SKIN TONE PENALTIES: QUASI-EXPERIMENTAL EVIDENCE ON COLORISM IN FOOTBALL

L. GUILLERMO WOO-MORA
®
DONIA KAMEL

WORKING PAPER N°2026/02

FEBRUARY 2026

WORLD INEQUALITY LAB

Skin Tone Penalties:

Quasi-Experimental Evidence on Colorism in Football[*]

L. Guillermo Woo-Mora[†] ⓡ Donia Kamel[‡]

First version: August 2023

This version: February 2026

Abstract

We provide causal evidence of skin tone discrimination using professional football (soccer) as a natural laboratory. Leveraging a computer-vision measure of skin tone and quasi-random variation in shot outcomes near the goal frame, we implement a Difference-in-Discontinuities design comparing narrowly scored goals to narrowly missed attempts. We find that Light-skinned players receive significantly larger boosts in post-match ratings than Tan- and Dark-skinned peers for identical actions. These disparities appear in both algorithmic and human-assigned evaluations and are concentrated in the subjective component of ratings. Season-level analyses reveal that biased evaluations translate into lower market valuations for darker-skinned players, despite equivalent performance. Evaluative bias, rather than differential treatment in contracts, emerges as a key driver of economic inequality in this high-information labor market. Our findings show how skin color discrimination can persist even in environments with transparent outcomes and extensive performance data.

JEL Codes: D63, F22, J15, J71, Z20, Z22

Keywords: Race, Colorism, Discrimination, Algorithms, Football

How to cite this working paper: Woo-Mora, L. G.*, Kamel, D., Skin Tone Penalties: Quasi-Experimental Evidence on Colorism in Football, World Inequality Lab Working Paper 2026/02

[*] Author order was randomized using the AEA Author Randomization Tool (code: _6oRQ8bv0JT2). 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, Joshua Angrist, Sebastian Bustos, Enzo Brox, Jacob Conway, Patricia Cortés, Christopher Dawes, Ellora Derenoncourt, Glen Ellison, Paolo Falco, Julio Garbers, 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, REP-LAC, 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. Refine.ink was used to proofread the paper for consistency and clarity, all remaining errors are our own.

[†] Paris School of Economics and EHESS, guillermo.woo-mora@psemail.eu

[‡] Paris School of Economics, donia.kamel@psemail.eu

1 Introduction

Skin tone disparities are well documented in representation (Adukia et al., 2023) and in outcomes such as income, human capital, and intergenerational mobility (Woo-Mora, 2026). Yet we know far less about whether these gaps reflect *direct discrimination* based on skin tone—known as colorism—as opposed to differences in underlying opportunities or cumulative disadvantage. High-profile performance settings, where success and failure are immediately observable, provide rare windows into this distinction.

In July 2021, three Black players missed penalties in the Euro final and faced immediate racist abuse (Taylor et al., 2021). A year later, a White player missed a decisive World Cup penalty without comparable backlash. Such incidents illustrate a core definition of discrimination in economics: members of minority groups are treated less favorably than otherwise identical majority-group members in similar circumstances (Bertrand and Duflo, 2017). Empirically isolating this unequal treatment remains a persistent challenge.

Empirical research on discrimination has long confronted a fundamental trade-off. Audit studies randomize identity signals but cannot observe true productivity (Bertrand and Mullainathan, 2004). Observational studies measure productivity but face selection into who is observed and how performance is recorded (Guryan and Charles, 2013; Lang and Spitzer, 2020). Few settings allow researchers to hold output constant while identifying how evaluators translate identical performance into differential rewards. This observability problem is particularly acute for colorism because skin tone is rarely captured in administrative data and, when recorded, often reflects subjective enumerator assessments rather than objective phenotypic variation (Dixon and Telles, 2017; Monk, 2021).

This paper provides causal evidence of skin tone discrimination in a setting where productivity is fully observed, directly comparable, and subject to quasi-random variation in measured success. We study post-match player ratings in professional mens football (soccer), leveraging the mechanical similarity of shots that narrowly score versus narrowly miss the goal. Specifically, we implement a regression discontinuity design based on the Euclidean distance from the balls final position to the edges of the goal frame—the scoring boundary. Within a narrow window of approximately one ball circumference, whether a shot becomes a goal depends on micro-variation in trajectory, spin, or bounce that players cannot precisely control. This design enables us to estimate how algorithmic and human evaluators reward identical on-field actions, and whether those rewards vary systematically by skin tone.

Our empirical strategy proceeds in three steps. First, we follow Adukia et al. (2023) and develop a computer-vision algorithm to construct a continuous, pixel-based measure of skin tone from high-resolution player headshots. The algorithm isolates facial regions, identifies dominant skin-color composition, and computes the Individual Typology Angle (ITA)—a dermatological metric combining lightness and undertone that closely aligns with human visual perception (Chardon et al., 1991). Unlike categorical race, which recent work shows is endogenously constructed along socioeconomic lines (Adukia et al., 2025), ITA-based skin tone is a stable physical trait—measured independently of outcomes or socioeconomic status—that provides a consistent basis for cross-sectional comparisons.

Second, we combine this measure with geolocated shot data covering over 76,000 attempts across seven major European leagues. Using bias-corrected regression discontinuity methods (Calonico et al., 2014), we estimate the causal effect of scoring a marginal goal on post-match player ratings. We then implement a Difference-in-Discontinuities design (Grembi et al., 2016; Picchetti et al., 2024), estimating this local effect separately for Light, Tan, and Dark players and formally comparing them using recent advances in heterogeneous treatment effect estimation for RD designs (Calonico et al., 2025*a*). Our identifying assumption is that, within a narrow bandwidth around the goal line, shot outcomes are as good as randomly assigned—an assumption we assess using density tests, covariate balance checks, and event-study analyses.

Third, we examine both mechanisms and downstream consequences. To understand the origins of bias, we replicate our design using journalist-assigned post-match ratings and show that human evaluators exhibit nearly identical skin tone gradients as the algorithm, suggesting that algorithmic bias likely reflects inherited human preferences. To assess economic consequences, we link player-season ratings to data on wages and market valuations. Using distance-weighted near-goals as a quasi-random shock to ratings, we estimate the causal effect of marginal rating improvements on labor market outcomes and test whether these returns differ by skin tone.

Our findings reveal large and robust disparities in how marginal goals are rewarded. In the pooled sample, narrowly scoring rather than narrowly missing increases a players rating by approximately 0.57 standard deviations. However, Light-skinned players receive a premium of 0.81 standard deviations, compared with 0.50 for Tan players and 0.45 for Dark players. The Light–Dark gap of 0.36 SD ($p = 0.037$) represents a 44% difference in causal recognition for identical performance. When we replace categorical groupings with our continuous ITA measure, we find that each one standard deviation increase in skin lightness corresponds to a 0.13 SD larger rating reward for scoring. Moving from very Dark to very Light skin tone—a four-SD shift—yields a roughly 0.52 SD increase in recognition for the same performance event.

These disparities are highly robust. Pre-shot and pre-match covariates selected via Double LASSO (Belloni et al., 2014) are balanced across the scoring threshold, and the results persist when conditioning on them. While some within-match performance metrics shift at the threshold and may differ by skin tone, the disparities remain stable even when conditioning on these outcomes. The gradient holds under narrow-window local randomization inference (Cattaneo et al., 2015), does not appear at placebo thresholds, and is not driven by selection into the sample of shooters or differences in shot difficulty. We replicate the pattern using an alternative parametric strategy that compares goals to goalkeeper saves within spatially matched locations. Importantly, no other observable player characteristic—including nationality, age, height, weight, or preferred foot—exhibits comparably large or statistically significant heterogeneity in the goal premium after adjusting for multiple testing. The gradient also persists across match contexts (home vs. away, wins vs. losses), further confirming its robustness.

To understand the origins of these disparities, we decompose algorithmic ratings into two components: an objective performance score predicted by over 40 match statistics using a high-accuracy machine learning model, and a residual component capturing unexplained variation. While predicted ratings show no systematic skin tone gradient, the residuals display a clear and monotonic pattern: Light-skinned players receive the highest recognition, followed by Tan, then Dark-skinned players. Only 33% of the Light-Dark disparity is attributable to observable performance, with the remaining 67% driven by residual differences. This pattern suggests that algorithmic ratings reflect not only what players do on the pitch, but also subjective or unmeasured factors correlated with skin tone.

We then test whether this residual bias originates in human evaluations. Replicating our design with journalist-assigned ratings from three top European leagues, we find a nearly identical skin tone gradient, indicating that human evaluators reward marginal goals unequally by appearance. Linking journalist ratings to algorithmic outputs for the same players and matches, we show that journalist bias—defined as deviation from predicted performance—strongly predicts algorithmic residuals, particularly for lighter-skinned players. These findings provide direct evidence that algorithms inherit and amplify biases embedded in human judgment, formalizing them into scalable and reproducible disparities.

Do biased evaluations translate into economic inequality? Using player-season panel data and quasi-random match-level variation in close shots, we show that unequal recognition for identical performance has real labor market consequences. Our match-level DiDC estimates indicate that lighter-skinned players receive systematically greater recognition for marginal on-field success. Assuming these evaluation gaps accumulate into season-average ratings, we show that they translate into significantly higher market valuations in the following season. Crucially, we find no evidence of direct discrimination in the valuation process: conditional on ratings, market values do not differ by skin tone. We also find no evidence of statistical discrimination: the semi-elasticity of market value with respect to ratings is statistically identical across the skin tone distribution, implying that ratings are weighted equally regardless of player appearance.

This pattern reflects systemic discrimination (Bohren, Hull and Imas, 2025): downstream inequality arises entirely from biased evaluations, not differential market responses. The effects are economically meaningful. Our main DiDC estimates show that a four-standard-deviation shift in skin tone yields a 0.52SD increase in match-level recognition for identical performance. Regressing season-average ratings on skin tone reveals gaps of up to 0.25SD under minimal controls—implying a 2-4% valuation difference, given a 0.161 semi-elasticity of market value with respect to ratings. By contrast, wages—constrained by contracts and institutional frictions—do not respond to marginal performance updates. Distorted performance signals, shaped by both algorithmic and human evaluators, are thus faithfully transmitted through labor markets and amplified via crowd-based valuation systems.

The main contribution of this paper is to advance empirical research on discrimination by providing quasi-experimental evidence that differential treatment based on visually salient phenotypic traits, namely skin color, persists even when productivity is perfectly observable (Becker, 1957; Phelps, 1972; Arrow, 1973; Bertrand and Duflo, 2017). While a large literature documents disparities across racial groups in labor markets, education, credit access, and criminal justice (Bertrand and Mullainathan, 2004; Arceo-Gomez and Campos-Vazquez, 2014; Arnold et al., 2018; Lang and Spitzer, 2020; Arnold et al., 2022; Kline et al., 2022; Derenoncourt et al., 2023; Bohren, Hull and Imas, 2025), many empirical strategies necessarily bundle multiple identity dimensions. Methodologically, our approach is closely related to influential work in the gender literature that exploits settings with largely observable performance to study differential evaluation (Goldin and

Rouse, 2000; Sarsons, 2022). Our setting isolates skin tone under near-experimental conditions, allowing us to measure differential evaluation while holding output constant. Because evaluators observe rich performance metrics, we can also distinguish taste-based discrimination from statistical discrimination and other related mechanisms (Bohren, Haggag, Imas and Pope, 2025).

