won | -0.114 | [1.000] | -0.107 | [0.776] | 0.134 | [0.892] |
| Last man tackle | -0.134 | [1.000] | -0.018 | [1.000] | -0.065 | [0.892] |
| Missed penalty | -0.031 | [0.061] | -0.039 | [0.018] | -0.029 | [0.009] |
| Own goal | 0.006 | [1.000] | -0.008 | [0.142] | -0.003 | [0.892] |

*Notes:* This table tests whether scoring a goal causes discontinuous changes in within-match performance metrics, estimated separately for each skin tone group. Each coefficient represents the bias-corrected RDD estimate of the jump at the scoring threshold. All metrics except binary indicators are standardized (z-scores). $q$-values are computed using group-specific FDR correction (Anderson 2008), controlling the false discovery rate at 10% within each skin tone group across 35 performance metrics. Metrics with significant discontinuities in any group are potential "bad controls" that reflect post-treatment outcomes and should be included as covariates when testing for discrimination in post-match ratings. Robust standard errors clustered at player level. Bandwidth: 114 cm.

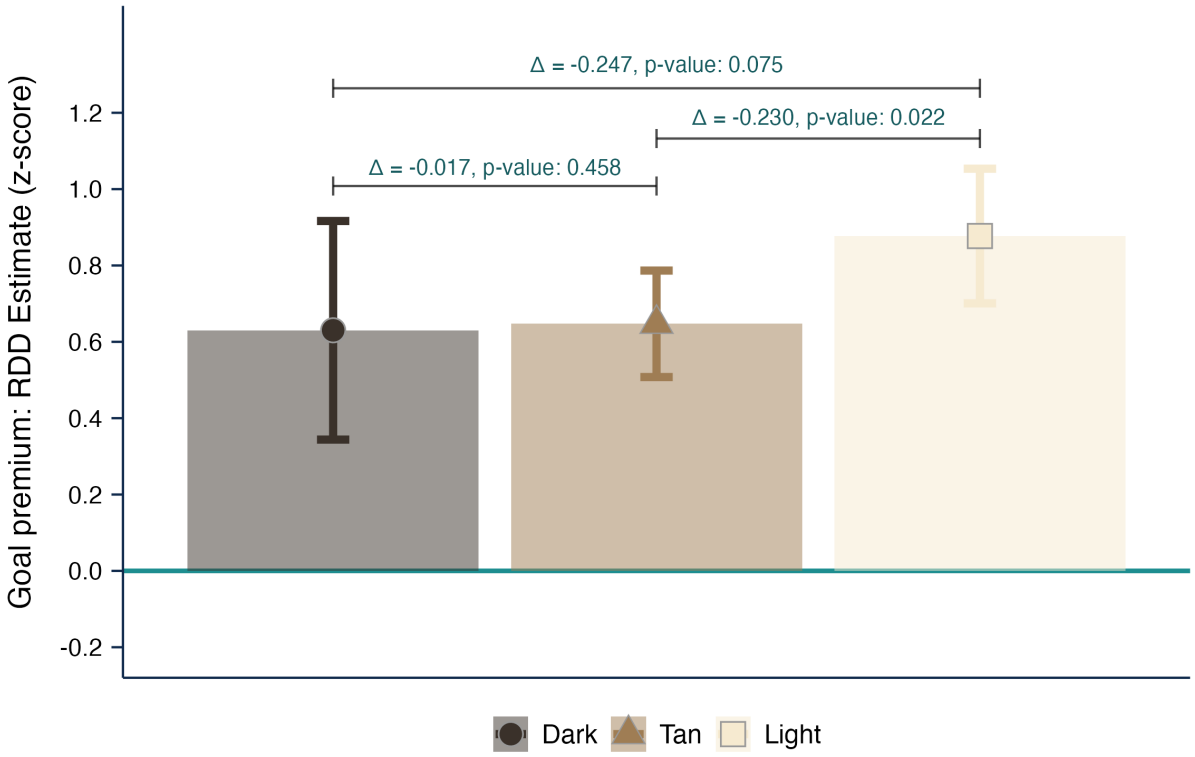



Figure B.7: Difference-in-Discontinuities Controlling for Post-Treatment Performance Metrics

*Notes:* This figure displays group-specific RDD estimates of the goal premium (in z-scores), controlling for three non-mechanical post-treatment variables: accurate long balls, dispossessions, and additional goals scored. Estimates are obtained using the `rdhte` package (Calonico et al., 2025*b*), with 95% confidence intervals clustered at the player level. Brackets show pairwise differences ($\Delta$) and one-sided $p$-values. All models use the MSE-optimal bandwidth ($h = 114$ cm), an Epanechnikov kernel, and include Double LASSO-selected covariates, with weights for the number of shots per player-match.

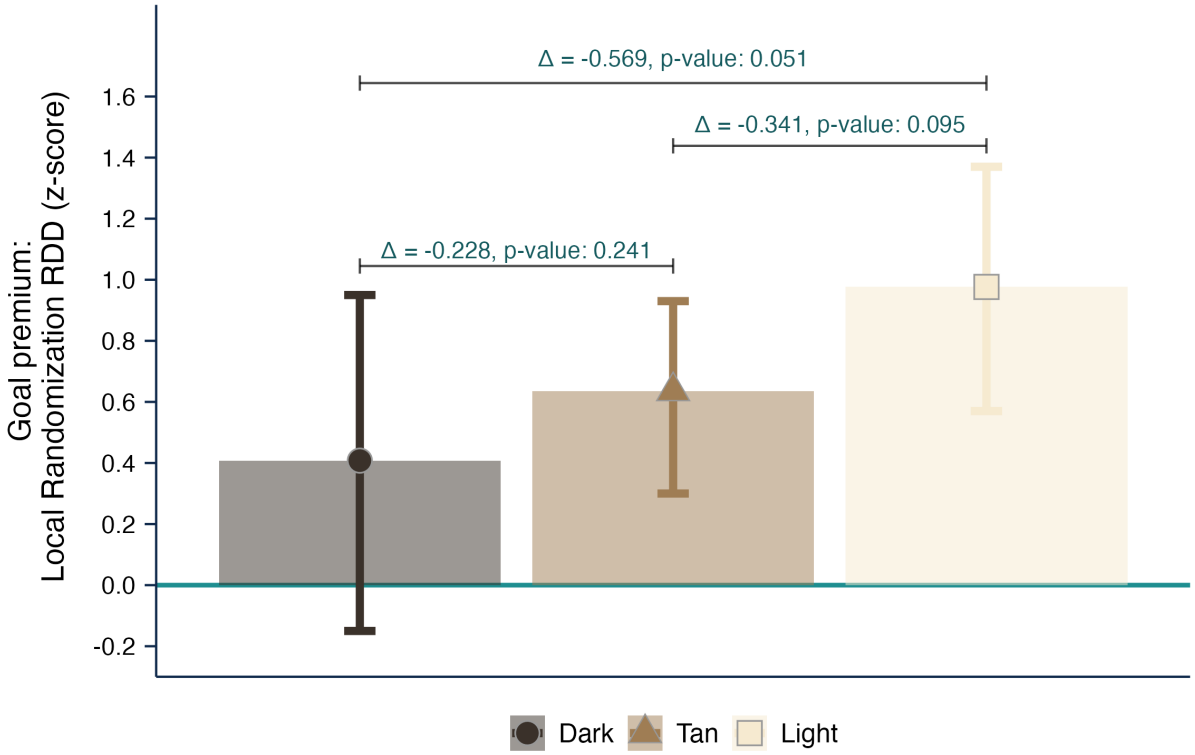



Figure B.8: Local Randomization RDD: Goal Premium by Skin Tone

*Notes:* This figure displays Local Randomization RDD estimates of the goal premium—the causal effect of scoring a goal on algorithmic post-match player ratings—for each skin tone category. Unlike the continuity-based RDD in Figure 4, this approach assumes quasi-random assignment within a narrow window around the threshold rather than relying on smoothness of potential outcomes. The y-axis shows goal premium estimates in z-scores. Point estimates are computed using randomization inference within a ś15 cm window (approximately the radius of a regulation football), with 95% confidence intervals constructed via permutation tests that account for mass points in the running variable (Cattaneo et al., 2015). The narrow window contains shots most plausibly affected by quasi-random variation in ball placement: those missing or scoring by just a few centimeters where precision in striking is paramount. Brackets denote pairwise differences $\Delta = \hat{\tau}^{ST_{\text{darker}}} - \hat{\tau}^{ST_{\text{lighter}}}$, with one-sided $p$-values computed from the distribution of differences under the null hypothesis of no discrimination ($H_0$: $\Delta \geq 0$ vs. $H_1$: $\Delta < 0$). Standard errors for pairwise comparisons are computed as $\sqrt{SE^2_{\text{group 1}} + SE^2_{\text{group 2}}}$, where group-specific SEs are derived from the 95% confidence intervals: $SE \approx (UCI - LCI)/(2 \times 1.96)$. The Local Randomization framework provides a robustness check on our main continuity-based estimates by employing a more restrictive identification assumption (randomization rather than continuity) but avoiding functional form assumptions about the conditional expectation of potential outcomes.