Second, we contribute to the literature on colorism and the economic salience of phenotypic variation. Prior work shows that skin tone predicts wages, wealth, mobility, and educational attainment even within racial groups (Goldsmith et al., 2006, 2007; Monk, 2021; Adukia et al., 2023). Recent research conceptualizes race as a socially constructed bundle of attributes whose salience varies across contexts (Sen and Wasow, 2016; Davenport, 2020; Rose, 2023). Adukia et al. (2023) use computer vision and natural language processing to document phenotypic representation in childrens books, showing that darker-skinned characters are underrepresented and receive less symbolic recognition, even conditional on race and gender. We extend this work from representation to evaluation: using high-resolution skin tone measurement and quasi-random variation in performance, we show that darker-skinned individuals receive systematically less credit for identical success. Our findings align with recent observational evidence from Woo-Mora (2026), who document skin tone gradients in income, education, and mobility across 25 Latin American countries, even *within* racial categories. The alignment between these correlational patterns and our causal estimates underscores the structural persistence of colorism in both symbolic and material domains.

Third, we contribute to the literature on algorithmic fairness and the distributional consequences of predictive systems. Prior work shows that algorithms can reproduce or amplify group disparities when trained on biased data (Kleinberg et al., 2018; Lambrecht and Tucker, 2019; Obermeyer et al., 2019; Arnold et al., 2021). Unlike settings with selective outcome observability (Arnold et al., 2025), our context allows us to study algorithmic bias under near-experimental conditions where output is standardized and fully observed. We find that algorithmic evaluations exhibit a monotonic colorism gradient that mirrors journalist-based ratings, providing quasi-experimental evidence of algorithmic colorism in a real labor market setting.

Finally, we contribute to a growing literature using football (soccer), and more broadly sports, as a natural laboratory for economic questions (Palacios-Huerta, 2003; Kleven et al., 2013; Depetris-Chauvin et al., 2020; Alrababa'H et al., 2021; Palacios-Huerta, 2025). Prior work documents discrimination in refereeing, labor markets, and disciplinary sanctions (Price and Wolfers, 2010; Reilly and Witt, 2011; Gallo et al., 2013; Szymanski, 2000; Deschamps and De Sousa, 2021; Caselli et al., 2023; Faltings et al., 2023). Three recent studies examine evaluator responses to player identity in football: Principe and van Ours (2022) find lower newspaper ratings for Black players conditional on performance; Alrababah et al. (2024) find limited evidence of harsher punishment for underperformance; and Colombe et al. (2025) document overcorrection under media scrutiny. We differ by exploiting quasi-random variation in narrowly scored versus narrowly missed shots, enabling within-match causal identification of how evaluators reward identical performance across skin tone groups.

The paper proceeds as follows. Section 2 introduces the institutional context. Section 3 describes the data and skin tone measurement. Section 4 outlines the empirical strategy. Section 5 presents main findings. Section 6 explores mechanisms. Section 7 examines labor market consequences.

Section 8 concludes.

2 Setting

Professional football offers a uniquely powerful context for studying discrimination in performance evaluation. Individual contributions are observable, quantifiable, and systematically assessed by multiple actors. During each 90-minute match, players perform hundreds of discrete actions—passes, shots, tackles, dribbles—which are captured by optical tracking systems and classified by trained coders. These granular event logs feed into post-match player ratings: single-number scores (typically 1–10) summarizing overall performance.

Two parallel systems generate these ratings. *Algorithmic ratings*, distributed by platforms such as FotMob, WhoScored, and Sofascore, apply proprietary statistical rules to transform event-level data into performance scores. FotMob, a Norwegian platform with over 20 million monthly users (Pathak, 2025), has emerged as one of the most widely consulted sources. According to Norwegian media, professional scouts routinely rely on these ratings when evaluating players (Wilhelmsen, 2025).

By contrast, *journalist ratings*, published by legacy outlets such as *Sky Sports* or *LÉquipe*, are based on expert human judgment. After each match, journalists assign 1–10 scores to every player, typically within hours of the game ending. These assessments often provoke intense public scrutiny. For instance, during Frances 4-2 victory in the 2018 World Cup final, *LÉquipe* gave NGolo Kanté a 3/10, prompting national debate over the subjectivity and opacity of such evaluations. As reported by James and Harpur (2025), these ratings are typically assigned by a small team of in-stadium journalists without access to replays or advanced metrics. Despite their subjective nature, journalist ratings remain deeply embedded in football discourse and influence both media narratives and public perceptions.

Though distinct in method, both algorithmic and human ratings evaluate the same performances and circulate widely within the football ecosystem. While these scores carry no formal financial reward, they are far from symbolic. They act as *informal yet influential labor market signals.* High ratings can raise a players visibility, enhance contract negotiations, or influence scouting and recruitment decisions.

Reputational capital formed through ratings also feeds into crowd-sourced platforms like Transfermarkt, which now serve as de facto benchmarks in the football economy. Originally launched as a fan project, Transfermarkt has grown into a global database estimating the market value of more than 800,000 players. Though based on volunteer contributions moderated by editors, its valuations are widely cited in media, club financial documents, and even legal proceedings. Transfermarkt estimates are not just reflections of market trends, they help shape them (Smith, 2021; Coates and Parshakov, 2022).

Beyond this ratings economy, football continues to grapple with persistent forms of racial discrimination. Players of color face racial abuse both on and off the field, from monkey chants in stadiums to derogatory comments on social media (Taylor et al., 2021; Panenka Magazine, 2023). These patterns underscore that performance signals in football—though highly structured—do not operate in a social vacuum.

Far from being a niche context, football thus provides a rare empirical setting for testing core economic theories of discrimination. It offers what Palacios-Huerta (forthcoming) calls a natural laboratory, with transparent, high-stakes evaluation mechanisms and standardized measures of performance. We leverage this structure to test the colorism hypothesis using one of the most stringent definitions of discrimination proposed in economics (Bertrand and Duflo, 2017).

This paper focuses on one specific type of action: shots that narrowly score or narrowly miss the goal. These events are mechanically similar in execution but evaluated differently in outcome, offering a sharp test of whether small differences in performance generate unequal rewards across player groups.

More broadly, professional football allows us to examine how discrimination can operate systemically (Bohren, Hull and Imas, 2025). Standardized evaluations flow directly into public ratings, media narratives, and market valuations, shaping players' career trajectories. The sports transparency, data richness, and economic significance make it a rare natural laboratory for testing fundamental questions about fairness and bias in performance evaluation.

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 FotMob, using over 300 individual statistics sourced from Opta.[1] FotMob confirmed the algorithmic basis of these ratings in a tweet on February 17, 2018. See Figure A.2.

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. Appendix Figure A.1 shows a strong positive correlation between average FotMob ratings and Sofifa ratings used in EA Sports football video games, which aggregate fan evaluations of player performance.

We extracted the data using the `worldfootballR` package (Zivkovic, 2023), an open-source R wrapper for retrieving structured football data from multiple online sources. At the time of our initial data collection, the package enabled access to both player-level ratings and detailed event-level data from FotMob. However, the repository was archived in September 2025 and is no longer maintained.[2] As of version 0.6.4, the package ceased providing access to FotMob data due to a change in their terms of service.[3] Consequently, we are unable to extract data beyond the 2022/23 season using this method.

[1] See the FAQ at https://www.fotmob.com/es/faq, which states that player ratings are based on Opta stats and calculated via FotMobs own model.

[2] See the archived GitHub repository at https://github.com/JaseZiv/worldfootballR.

[3] See FotMobs updated terms at https://www.fotmob.com/tos.txt.

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.

Geolocated shot data are 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 final dataset spans three complete seasons (2020/21 to 2022/23)[4] and includes over 76,000 shots on goal. These data are enriched with covariates at multiple levels: event-level (e.g., shot coordinates, outcome, match minute), player-level (e.g., position, team, minutes played), and match-level (e.g., date, stadium, home and away teams, final score).

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 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. 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.4 shows an example of one

[4] Figure A.3b displays a screenshot of the FotMob interface with shot-level data.

[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

of these images displaying post-match journalist ratings in Ligue 1. Despite limited formatting, we achieve approximately 75% coverage of relevant players. Figure A.5 shows the distribution of normalized scores by league.

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 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 on the cultural context and the enumerator or observer. We proceed by explaining our algorithm. Figure 1 illustrates descriptively the steps we take and the outcome.

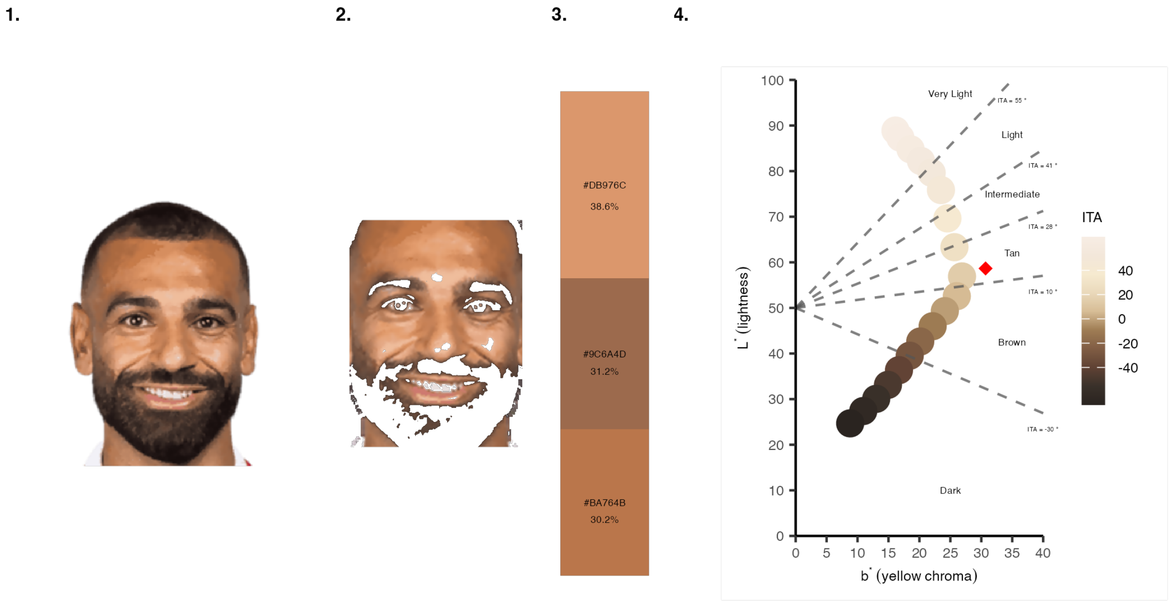


Figure 1: Skin tone segmentation and classification algorithm.

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 stored as RGB images, represented as three-dimensional arrays of shape *height* × *width* × 3, where the third dimension encodes red, green, and blue color channels. A typical resolution is 224 × 224 pixels. Our approach starts with image segmentation. First, we separate each image into two classes—foreground and background—using

when players have the same or similar last names.

the method of Otsu (1979), based on pixel-level grayscale intensity. From this, we extract a *mask* for the targeted facial region. We then apply the algorithm of Kolkur et al. (2017) for skin detection, which incorporates combinational ranges, texture, and edge features to produce a more accurate segmentation of skin areas.