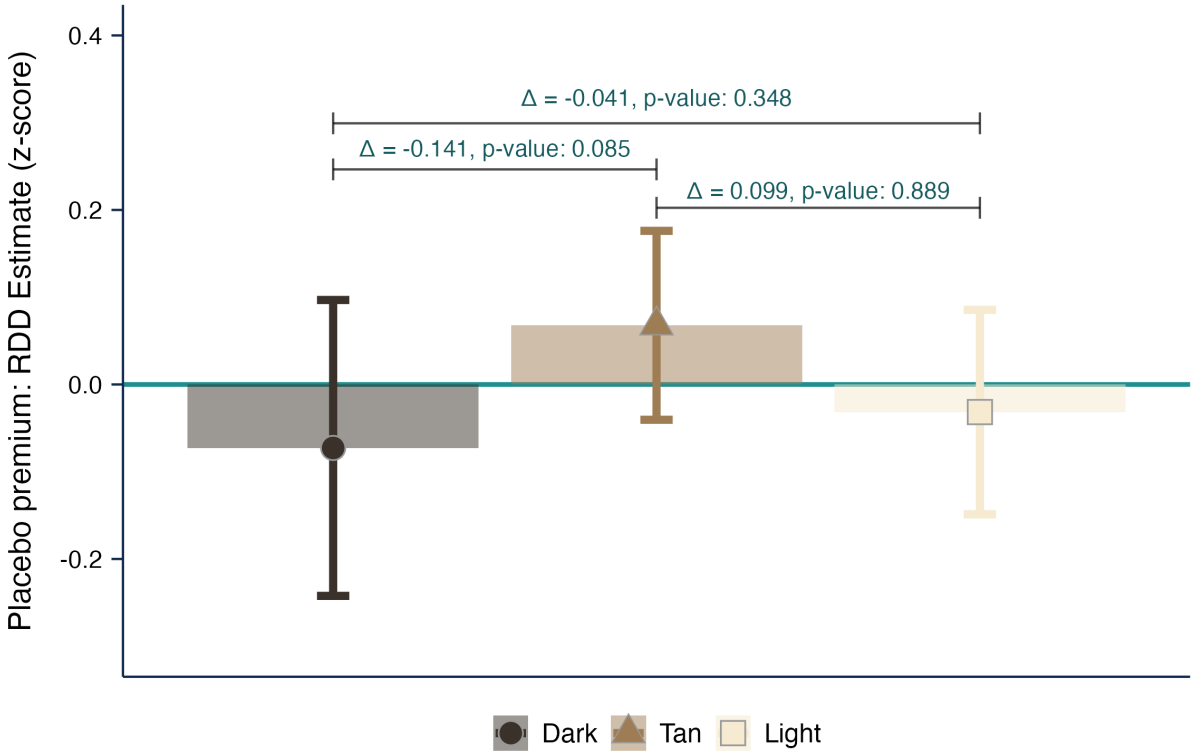



Figure B.9: Placebo Test: Artificial Threshold Among Scored Goals

*Notes:* This figure presents a placebo test to assess whether the observed colorism gradient reflects genuine discrimination at the scoring threshold or spurious differences in rating trajectories by skin tone. We estimate an artificial RDD centered at the mean distance of scored goals from the goal frame (approximately 26 cm inside the goal line) using the same MSE-optimal bandwidth (114 cm), controls (Double LASSO-selected covariates), weighting scheme (inverse shots per player-match), and specification as our main analysis. The y-axis shows placebo "premium" estimates in z-scores—the estimated rating discontinuity at this artificial threshold where no actual change in shot outcome occurs. Under the null hypothesis that our main findings are not driven by arbitrary differences in rating levels or trends across skin tone groups, we should observe no systematic discontinuities at this placebo cutoff. Point estimates are shown with 95% confidence intervals (robust to heteroskedasticity, clustered at player level). Brackets denote pairwise differences with one-sided $p$-values testing for discrimination. All three groups exhibit estimates close to zero: Dark ($-0.07$ [$-0.25$, 0.11]), Tan (0.07 [$-0.05$, 0.18]), Light ($-0.03$ [$-0.12$, 0.07]). Pairwise differences are substantively small and statistically insignificant: Dark vs. Light ($\Delta = -0.041$, $p = 0.348$), Tan vs. Light ($\Delta = 0.099$, $p = 0.889$), Dark vs. Tan ($\Delta = -0.141$, $p = 0.085$). The absence of rating differences at this artificial threshold—where crossing the cutoff carries no performance implications—demonstrates that the colorism gradient in our main results emerges specifically at the scoring threshold where algorithmic evaluations respond to goal outcomes, rather than reflecting pre-existing differences in how ratings evolve with shot placement across skin tone groups.

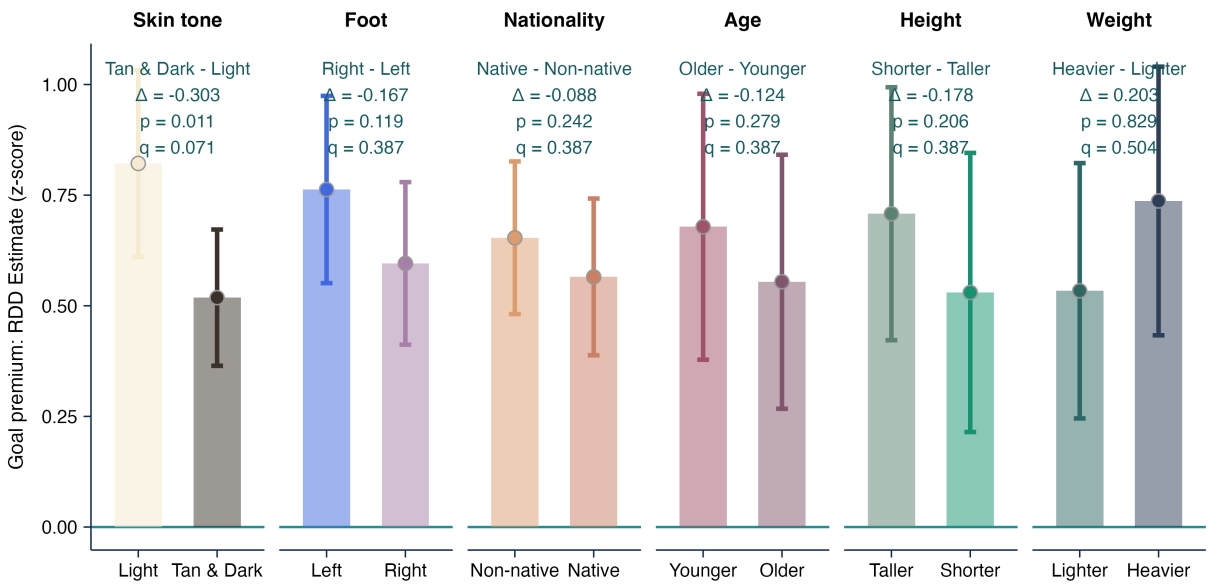



Figure B.10: Heterogeneity Analysis: Goal Premia Across Player Dimensions

*Notes:* This figure presents heterogeneous RDD estimates of the goal premium across six player dimensions to assess whether discrimination extends beyond skin tone. Each panel displays bias-corrected estimates for two groups within a dimension, using identical specifications to our main analysis: MSE-optimal bandwidth (114 cm), Epanechnikov kernel, Double LASSO-selected covariates, player-level clustering, and inverse shot-count weighting. For age, height, and weight analyses, we merge player demographics from the SoFIFA database using fuzzy name and nationality matching, then create binary indicators for above/below-mean values. Point estimates show the goal premium in z-scores with 95% confidence intervals. Annotations report pairwise differences ($\Delta$) with one-sided $p$-values and FDR-corrected $q$-values computed across all six comparisons using the Storey (2002) method at the 10% level.

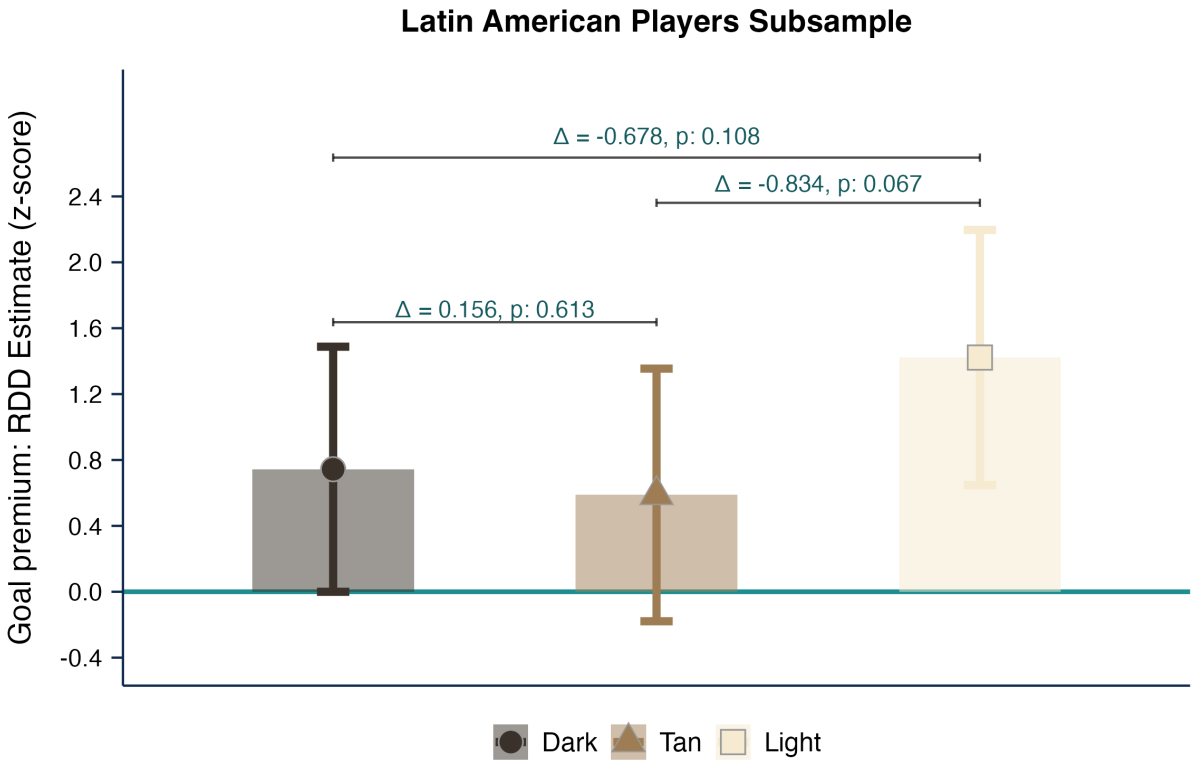
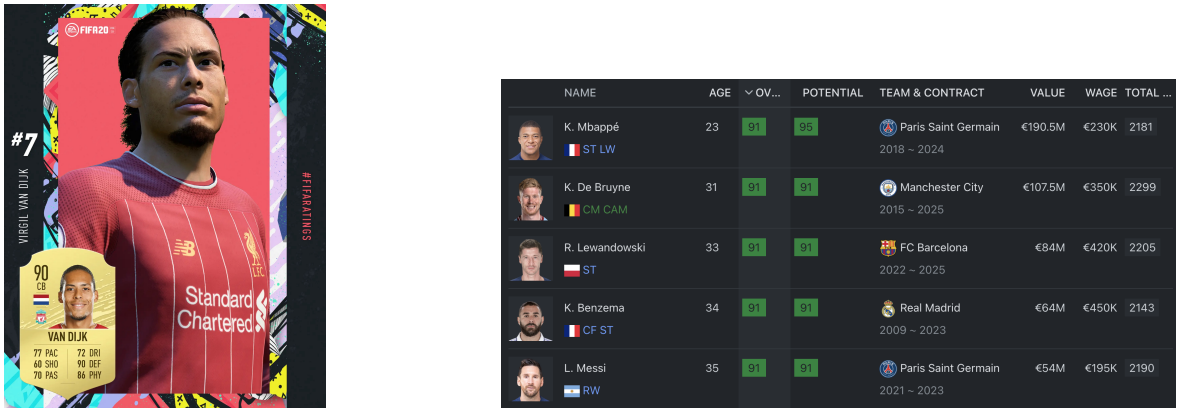