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: $\text{ITA} > 55°$; Light: $41° < \text{ITA} \leq 55°$; Intermediate: $28° < \text{ITA} \leq 41°$; Tan: $10° < \text{ITA} \leq 28°$; Brown: $-30° < \text{ITA} \leq 10°$ and Dark: $\text{ITA} \leq -30°$.

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 1, 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.6 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 local-polynomial RDD specification:

$$y_{pm(ls)} = \alpha_l + \alpha_s + f(z_{ipm}) + f(z_{ipm}) \cdot \mathbf{1}(z_{ipm} > 0) + \tau \cdot \text{Goal}_{ipm} + \mathbf{X}'_{ipm}\gamma + \varepsilon_{ipm(ls)} \qquad (1)$$

where y_{pm} is the post-match rating of player p in match m, and Goal_{ipm} is an indicator equal to one if shot i resulted in a goal. The running variable z_{ipm} is a signed Euclidean distance (in centimeters) from the shot location to the goal frame, where negative values denote near misses and positive values denote goals. By construction, $z_{ipm} = 0$ corresponds to the scoring threshold—that is, the outer boundary of the goal frame.

The function $f(z_{ipm})$ denotes a flexible local polynomial in the running variable. Interacting this polynomial with an indicator for being weakly above the cutoff allows the conditional mean of the outcome to vary smoothly with distance while permitting different slopes on either side of the threshold. All specifications include league (α_l) and season (α_s) fixed effects.

The vector $\mathbf{X}_{ipm}$ contains pre-determined shot- and match-level covariates selected using the Double LASSO procedure 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 τ 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 68 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, corresponding to the optimal bias-correction bandwidth used in the local-polynomial RDD. Because players may take multiple such 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. This ensures that each player–match contributes equally to the estimation. Since shots within this 114 cm window are already rare, and multiple qualifying shots by the same player in the same match are even rarer, this weighting choice has little influence on the results. As a robustness check, we replicate our main specifications using an aggregated sample with only one shot per player–match, confirming that our findings are not driven by this feature.

To assess the plausibility of the continuity assumption, we examine the density of the running variable around the scoring threshold. Figure B.1 presents two density plots. Panel (a), which includes shots that hit the post or crossbar, reveals a visible spike just below the cutoff. This bunching is mechanical rather than behavioral: it reflects the geometry of the goal frame, not strategic shot placement. In fact, post-hitting shots are among the most clearly quasi-random misses in our data, resulting from millimeters of trajectory rather than intentional manipulation. Panel (b) excludes these shots and displays a smooth, continuous density around the cutoff, with no indication of sorting. Importantly, our baseline RDD and DiDC estimates include post shots. We show in robustness checks that excluding them yields nearly identical treatment effects, confirming that this mechanically defined subset does not bias the estimates or threaten identification.

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

$$y^{ST}_{pm(ls)} = \alpha_l + \alpha_s + f(z_{ipm}) + f(z_{ipm}) \cdot \mathbf{1}(z_{ipm} > 0) + \tau^{ST} \cdot \text{Goal}_{ipm} + \mathbf{X}'_{ipm}\gamma + \varepsilon^{ST}_{ipm(ls)} \tag{2}$$

where all variables are defined as in Equation 1. We estimate a separate treatment effect $\hat{\tau}^{ST}$ for each skin tone group, reflecting the causal impact of scoring a marginal goal on post-match ratings for players in that category.

To assess whether goal recognition varies systematically by skin tone, we compute pairwise differences in these effects:

$$\Delta = \hat{\tau}^{ST_1} - \hat{\tau}^{ST_2} \tag{3}$$

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

Conceptually, Equation 3 resembles a Mincer-type earnings regression, where Δ plays a role analogous to the "returns" or "discrimination" coefficient in an Kitagawa-Oaxaca-Blinder decomposition (Kitagawa, 1955; Blinder, 1973; Oaxaca, 1973). 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 tailored to 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.[6] To conduct inference on the difference-in-RD coefficients Δ across groups, we supplement the main estimates with a two-way clustered Bayesian bootstrap (Rubin, 1981).

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

[6] The current implementation of `rdhte` does not support two-way clustering. In our setting, potential correlation arises from two sources: player-level clustering, because skin tone is fixed at the player level; and match-level clustering, because multiple players ratings within the same match may share unobserved shocks, and each players rating is determined after all match events are observed.

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 Δ valid. We implement a two-way Bayesian bootstrap 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

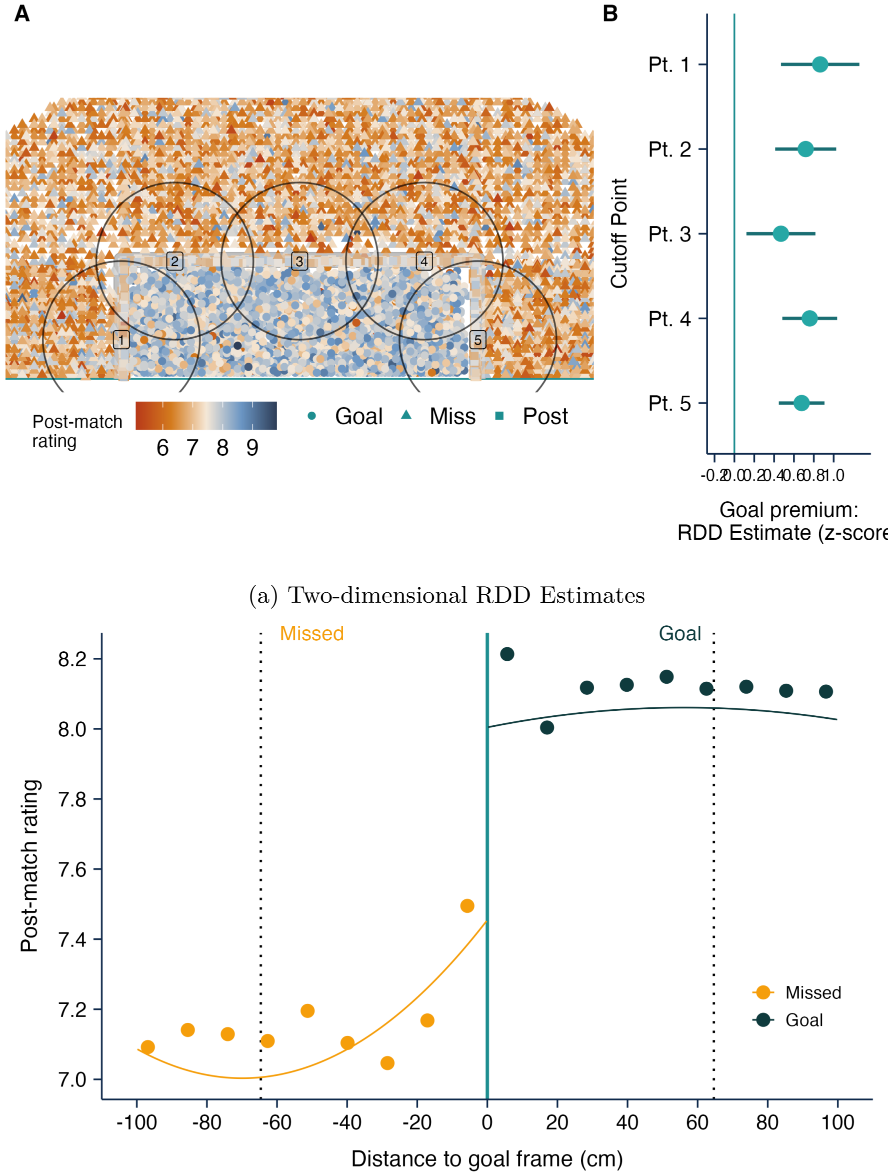
We begin by validating the first stage of our identification strategy: do narrowly scored goals result in higher post-match ratings than narrowly missed attempts? Before introducing heterogeneity by skin tone, we estimate the average causal effect of scoring on player ratings using a pooled RDD.

Figure 2a illustrates our empirical strategy. First, it provides a visual summary of the FotMob 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. It 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). In all five locations, scoring leads to statistically significant increases in player ratings. This supports the central identifying assumption that near-threshold scoring outcomes are quasi-random and suitable for causal inference.

Figure 2b presents the pooled, unidimensional regression discontinuity design (RDD). The outcome variable is the players algorithmic post-match rating, and the running variable is the shots Euclidean distance to the goal frame, with zero marking the scoring threshold. The plot reveals a clear discontinuity at the cutoff: narrowly scoring increases the average rating by approximately 0.5 points, or 0.575 standard deviations—around 7% increase relative to the mean.

This estimate is statistically significant at the 1% level and remains robust across a range of alternative specifications. Appendix Table B.1 reports the RDD estimates under different model configurations. The estimated goal premium is stable when controlling for a rich set of shot- and player-level covariates (using both clean controls and Double LASSO), and when including match fixed effects. Results are also robust to alternative smaller and bigger bandwidths, kernel



(a) Two-dimensional RDD Estimates

(b) One-dimensional RDD Plot

Figure 2: Goal Scoring Premium

Notes: Panel (a) shows the spatial distribution of shots relative to the goal frame, with color indicating the shooters post-match rating (from algorithmic sources). Each vertical segment marks a scoring threshold (e.g., post or crossbar). Using local two-dimensional RDDs (Cattaneo et al., 2025), we compare outcomes for shots just inside versus just outside each boundary. All five cutoffs show statistically significant increases in ratings for goals. Panel (b) presents the pooled, one-dimensional RDD using Calonico et al. (2014), with Euclidean distance to the goal frame as the running variable. The solid vertical line denotes the scoring cutoff. Ratings exhibit a clear discontinuity: narrowly scoring increases the average post-match rating by approximately 0.5 points, or 0.575 standard deviations—a 7% increase relative to the mean. Estimates use a 68 cm MSE-optimal bandwidth. All ratings are algorithm-based.

functions, and polynomial orders. Across all specifications, the estimated goal premium ranges from 0.45 to 0.58 standard deviations.

Consistent with Gauriot and Page (2019), 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 68 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 appears more consistent with a reward for scoring than with a penalty for missing. Appendix Figure B.2 presents suggestive evidence on this point by showing the distribution of post-match ratings for three groups: (i) the full universe of player ratings, including players regardless of shot-taking; (ii) players who missed a close-range shot on goal; and (iii) players who scored a goal. The distribution for missed shots closely resembles the full rating distribution, while the distribution for goal-scoring players is clearly shifted to the right, with a higher mean. Although this comparison is not causal, it suggests that narrowly missing a goal does not lead to substantial rating penalties, and that the observed goal premium in our RDD results likely reflects positive recognition for success rather than punishment for failure.

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 $\mathrm{ST} \in \{\mathrm{Dark}, \mathrm{Tan}, \mathrm{Light}\}$. We then compare the estimated effects across groups using the Difference-in-Discontinuities framework described in Equations 2 and 3.

For each group, we estimate the group-specific goal premium $\hat{\tau}^{ST}$. 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.[7]

To formally test for discrimination, we conduct one-sided hypothesis tests based on pairwise differences defined as $\Delta = \hat{\tau}^{ST_{\text{darker}}} - \hat{\tau}^{ST_{\text{lighter}}}$. For each comparison, we test $H_0 : \Delta \geq 0$ against $H_1 : \Delta < 0$, where rejection of the null indicates that 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 stan-

[7] Throughout the heterogeneous RD analysis, we follow the implementation of `rdhte` (Calonico et al., 2025*a*), which differs slightly from the conventional `rdrobust` setup. In particular, `rdhte` relies on built-in least squares base commands and, by default, sets the main estimation bandwidth equal to the bias-correction bandwidth ($h = b$), rather than using the smaller MSE-optimal bandwidth for point estimation and a larger bandwidth only for bias correction. In our setting, the MSE-optimal bandwidth for the pooled RDD is approximately 68 cm, with a corresponding bias-correction bandwidth of 114 cm. Because the heterogeneous analysis requires estimating separate RDDs within mutually exclusive skin tone groups—substantially reducing effective sample sizes near the cutoff—we adopt the larger bandwidth ($h = b = 114$ cm) as a conservative benchmark to ensure stable estimation and valid inference across subgroups. Importantly, all heterogeneity results are robust to narrower bandwidth choices, including local randomization inference using ±15 cm windows, which lie well within the region of quasi-random assignment.

dard errors clustered at the player–match level.

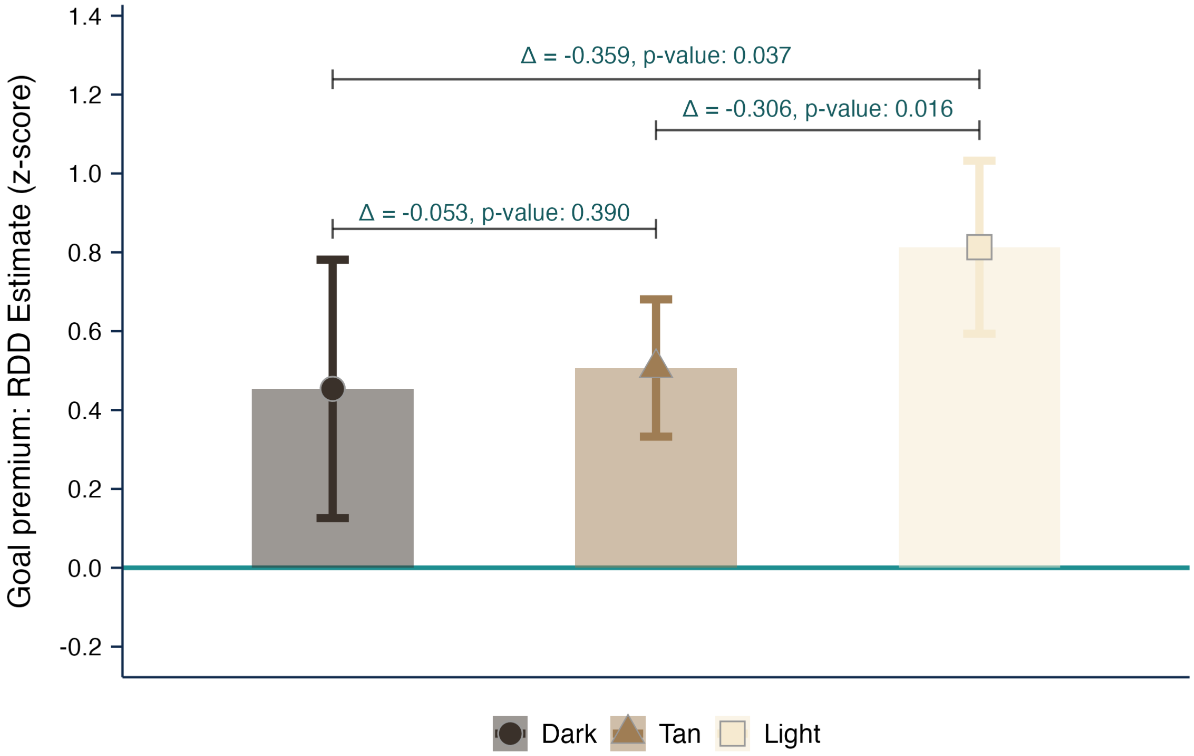


Figure 3: Difference-in-Discontinuities: Goal Premium by Skin Tone

Notes: This figure reports heterogeneous bias-corrected RDD estimates of the goal premium—the causal effect of scoring a marginal goal on algorithmic post-match ratings—by skin tone group, estimated using the `rdhte` package (Calonico et al., 2025*b*). The y-axis shows treatment effects in standardized rating units; points denote group-specific estimates $\hat{\tau}$ from Equation 1, with 95% confidence intervals clustered at the player–match level. Brackets report pairwise differences defined as $\Delta = \hat{\tau}^{ST_{\text{darker}}} - \hat{\tau}^{ST_{\text{lighter}}}$, with one-sided p-values from tests of $H_0 : \Delta \geq 0$ against $H_1 : \Delta < 0$. Under this convention, negative values of Δ indicate greater recognition for lighter-skinned players. Following the default `rdhte` implementation, estimation uses a bandwidth of $h = 114$ cm (equal to the bias-correction bandwidth in the pooled RDD), which improves precision when estimating subgroup-specific effects. All specifications use a triangular kernel, include Double LASSO-selected covariates, and weight observations by the inverse number of within-bandwidth shots per player–match.

Figure 3 presents the main results. Light-skinned players receive a goal premium of 0.81 standard deviations; Tan-skinned players receive 0.50 standard deviations; and Dark-skinned players receive 0.45 standard deviations. The one-sided tests reveal significant disparities: Light-skinned players receive 0.36 standard deviations more recognition than Dark-skinned players ($p = 0.037$), and 0.31 standard deviations more than Tan-skinned players ($p = 0.016$). While Tan-skinned players also receive more recognition than Dark-skinned players ($\Delta = 0.053$), the difference is not statistically significant ($p = 0.390$).

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 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.3 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 lighter skin tone leads to a 0.13 standard deviation higher goal premium (two-sided $p = 0.059$).

To illustrate the magnitude, moving from the Dark ITA category to the Very Light category represents roughly a four–standard deviation shift. Over this range, the estimated goal premium rises from about 0.23 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 the universe all players who received post-match ratings during the same period.

Appendix Figure B.4 shows substantial overlap between the two distributions. A Kolmogorov-Smirnov (KS) test fails to reject equality (D = 0.020, p = 0.349), indicating no significant compositional differences. This supports the validity of our research design and suggests that observed disparities reflect differential treatment, not differential selection.

Balance in Shot Difficulty and Skin Tone at the Cutoff A second threat to identification is that players of different skin tones may systematically attempt shots from different locations, so that marginal goals and marginal misses differ in difficulty across groups. If this were the case, observed rating disparities could reflect endogenous shot placement rather than unequal evaluation of identical performance. As a first diagnostic, we test whether skin tone itself changes discontinuously at the scoring threshold by estimating an RDD with standardized ITA as the outcome. Appendix Figure B.5 shows no evidence of a discontinuity: average skin tone evolves smoothly through the cutoff, indicating that marginal goals and marginal misses are taken by observationally similar players in terms of skin tone.

Second, conditional on scoring, we test whether players of different skin tones place shots in systematically different locations within the goal frame. We compare the distribution of distance to the goal frame edge among successfully scored goals within the RDD bandwidth. Appendix Figure B.6 shows substantial overlap across groups, and Kolmogorov-Smirnov tests fail to reject equality in all pairwise comparisons. Together, these results indicate that skin tone is neither sorted at the cutoff nor correlated with shot placement among scorers, reinforcing that observed rating gaps reflect differential evaluation of equivalent performance rather than differences in shot difficulty.

Robustness to Specification Choices and Unit of Analysis. Our DiDC are robust to a wide range of alternative specifications, as shown in Appendix Table B.2. To address concerns that unobserved factors—such as biased refereeing or match-specific dynamics—may influence both scoring and ratings, we re-estimate our model using match fixed effects (demeaned within match) and player fixed effects (demeaned within player). These specifications absorb any performance shifts due to referees or venue context, and account for stable differences in players' skill at converting marginal chances. In all cases, the skin tone gradient in recognition persists. Results are also robust to alternative kernel choices (Epanechnikov and uniform), exclusion of shots hitting the post, and a design that aggregates to a single within-bandwidth shot per player-match. This last check addresses potential attenuation bias from multiple shots receiving the same post-match rating. Across all specifications, we continue to find that Light-skinned players receive significantly higher rating premia than Dark- and Tan-skinned players for identical scoring events.

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 (±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 4 shows results from the binned event-study analysis. The left panel plots predicted post-match ratings for near-miss shots by skin tone group. Although the pre-threshold lines are somewhat noisy, they overlap closely and show no statistically detectable differences. A joint F-test fails to reject equality across groups ($p = 0.766$), consistent with the identifying assumption that, near the cutoff, ratings would have evolved similarly absent scoring.

The right panel reveals sharp rating increases at the scoring threshold, with systematic divergence: ratings rise to roughly 8.2 for light-skinned players, 8.0 for tan-skinned players, and 7.8 for dark-skinned players. A post-threshold F-test rejects equality ($p = 0.006$), indicating heterogeneous recognition.

Importantly, the gradient attenuates beyond the MSE-optimal window, where groups begin to converge. This pattern suggests that disparities are concentrated among marginal goals, where evaluator discretion is greatest and causal identification is strongest.

Balance Tests for Pre-Determined Covariates To assess the plausibility of the RDD identifying assumption, we test for discontinuities in 13 pre-determined covariates at the scoring threshold. Under quasi-random assignment, characteristics determined prior to the shot—such as shot coordinates, match minute, or home status—should evolve smoothly around the cutoff. We estimate group-specific RDDs by skin tone following Calonico et al. (2025*b*) and apply false discovery rate (FDR) adjustments within each group (Anderson, 2008). Continuous variables are standardized; binary variables are left unchanged.

Appendix Table B.3 reports the results. Of the 39 covariate–group comparisons, 38 show no statistically significant discontinuity after FDR correction. The sole exception is the vertical shot coordinate (Shot y) for Tan-skinned players, which exhibits a statistically significant but isolated