Figure B.11: RDD and DiDC Estimates: Latin America Subsample

*Notes:* This figure reports RDD and DiDC estimates for the sample of players from Brazil, Argentina, Uruguay, Colombia, and Mexico. These countries are overwhelmingly Christian, allowing us to hold religion relatively constant while preserving variation in skin tone. The persistence-and even amplification-of racial disparities in this subsample suggests that religion is not the primary driver of the results.




(a) Rating card of Virgil van Dijk

(b) Snapshot of Sofifa.com

Figure B.12: EA Sports FIFA: Ratings

*Notes*: These are fan- driven ratings. Overall rating is determined by six key statistics, and international recognition. Panel (b) provides a snapshot of Sofifa.com with extensive information on players.

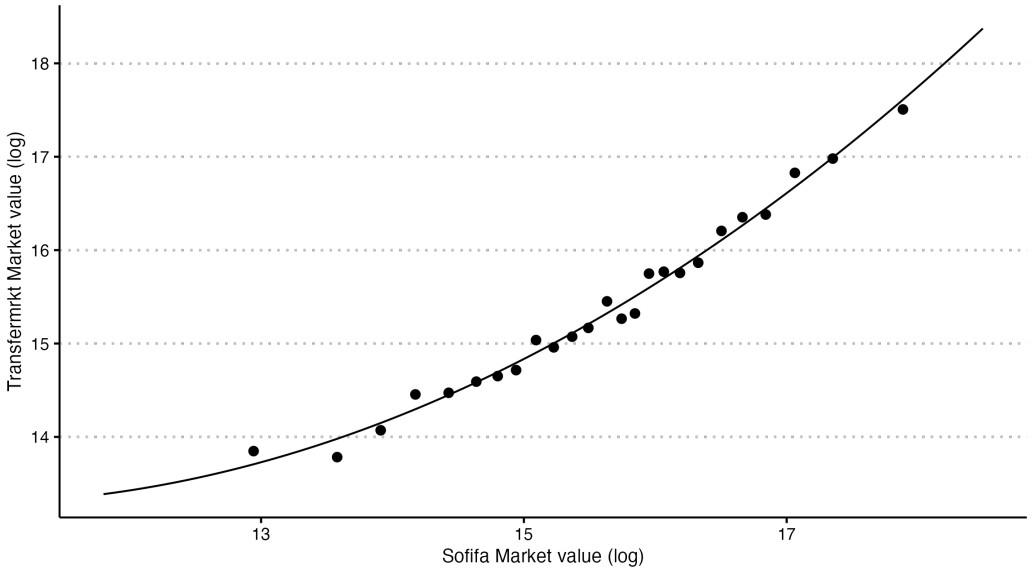



Figure B.13: Market values: Sofifa.com and Transfermarkt

*Notes*: Logged player market values from Sofifa.com and FotMob lie on the x- and y-axis, respectively. The binned scatterplot demonstrates the correlation of market values (logged) reported on Sofifa.com and Transfermarkt. We compare players' valuations between the two sources by implementing fuzzy matching based on player name and nationality. The match distance is up to 0.1, i.e., matches based on 90% or above similarity.