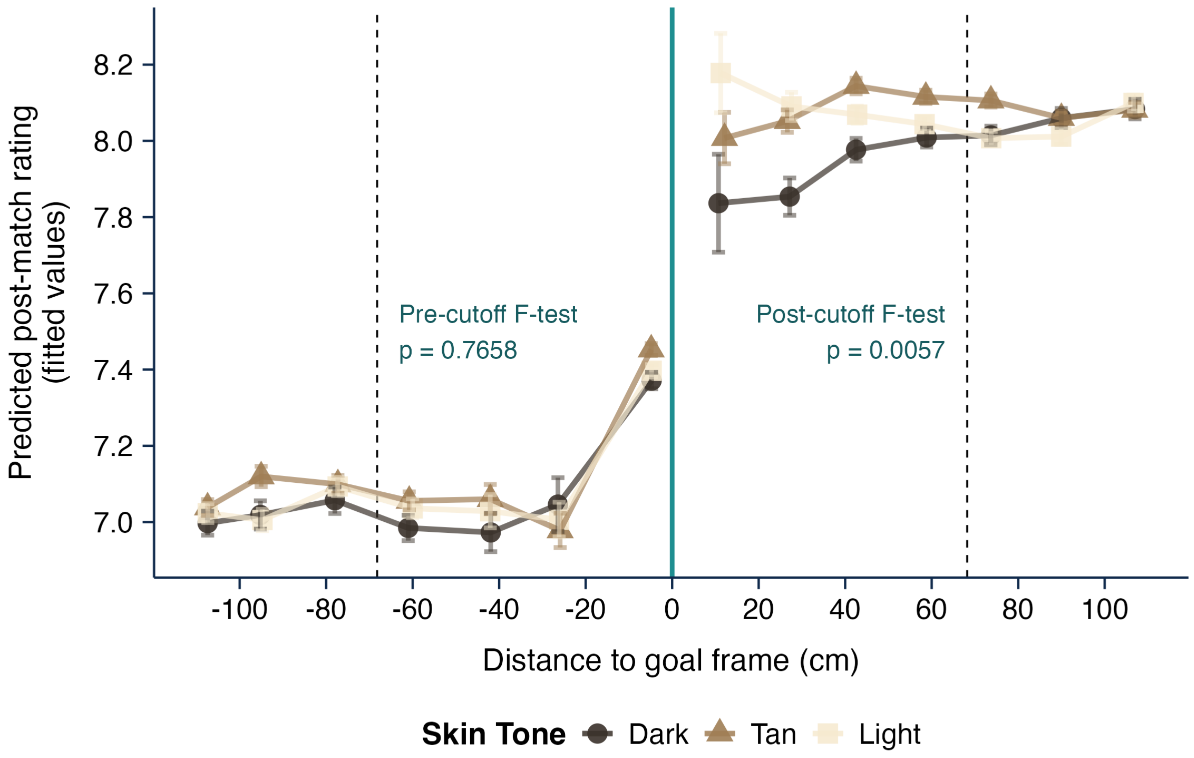


Figure 4: Binned Event-Study: Parallel Trends and Heterogeneous Treatment Effects

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 (±68 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.

jump at the threshold. Importantly, shot coordinates—including Shot y—are included as controls in all baseline RDD and DiDC specifications. No other shot-, player-, or match-level covariate displays evidence of imbalance.

Taken together, these results indicate that observable characteristics are overwhelmingly smooth at the cutoff, and that the single detected imbalance is transparently addressed in our estimation. This supports the validity of the DiDC design and suggests that estimated differences in goal premia are unlikely to be driven by confounding discontinuities in pre-determined covariates.

Within-Match Performance Metrics While pre-determined characteristics should evolve smoothly at the scoring threshold, scoring a goal may mechanically or behaviorally affect other within-match performance metrics—such as shots, passes, or defensive actions—that also enter the construction of post-match ratings. If such post-treatment outcomes respond differently across skin tone groups, they could in principle mediate part of the observed rating premium. At the same time, conditioning on these variables in the main DiDC specification would constitute post-treatment bias, as they are themselves affected by scoring (i.e., "bad controls" in the sense of Angrist and Pischke, 2009).

We therefore treat within-match performance metrics as *post-treatment mediators* and examine them only in an ancillary robustness exercise. We begin by testing whether marginally scoring induces discontinuous changes in 35 within-match performance variables spanning offense, passing, defense, and discipline. Using group-specific RDDs and controlling the false discovery rate at 10% within each skin tone group, we identify which metrics respond sharply at the scoring threshold.

Appendix Table B.4 shows that six variables exhibit statistically significant jumps in at least one group: the dependent variable (post-match rating), goals scored, big chances missed, missed penalties, accurate long balls, and dispossessions. Among these, goals scored and big chances missed are mechanically linked to the scoring outcome and therefore do not reflect meaningful behavioral responses.[8]

We therefore focus on the remaining non-mechanical outcomes—accurate long balls and dispossessions—which plausibly capture changes in on-field behavior or evaluator-relevant performance following a goal. In addition, we control for the total number of goals scored in the match to absorb residual mechanical exposure.

Appendix Figure B.7 presents DiDC estimates re-estimated with these three post-treatment variables (goals scored, accurate long balls, and dispossessions) included as controls. The colorism gradient remains clearly visible. If anything, controlling for these outcomes slightly increases the estimated group-level differences. This robustness check reinforces our interpretation that differential recognition by skin tone is not driven by downstream changes in observable performance, but instead reflects differences in how identical performance shocks are evaluated.

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 10,000 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.8 shows the resulting goal premia and pairwise differences. Light-skinned players receive significantly more recognition than both Tan- and Dark-skinned players, with estimated differences of 0.41 standard deviations (vs. Tan, $p = 0.013$) and 0.37 (vs. Dark, $p = 0.063$). The Tan-Dark gap is small and statistically insignificant (0.04, $p = 0.574$). These results confirm that the main heterogeneity is between Light-skinned players and the other two groups, consistent with a colorism gradient.

Appendix Figure B.9 displays the bootstrap distributions. Panel A shows kernel densities of the estimated goal premia by group. Panel B shows ECDFs: Light-skinned players dominate at all quantiles, indicating first-order stochastic dominance. Kolmogorov-Smirnov tests confirm these

[8] While missed penalties is also post-treatment, it is deterministically defined for penalty shots based on whether the shot is scored, and thus not treated as a behavioral mediator. In practice, missed penalties are rare in the local neighborhood of marginal goals and are not included in any robustness specification.

differences are statistically significant between Light vs. Dark and Light vs. Tan ($p < 0.001$), with smaller and less robust differences between Dark and Tan. These findings reinforce our main result: algorithmic recognition for goals is systematically biased in favor of lighter skin tones.

Robustness: Ruling Out Religious Discrimination To the extent that skin tone may proxy for religious identity in European leagues, a potential concern is that the estimated skin-tone gradient captures religious discrimination rather than colorism. In the absence of direct measures of religion, we re-estimate our DiDC design restricting the sample to players from Latin American countries—Brazil, Argentina, Uruguay, Colombia, and Mexico—where the vast majority of players are Christian, holding religion relatively constant while preserving variation in skin tone. As shown in Appendx Figure B.10, heterogeneous goal recognition persists within this subsample: Light-skinned players receive systematically higher rating premia than Tan- and Dark-skinned players for identical scoring events. This persistence suggests that religion is unlikely to be a primary driver of the results and reinforces an interpretation centered on colorism.

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 (114 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 circumference) and assume as-if random assignment within this range. This approach avoids functional form and bandwidth selection but imposes stronger assumptions about local randomization.

Appendix Figure B.11 reports the group-specific estimates: Dark-skinned players receive a goal premium of 0.45 standard deviations (not statistically significant), Tan-skinned players 0.68, and Light-skinned players 1.01. These results echo our continuity-based findings.

Pairwise differences reinforce the pattern: Light-skinned players receive significantly more recognition than Dark-skinned players ($\Delta = 0.565$, $p = 0.046$) and marginally more than Tan-skinned players ($\Delta = 0.339$, $p = 0.082$); the Tan–Dark gap remains small and statistically insignificant ($\Delta = 0.226$, $p = 0.242$). Despite the tighter window and stricter assumptions, the colorism gradient persists, reinforcing the robustness of our main 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 median 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.12 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.

Heterogeneity in Other Player Characteristics A potential concern 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 shape algorithmic evaluations. If the algorithm exhibited preferential treatment along multiple correlated dimensions, our skin-tone results might reflect broader patterns of heterogeneity rather than colorism per se. To assess this, we examine heterogeneous goal premia across five alternative player dimensions: shooting foot, nationality (native vs. non-native to the league country), age, height, and weight. For comparison, we also report a two-group skin-tone contrast (Light vs. Tan & Dark).[9]

For each dimension, we estimate an RDD specification identical to our main model and test for differences in the goal premium across groups. Appendix Figure B.13 reports the results. Skin tone stands out in terms of statistical robustness: Light-skinned players receive a goal premium of 0.81 standard deviations, compared to 0.51 for Tan & Dark players, yielding a difference of 0.30 ($p = 0.014$), which remains marginally significant after false discovery rate adjustment ($q = 0.092$).

By contrast, several other dimensions—most notably height and footedness—exhibit point estimates of similar magnitude, but with substantially wider confidence intervals and no statistical significance after multiple-testing correction. None of the alternative characteristics display robust evidence of heterogeneous recognition comparable to skin tone.

Taken together, these results suggest that while heterogeneity in goal recognition may exist along other dimensions, skin tone is the only characteristic for which disparities are both economically meaningful and statistically robust in our data. This pattern is consistent with algorithmic colorism rather than a more general tendency to reward particular physical or demographic traits.

Alternative Estimation: Parametric Spatial Matching While our RDD and DiDC estimates provide quasi-experimental identification by exploiting random variation near the scoring threshold, they necessarily rely on a narrow bandwidth, continuity assumptions, and nonparametric estimation methods. To assess robustness to parametric specification, we implement an alternative identification strategy using a different sample: the universe of on-target shots within the goal frame (i.e., goals and goalkeeper saves).

This parametric design compares post-match ratings for all goals versus saves, matched by exact spatial position within the goal frame. By leveraging this full sample and introducing spatial fixed effects, we relax continuity assumptions, increase statistical power, and allow a direct test of the mechanism suggested by Figure 4: whether discrimination is most pronounced near the goal frame edge, where randomness dominates, and attenuates toward the center, where goalkeeper skill plays a greater role.

We divide the goal frame (7.32 × 2.44 meters) into a 20 × 8 rectangular grid (160 cells), matching the natural 3:1 width-to-height ratio. Within each cell, we compare post-match ratings for goals

[9] For age, height, and weight, we merge our data with the SoFIFA player database using fuzzy name and nationality matching, harmonize units, and construct binary indicators for above- and below-mean values.

versus saves, controlling for match fixed effects, player role, shot timing, and shot origin. Distance to the goal frame edge is defined as the minimum distance (in centimeters) to the left post, right post, or crossbar, excluding the ground, which reflects goalkeeper positioning rather than a structural boundary. All models weight by the inverse number of shots per player–match and cluster standard errors at the player–match level.

Table 1 presents the results. Column (1) reports a baseline goal premium of 1.09 SD. Column (2) adds spatial cell fixed effects; the premium remains 1.06 SD, indicating large rating gains even within identical shot locations. Column (3) introduces our key test: an interaction between goals and standardized skin tone (ITA). The coefficient of 0.010 implies that each one-standard-deviation increase in skin lightness yields an additional 0.010 SD boost in ratings. Moving from very dark to very light skin (a four SD shift) implies a cumulative advantage of 0.040 SD—approximately 4% of the baseline goal premium.

Table 1: Parametric Estimation: Spatial Matching of Goals vs. Saves

	Post-match rating (z-score)							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Goal	1.09	1.06	1.06	1.06	1.06	1.05	1.08	1.06
	(0.007)	(0.008)	(0.008)	(0.019)	(0.019)	(0.019)	(0.020)	(0.021)
Goal × Skin tone (ITA, z)			0.010	0.009	0.032	0.044	0.033	0.031
			(0.006)	(0.006)	(0.013)	(0.014)	(0.014)	(0.014)
Goal × Skin tone × Distance					−0.0002	−0.0002	−0.0002	−0.0002
					(9.94×10^{-5})	(0.0001)	(0.0001)	(0.0001)
Observations	91,666	91,666	91,666	91,666	91,666	91,666	91,666	91,666
R^2	0.2017	0.2049	0.2049	0.2049	0.2049	0.3005	0.3377	0.3411
Within R^2		0.1570	0.1570	0.1570	0.1570	0.1546	0.1410	0.1340
Spatial FE (20×8)		✓	✓	✓	✓	✓	✓	
Spatial FE (40×16)								✓
Match FE						✓	✓	✓
Shot controls							✓	✓

Notes: Dependent variable is the standardized post-match rating (FotMob). Sample includes all on-target shots within the goal frame (goals and goalkeeper saves). Skin tone is measured by ITA (Individual Typology Angle), standardized to mean zero and unit variance. Distance to edge is the minimum distance to the left post, right post, or crossbar (in cm). All models are weighted by the inverse number of within-match shots per player. Standard errors are clustered at the player–match level. Columns (1)–(7) use a 20×8 spatial grid; column (8) uses a 40×16 grid.

Column (5) introduces the triple interaction that directly tests whether the skin tone gradient varies with proximity to the goal frame edge. The negative coefficient on Goal × Skin tone × Distance (−0.0002, $p = 0.034$) confirms the pattern observed in Figure 4: discrimination is strongest near the structural boundaries of the goal frame (0.032 SD per SD of ITA) and diminishes with distance—falling to 0.022 SD at 50 cm and approaching zero by 100 cm. Columns (6)–(8) demonstrate robustness to match fixed effects, shot-level controls, and finer spatial resolution (40×16 grid). The estimated gradient remains stable across specifications, reinforcing the broader pattern—identified in both RDD and parametric analyses—that evaluative bias is most pronounced in marginal scoring situations where randomness plays a larger role in outcomes.

These findings validate our RDD and DiDC estimates using a complementary, parametric framework, and provide direct causal evidence that evaluative bias emerges most strongly in contexts where outcomes are determined by marginal luck rather than observable skill.[10]

5.2.2 Heterogeneity in main results

We explore whether the colorism gradient in goal-scoring premiums varies across match circumstances. We conduct heterogeneity analyses using the Local RDD framework with a narrow window of ±15 cm around the scoring threshold.[11]

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 that is not statistically significant. However, the 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 may amplify algorithmic bias, potentially reflecting greater responsiveness to fan reactions, media narratives, or other contextual 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. Appendix Figure B.15 shows that the light-dark gap in algorithmic recognition remains across both match outcomes. When teams lose or draw, light-skinned players receive 0.212 standard deviations more recognition than dark-skinned players for scoring ($p = 0.061$). When teams win, the gap is slightly smaller at 0.187 standard deviations ($p = 0.088$). While these point estimates differ slightly, their magnitudes are similar and not statistically distinguishable. Thus, rather than indicating true heterogeneity, the results highlight the persistence of discriminatory recognition regardless of match outcome. If anything, algorithmic bias remains stable across contexts—even when teams succeed. Any potential amplification when teams underperform should be interpreted cautiously and is not supported by a formal test of heterogeneity.

6 Mechanisms

What explains the skin tone disparities in algorithmic ratings documented in Section 5? We investigate two complementary mechanisms. First, we assess whether the observed differences

[10] While the parametric estimates confirm the presence of a colorism gradient, the magnitude is smaller than in the DiDC analysis. This likely reflects differences in the comparison group: the DiDC contrasts goals with near misses, which receive lower ratings on average, while the parametric design compares goals to goalkeeper saves—already high-recognition events. With both outcomes rated highly, the scope for differential reward is narrower, yielding a smaller but directionally consistent gradient.

[11] 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.

reflect underlying performance variation that the algorithm captures accurately. Second, we test whether algorithmic ratings inherit bias from human evaluations embedded in the broader football ecosystem.

6.1 Decomposing Algorithmic Discrimination

To distinguish between performance-based and bias-based explanations, we implement a prediction residual decomposition. We decompose observed ratings into two components:

$$y_{pm} = \hat{f}(\mathbf{X}_{pm}) + \hat{\varepsilon}_{pm} \tag{4}$$

where y_{pm} is the observed algorithmic rating for player p in match m, $\hat{f}(\mathbf{X}_{pm})$ represents the predicted rating based on observable performance metrics, and $\hat{\varepsilon}_{pm}$ is the residual component capturing unexplained variation.

If algorithmic ratings are fully driven by observable match statistics, then a predictive model trained on those statistics should replicate any skin tone disparities. Conversely, if residuals from this model differ by skin tone, this suggests that darker-skinned players are systematically underrated, conditional on performance.

We train an XGBoost model, a gradient-boosted decision tree, to estimate $\hat{f}(\cdot)$ and predict post-match ratings using all available performance metrics $\mathbf{X}_{pm}$: goals, assists, passes, tackles, dribbles, fouls, and over 40 additional variables. We exclude features recorded in fewer than 5% of observations, train on 75% of the data, and test on the remaining 25%. The model achieves $R^2 = 0.875$, indicating that most of the variation in ratings can be explained by objective data. We compute residuals as $\hat{\varepsilon}_{pm} = y_{pm} - \hat{f}(\mathbf{X}_{pm})$ for all observations.

We then estimate Equation 2 separately for each component in Equation 4: (i) predicted ratings $\hat{f}(\mathbf{X}_{pm})$, (ii) actual ratings y_{pm}, and (iii) residuals $\hat{\varepsilon}_{pm}$.

Figure 5 presents the results. In predicted ratings based solely on objective performance (left panel), there is no clear or monotonic skin tone gradient in goal premia. The Light–Dark difference is small and statistically insignificant (0.117 SD, $p = 0.244$), and Tan players are statistically indistinguishable from both Light and Dark players. This suggests that observable match statistics alone do not generate systematic differences in recognition by skin tone.

In contrast, actual algorithmic ratings (center panel) replicate our main results: Light-skinned players receive a substantially larger goal premium than both Tan and Dark players. Thus, disparities arise in realized evaluations, but not in predicted performance.

The residual component (right panel) reveals the sharpest and most monotonic pattern. After netting out all observable performance, the goal premium declines stepwise with darker skin tone: Light-skinned players receive the largest residual boost, Tan players a smaller one, and Dark-skinned players the least. The Light–Dark residual gap is 0.241 SD ($p = 0.006$), with a smaller but still notable Light–Tan difference (0.092 SD, $p = 0.099$).

Based on these estimates, the observed gap in goal premia between skin tone groups can be decomposed as:

$$\Delta_y = \Delta_{\hat{f}(\mathbf{X})} + \Delta_{\hat{\varepsilon}} \tag{5}$$

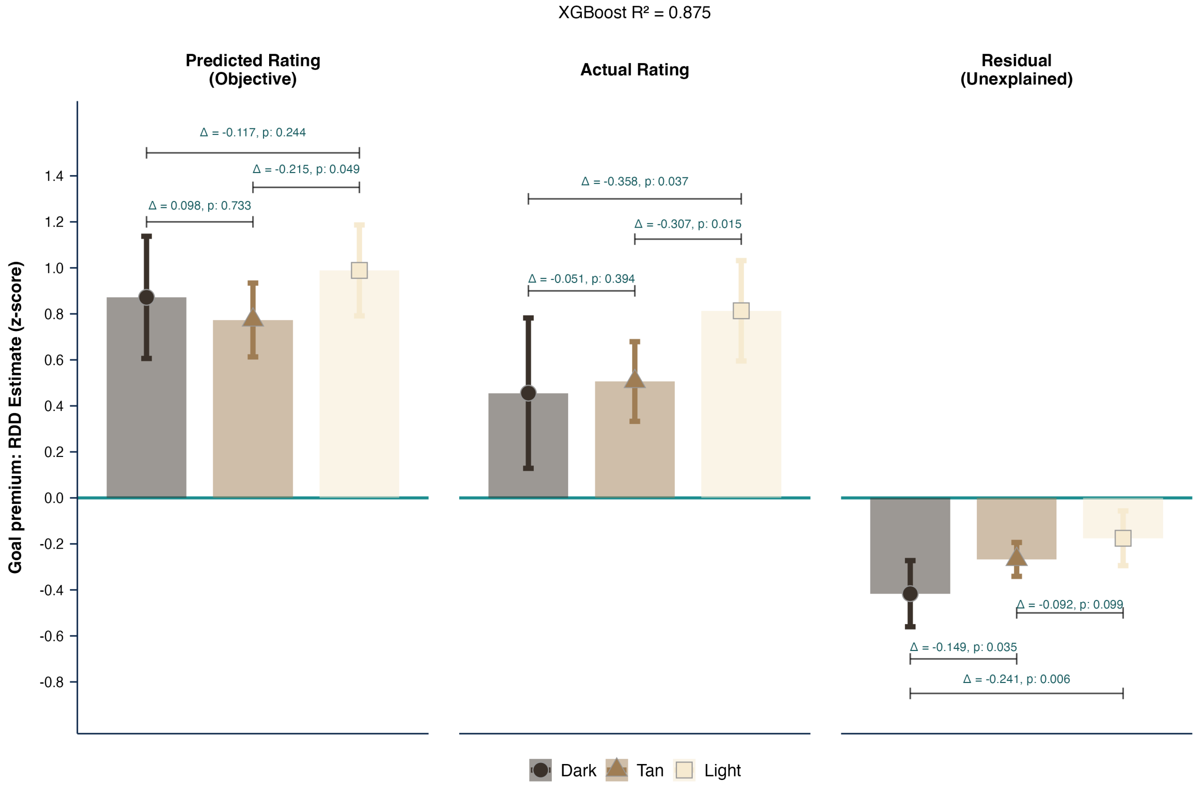


Figure 5: Decomposition of Skin Tone Gap in Goal Premia: Objective Performance vs. Algorithmic Bias

Notes: This figure decomposes the skin tone gradient in goal premia into the components defined in Equation 4. We train an XGBoost model ($R^2 = 0.875$) to predict post-match ratings using all available performance statistics (goals, assists, passes, tackles, dribbles, fouls, and 40+ additional metrics), then re-estimate Equation 2 separately for: (i) predicted ratings $\hat{f}(\mathbf{X}_{pm})$ based only on objective performance (left panel), (ii) actual algorithmic ratings y_{pm} (center panel), and (iii) residuals $\hat{\varepsilon}_{pm}$ representing the unexplained component (right panel). Each panel displays RDD estimates of the goal premium by skin tone group within a 114 cm bandwidth around the goal line, with 95% confidence intervals clustered at the player–match level. Brackets indicate pairwise differences between skin tone groups with corresponding p-values from linear hypothesis tests.

where Δ_y represents the total difference in goal premia, $\Delta_{\hat{f}(\mathbf{X})}$ captures the component explained by observable performance metrics, and $\Delta_{\hat{\varepsilon}}$ reflects the unexplained residual component.

We infer that approximately 32.7% of the Light–Dark disparity is attributable to differences in recorded performance, $\Delta_{\hat{f}(\mathbf{X})}$, with the remaining 67.3% unexplained by observable metrics, $\Delta_{\hat{\varepsilon}}$.[12] Given the model's high predictive power ($R^2 = 0.875$), and the quasi-random variation in the DiDC, this unexplained residual gradient is unlikely to reflect omitted performance variables. Instead, it points to systematic algorithmic bias operating through subjective or non-performance-based channels. In short, darker-skinned players receive less recognition than equally performing peers, and this unequal treatment emerges precisely where evaluator discretion is greatest.