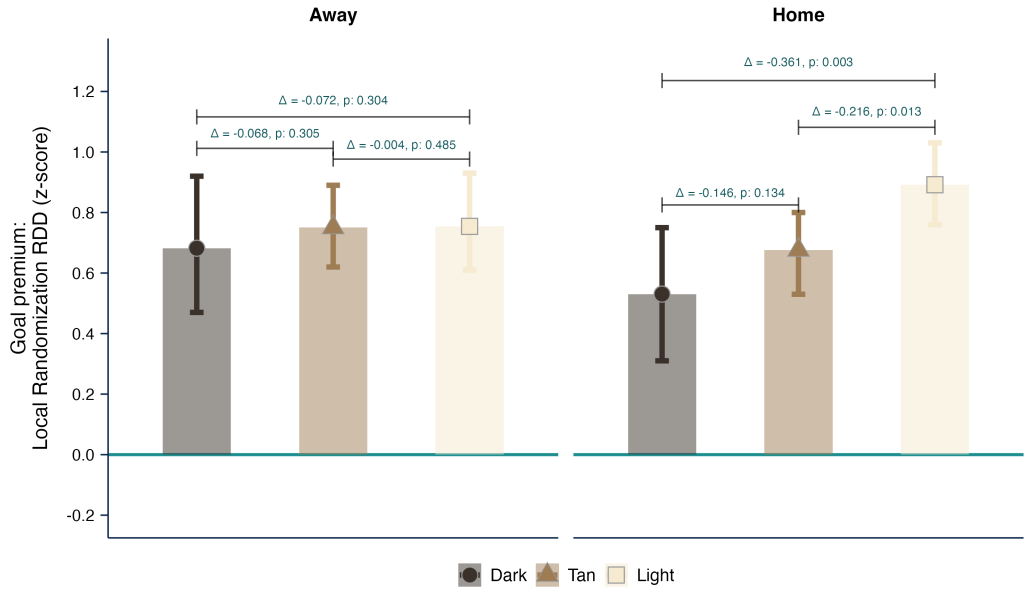



Figure B.14: RDD and Diff-in-Disc Estimates: Away vs. Home Matches.

*Notes:* The left (right) panel shows the estimated coefficients for the subsample of matches where the player's team played an away (home) game.

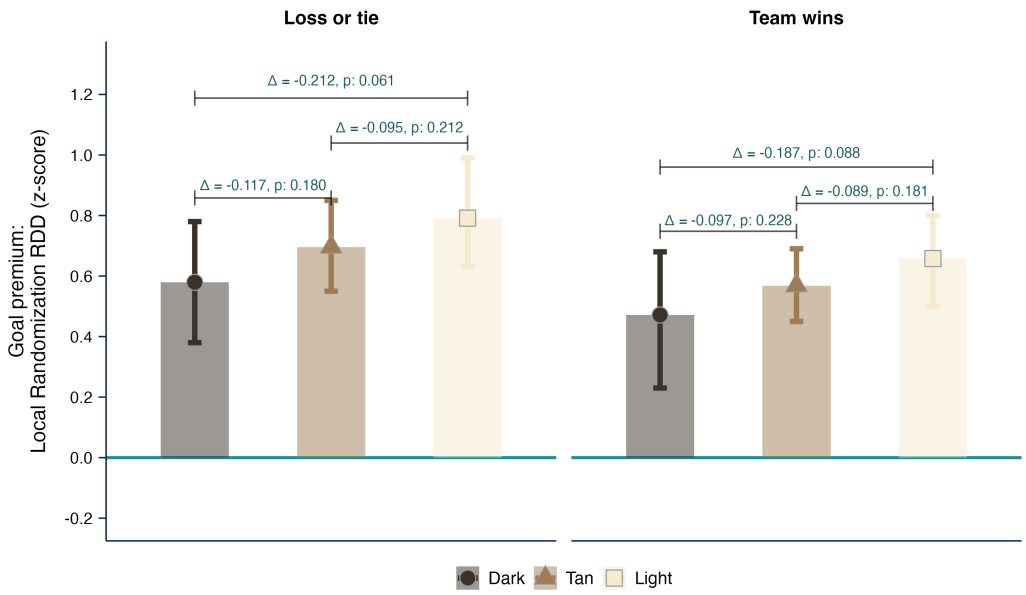



Figure B.15: RDD and Diff-in-Disc Estimates: Match Result

*Notes:*The figure reports RDD goal-scoring premiums (in standard deviations) by skin tone group, estimated separately for matches that end in a win (right panel) and those that end in a loss or tie (left panel). Estimates are derived using the Local Randomization RDD framework. In matches won by the player's team, we observe large and statistically significant disparities in treatment effects, with dark- and brown-skinned players receiving smaller post-match rating premiums than light-skinned players. In contrast, no significant differences are observed when the match ends in a loss or tie. P-values are ...