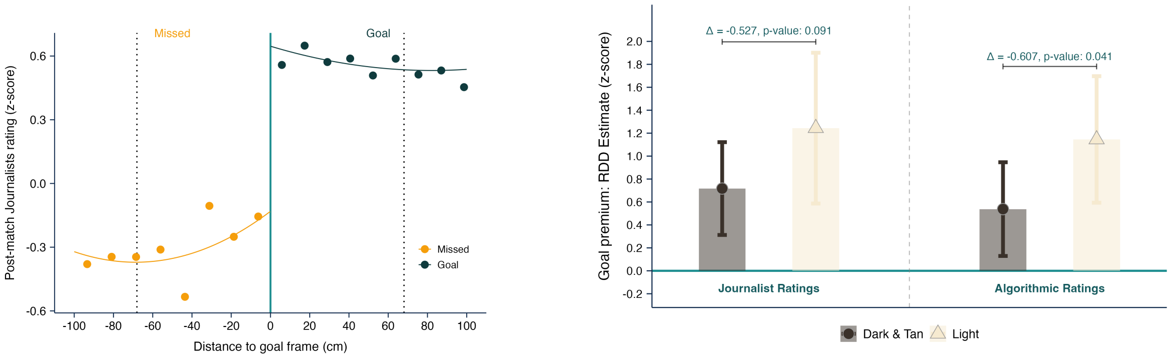
[12] This decomposition share should be interpreted with caution: the 32.7% figure is a point estimate derived from an imprecise predicted component, and is not statistically distinguishable from zero. Importantly, the monotonic skin tone gradient appears only in the residual component, not in the predicted one.

6.2 Journalist-Based RDD Analysis

If algorithms are designed to reflect objective performance, what explains the residual bias? One possibility is that algorithms absorb human evaluation biases embedded in training data or calibration routines. To assess this, we test whether human-assigned post-match ratings exhibit similar disparities by skin tone.

We replicate our RDD and DiDC (Equations 1 and 2) design using journalist-assigned ratings from the Premier League, Serie A, and Ligue 1. These ratings, published shortly after each match by major media outlets, serve as a public-facing benchmark for expert evaluation.

We re-estimate the MSE-optimal bandwidth using the journalist subsample and demean ratings within match to remove level shifts across journalists. Due to smaller sample sizes—arising from fewer leagues and seasons with available journalist data—we merge the *Tan* and *Dark* categories into a single group. This grouping preserves statistical power while still allowing for meaningful comparison with Light-skinned players. All other aspects of the empirical design remain consistent.



(a) RDD: Journalist Ratings

(b) DiDC: Journalist vs. Algorithmic Ratings

Figure 6: Journalist-Based RDD and DiDC Estimates

Notes: Panel (a) shows the RDD estimate of the journalist-based goal premium, using an MSE-optimal bandwidth re-estimated on the journalist dataset. Panel (b) 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. Δ values represent pairwise differences in the estimated goal premium, defined as Dark & Tan minus Light. Negative values of Δ therefore indicate a larger goal premium for Light-skinned players. p-values correspond to 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.

Figure 6 (left) shows that journalist ratings exhibit a clear RDD jump at the scoring threshold, similar in magnitude to algorithmic ratings. In the DiDC analysis (right), journalist-assigned ratings also display a substantial skin tone gradient: Light-skinned players receive a 0.53 SD larger goal premium than Dark and Tan players—a gap nearly identical to that in algorithmic evaluations.

6.3 Testing Bias Inheritance: Journalist Ratings and Algorithmic Residuals

To directly test whether algorithms inherit bias from human judgments, we exploit the subsample of 4,130 player–matches (across 2,223 matches and 1,038 players) with both journalist and algorithmic ratings. We decompose the player-match algorithmic ratings y_{pm} into (i) XGBoost predictions $\hat{f}(\mathbf{X}_{pm})$, the objective performance, and (ii) residuals $\hat{\varepsilon}_{pm}$, the unexplained variation, as defined in Equation 4. We then test whether journalist evaluations—both raw scores and *journalist bias* (journalist score minus predicted rating)—predict each component.

All regressions include player and match fixed effects and interact journalist evaluations with ITA, our continuous skin tone measure. Standard errors are two-way clustered by player and match. We report results in Table 2, which shows how journalist evaluations predict both algorithmic components: the objective performance estimate $\hat{f}(\mathbf{X}_{pm})$ and the residual component $\hat{\varepsilon}_{pm}$, where bias is most likely to manifest.

Table 2: Journalist Ratings Predict Algorithmic Bias: Evidence of Inheritance

	Algorithmic residual	XGBoost prediction	Algorithmic residual
	$\hat{\varepsilon}_{pm}$	$\hat{f}(\mathbf{X}_{pm})$	$\hat{\varepsilon}_{pm}$
	(1)	(2)	(3)
Journalist rating (z)	0.385	0.726	
	(0.031)	(0.027)	
Journalist rating × Skin tone (ITA, z)	0.088	-0.006	
	(0.031)	(0.025)	
Journalist bias (z)			0.423
			(0.033)
Journalist bias × Skin tone (ITA, z)			0.115
			(0.034)
Observations	4,130	4,130	4,130
R^2	0.8185	0.8798	0.8207
Within R^2	0.1257	0.4251	0.1362
Player FE	✓	✓	✓
Match FE	✓	✓	✓

Notes: This table tests whether algorithmic ratings inherit bias from journalist evaluations. Sample: 4,130 player–matches from three leagues (Premier League, Serie A, Ligue 1) covering 1,038 unique players across 2,223 matches. Dependent variables are standardized (z-scores). Columns (1)–(2) use raw journalist ratings as the predictor; column (3) uses journalist bias, defined as journalist rating minus XGBoost prediction. Algorithmic residuals (columns 1 and 3) capture unexplained rating variation $\hat{\varepsilon}_{pm}$ after controlling for all observable performance statistics; XGBoost prediction (column 2) captures the objective performance component $\hat{f}(\mathbf{X}_{pm})$. ITA is Individual Typology Angle measuring continuous skin tone (higher = lighter). All specifications include player and match fixed effects. Standard errors (in parentheses) are two-way clustered at player and match level.

Column (1) shows that journalist ratings significantly predict algorithmic residuals $\hat{\varepsilon}_{pm}$, with stronger effects for lighter-skinned players. Column (2) confirms that journalists also track objective performance $\hat{f}(\mathbf{X}_{pm})$, but without any skin tone gradient. Column (3) isolates bias transmission: when journalists rate a player 1 SD above their objective performance, algorithmic residuals rise by 0.42 SD, with a 3.4-fold stronger effect for very light-skinned players than very dark-skinned players.

These findings provide direct evidence that algorithms do not generate bias independently, but rather inherit and amplify colorism already present in human evaluations. Critically, this inheritance operates through the subjective, residual component $\hat{\varepsilon}_{pm}$ of ratings and is strongly moderated by skin tone.

While our specifications absorb within-player and within-match variation, we acknowledge the possibility of endogeneity in recent seasons, where journalist and algorithmic ratings may influence one another. However, the historical record suggests a clear direction of influence: journalist ratings have been published since the early 1990s, whereas algorithmic platforms emerged much later—FotMob in 2004, WhoScored in 2008, and Sofascore in 2010—with most systems active for only the past decade. This temporal precedence, combined with the strong association between journalist bias and algorithmic residuals, supports the interpretation that algorithmic systems inherited and formalized pre-existing human biases. By learning from or being calibrated to biased human inputs, algorithms transform informal discrimination into persistent, reproducible, and scalable disparities.

7 Labor Market Consequences and Systemic Discrimination

Do biased algorithmic ratings lead to economic inequality? Recent work on systemic discrimination shows that discrimination at one decision node can propagate into downstream disparities even when later stages apply neutral rules (Bohren, Hull and Imas, 2025).

We model the football labor market as a two-node system: (i) match-level evaluations (Node 1), which produce season-average ratings, and (ii) wage-setting and market valuations (Node 2), which act on those signals. Our DiDC evidence shows that Dark-skinned players receive smaller boosts for marginal goals. If these biased ratings flow into market outcomes, they can generate systemic discrimination without requiring bias at Node 2.

7.1 Data and Empirical Strategy

We focus on players who attempted at least one shot within the RDD bandwidth during 2020/21 to 2022/23. We aggregate to the player-season level and use three key variables: (i) a constructed distance-weighted near-goals, which we compute as our instrument; (ii) season-average player ratings from FotMob; and (iii) performance controls (non-RDD goals, assists, minutes played), also sourced directly from FotMobs player-season data. Distance-weighted near-goals provide quasi-random variation in seasonal ratings, capturing luck-driven marginal goals, while the other variables reflect the platform's own end-of-season summaries.

We match this dataset to wage (Capology/FBref) and market value (Transfermarkt) data using fuzzy string matching across league-season cells.[13] Final samples include 1,067 player-seasons with valuations and 1,095 with wages.

We estimate two-stage least squares. The first stage instruments season-average ratings with distance-weighted near-goals. The second stage regresses the season-to-season change in (log) market value or (log) weekly wage on instrumented ratings:

$$\begin{aligned} \text{First stage:} \quad \text{Rating}_{is} &= \alpha_{st(i)} + \alpha_{n(i)} + \alpha_{p(i)} + \pi_1 \text{NearGoals}_{is} + \mathbf{X}_{is}\boldsymbol{\gamma} + \varepsilon_{is} \\ \text{Second stage:} \quad \Delta\text{Outcome}_{is+1} &= \alpha_{st(i)} + \alpha_{n(i)} + \alpha_{p(i)} + \beta_1 \widehat{\text{Rating}}_{is} + \mathbf{X}_{is}\boldsymbol{\gamma} + \nu_{is} \end{aligned} \tag{6}$$

where $\Delta\text{Outcome}_{is+1}$ is the first difference in (log) market value or wage; Rating_{is} is the season-average rating; NearGoals_{is} is the instrument; $\mathbf{X}_{is}$ includes non-RDD goals, assists, minutes, age,

[13] We obtain a 99.7% overall match rate across 3,191 player-seasons: 94.7% exact matches, 3.2% fuzzy matches (Jaro-Winkler ≤ 0.15), and 2.0% team-validated matches.

and age squared; and fixed effects control for team×season, nationality, and position. Ratings are in levels to ensure correct timing between s (performance) and $s + 1$ (outcomes). Standard errors cluster at the player level.

We also test for differential treatment at Node 2 by interacting ratings and instruments with skin tone (ITA). A null interaction implies ratings are converted into compensation equally across groups, suggesting discrimination arises solely from biased signals (Node 1). A significant interaction would imply additional bias at the labor market stage (Node 2).

Our strategy relies on two key identifying assumptions. First, the relevance restriction requires that quasi-random goals—measured through distance-weighted near-goals—significantly affect season-average ratings. We document this strong first-stage relationship, consistent with our prior match-level evidence that narrowly scoring a goal increases match-level algorithmic evaluations.

Second, the exclusion restriction assumes that distance-weighted near-goals affect market valuations or wages only through their effect on player ratings, and not through any independent channel. This assumption is not directly testable. However, it is plausible in our setting. While quasi-random goals visibly shift match-level ratings—and, by aggregation, season-level ratings—they are rarely used directly in valuation or wage-setting processes. Market actors and platforms typically rely on summary statistics and crowd-sourced season-level metrics, not granular shot-level data. Thus, the exclusion restriction is more likely to hold: these marginal goals shape perception via ratings but are not salient enough to directly influence player valuation in their own right. Nonetheless, we acknowledge that any IV strategy rests on an untestable assumption, and we interpret our estimates with appropriate caution.

7.2 Results

Table 3 shows that quasi-random near-goals significantly raise season ratings. These higher ratings, in turn, causally increase market valuations: a 1-SD increase in ratings raises a players market value by 16.1% (Panel A, col 3). This relationship holds after adjusting for performance controls and using within-player differencing.

Panel B introduces skin tone interactions. The valuation response to ratings is statistically indistinguishable across players with different ITA scores, indicating that all players—regardless of skin tone—convert ratings into market value at the same rate. A 1-SD increase in rating raises valuation by roughly 16% for both very dark and very light players.

Importantly, these market valuations originate from fan-driven platforms such as Transfermarkt, where crowd perceptions influence a players estimated value (Smith, 2021; Coates and Parshakov, 2022). Although subjective, these values increasingly serve as reference points in actual transfer negotiations. Thus, while fans may not directly apply discriminatory valuation rules, they anchor their beliefs in the same biased performance signals produced by algorithmic and human evaluators. In this sense, unequal recognition at the evaluation stage compounds into downstream inequality through reputational and economic channels.

By contrast, we find no evidence that ratings affect wages (Panel A, col 6), nor any differential wage returns by skin tone (Panel B, col 6). These null effects likely reflect institutional and contractual rigidities in wage-setting. Unlike market valuations, which are continuously updated

Table 3: Labor Market Consequences of Algorithmic Bias

	Δ Market Value (log)			Δ Weekly Wage (log)		
	First Stage		Second	First Stage		Second
	Rating (z)	Rating×ITA (z)	Δ Value	Rating (z)	Rating×ITA (z)	Δ Wage
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Without Skin Tone Interactions						
Near-goals weighted (z)	0.218			0.227		
	(0.032)			(0.031)		
Season rating (z)			0.161			−0.092
			(0.061)			(0.133)
Observations	1,067		1,067	1,095		1,095
R^2	0.792		0.372	0.790		0.345
First-stage F	64.5		64.5	72.4		72.4
Panel B: With Skin Tone Interactions						
Near-goals weighted (z)	0.217	0.060		0.226	0.069	
	(0.031)	(0.048)		(0.031)	(0.046)	
Near-goals×ITA (z)	0.028	0.461		0.025	0.434	
	(0.021)	(0.066)		(0.020)	(0.065)	
Skin tone (ITA, z)	−0.023	0.051	0.002	−0.022	0.074	0.036
	(0.031)	(0.091)	(0.014)	(0.030)	(0.087)	(0.034)
Season rating (z)			0.159			−0.074
			(0.062)			(0.140)
Rating×ITA (z)			0.005			−0.056
			(0.024)			(0.059)
Observations	1,067	1,067	1,067	1,095	1,095	1,095
R^2	0.792	0.463	0.373	0.791	0.455	0.343
First-stage F (Rating)	32.9		32.9	36.8		36.8
First-stage F (Rating×ITA)		78.5	78.5		74.5	74.5
Team×Season FE	✓	✓	✓	✓	✓	✓
Nationality FE	✓	✓	✓	✓	✓	✓
Position FE	✓	✓	✓	✓	✓	✓
Performance controls	✓	✓	✓	✓	✓	✓

Notes: Player-season level analysis (seasons 2020/21–2022/23). Dependent variables: First difference in log market value (cols 1–3) and log weekly wage (cols 4–6). Instruments: distance-weighted near-goals (z-score) and its interaction with ITA (z-score). Panel A reports models without skin tone interactions; Panel B adds continuous ITA interactions (higher = lighter). Ratings remain in levels to align season s performance with season $s+1$ outcomes. Controls include non-RDD goals, assists, minutes played, age, and age^2. All models include team×season, nationality, and position fixed effects. Standard errors clustered at player level.

and partially driven by crowd sentiment, wages are governed by multi-year contracts and internal pay structures that insulate them from marginal performance updates. Consistent with classic models of competitive labor markets (Becker, 1957), and consistent with recent evidence (Principe and van Ours, 2022), firms have limited incentives to discriminate when wages are negotiated infrequently, observed productivity is high, and competition for talent is strong.

The divergence between wages and market valuations also suggests different mechanisms. Clubs may face legal, reputational, or competitive constraints on explicit wage discrimination. In contrast, market values—reflecting fan preferences and consumer demand—may translate skin tone biases into economic returns through higher merchandise sales, media visibility, or sponsorship appeal. These interpretations remain suggestive, but highlight the importance of distinguishing between formal and informal market signals.

Taken together, these results indicate that labor market disparities arise from distorted inputs—biased evaluations—rather than from differential treatment at the compensation stage. First, the null ITA interaction in market value regressions rules out direct discrimination at Node 2: conditional on ratings, market valuations respond identically across the skin tone distribution. Second, as shown in Section 5, the causal reward for a marginal goal rises monotonically with lighter skin: a four-point increase in standardize ITA (from -2 to $+2$) increases the match-level goal premium by 0.52SD. While this effect operates at the level of individual matches, we show that skin tone is also systematically associated with higher *season-average* ratings.

Using season-level regressions, a four-point standardize ITA difference corresponds to a 0.16-0.25SD gap in season ratings under minimal and opportunity-adjusted specifications (Table B.5), which—combined with a 0.161 semi-elasticity of market value with respect to ratings—implies a 2.5-4.0% valuation gap attributable to unequal recognition. These estimates provide a disciplined calibration of how localized evaluation bias accumulates over a season, rather than a mechanical extrapolation from a single event. Third, we find no evidence of statistical discrimination: if clubs discounted the information content of ratings for Dark-skinned players, valuation responses would be attenuated for those players, yet the rating semi-elasticity is constant across ITA

In sum, labor markets faithfully transmit biased performance signals. Crucially, these distortions originate not only from algorithmic evaluations, but also from journalist-assigned ratings. This reinforces a central insight: discrimination enters at the point of evaluation, not compensation.

8 Conclusion

This paper provides causal evidence of skin tone discrimination in professional football, a setting in which performance is highly observable and precisely measured. Exploiting quasi-random variation in marginal goals and a machine-derived skin tone scale, we show that Light-skinned players receive significantly larger boosts in match ratings than Tan- or Dark-skinned players for identical actions. These disparities arise in both algorithmic and journalist assessments and are concentrated in the residual, non-performance component of ratings, suggesting that algorithms formalize rather than originate human biases.

These distorted evaluations have real economic consequences. Season-average ratings, biased by skin tone, causally predict subsequent changes in market valuations. Crucially, we find no evidence of additional discrimination at the wage- or valuation-setting stage: labor markets respond to ratings, not skin tone directly. Moving from very dark to very light skin tone increases valuations by 2.5–4% for equivalent performance, entirely due to unequal recognition. While wages are contractually rigid and largely unresponsive to marginal updates, crowd-sourced valuationsdriven in part by fan demandamplify these disparities through market dynamics.

Beyond professional football, our findings speak to labor markets in which workers are first filtered through evaluative scores, ratings, or grades that serve as inputs into downstream decisions. In different hiring processes, algorithmic screening tools increasingly rank or filter applicants before human review, often based on automated assessments of résumés, video interviews, or behavioral signals (Dastin, 2018; The Economist, 2018; Gregg, 2025; Kessler, 2025; **?**). As in our setting, these scores function as performance signals that shape later outcomes. Our results highlight a general mechanism: bias introduced at an early evaluation stage can propagate through otherwise neutral market processes, generating inequality even in the absence of discriminatory labor demand. Just as biased match-level ratings in football accumulate into season-level valuation gaps without additional bias at the pricing stage, biased screening scores in hiring can translate into employment disparities through formally impartial downstream decisions.

While our setting offers rare observational precision, it also underscores important limitations. We cannot directly observe how evaluators form or update biased beliefs over time, nor how fan perceptions interact with algorithmic assessments. These remain key questions for future research. Our findings suggest that mitigating discrimination in rating-based labor markets may require interventions at the evaluation stage itself, including the development of non-discriminatory algorithms (Arnold et al., 2025) and tools aimed at reducing bias among human raters (Webb, 2025). More broadly, while we provide causal evidence of colorism in a high-profile, data-rich environment, future work should examine whether similar dynamics arise in other settings where performance is summarized into evaluative scores. Colorism, long documented in correlational studies (Woo-Mora, 2026), warrants continued scrutiny using experimental and quasi-experimental designs across diverse institutional contexts.

References

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*

Angrist, J. D. and Pischke, J.-S. (2009), *Mostly harmless econometrics: An empiricist’s companion*, Princeton University press.

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. (2021), ‘Measuring racial discrimination in algorithms’, *AEA Papers and Proceedings* **111**, 49–54.

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*

Arnold, D., Dobbie, W. and Hull, P. (2025), ‘Building nondiscriminatory algorithms in selected data’, *American Economic Review: Insights* **7**, 231–49.

Arnold, D., Dobbie, W. and Yang, C. S. (2018), ‘Racial bias in bail decisions’, *The Quarterly Journal of Economics* **133**, 1885–1932.
URL: *https://dx.doi.org/10.1093/qje/qjy012*

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.

Blinder, A. S. (1973), ‘Wage discrimination: Reduced form and structural estimates’, *The Journal of Human Resources* **8**(4), 436–455.
URL: *http://www.jstor.org/stable/144855*

Bohren, J. A., Haggag, K., Imas, A. and Pope, D. G. (2025), ‘Inaccurate statistical discrimination: An identification problem’, *The Review of Economics and Statistics* **107**, 605–620.
URL: *https://dx.doi.org/10.1162/rest*$_{a0}$1367

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*

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.

Dastin, J. (2018), ‘Amazon scraps secret AI recruiting tool that showed bias against women’, *Reuters* .
URL: *https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G*

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* .

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.

Gauriot, R. and Page, L. (2019), ‘Fooled by performance randomness: Overrewarding luck’, *The Review of Economics and Statistics* **101**, 658–666.
URL: *https://dx.doi.org/10.1162/rest*$_{a0}$0783

Goldin, C. and Rouse, C. (2000), ‘Orchestrating impartiality: The impact of b̈lindäuditions on female musicians’, *American Economic Review* **90**, 715–741.

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.

Gregg, A. (2025), ‘Why you shouldn’t count on humans to prevent AI hiring bias’, *The Washington Post* .
URL: *https://www.washingtonpost.com/business/2025/11/25/biased-ai-hiring-research-university-of-washington-study/*

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*

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

James, S. and Harpur, C. (2025), ‘8/10 for messi, 9/10 for mbappe, 3/10 for griezmann: Lequipe and the mad world of football ratings’, *The Athletic* .
URL: *https://www.nytimes.com/athletic/3219272/2022/04/07/2-10-for-pulisic-3-10-for-messi-inside-lequipe-and-the-mad-world-of-football-ratings/*

Kessler, S. (2025), ‘Employers are buried in A.I.-generated résumés’, *The New York Times* . DealBook Newsletter.
URL: *https://www.nytimes.com/2025/06/21/business/dealbook/ai-job-applications.html*

Kitagawa, E. M. (1955), ‘Components of a difference between two rates’, *Journal of the American Statistical Association* **50**, 1168.

Kleinberg, J., Ludwig, J., Mullainathan, S. and Rambachan, A. (2018), ‘Algorithmic fairness’, *AEA Papers and Proceedings* **108**, 22–27.

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*

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*

Oaxaca, R. (1973), 'Male-female wage differentials in urban labor markets', *International Economic Review* **14**, 693.

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*

Palacios-Huerta, I. (forthcoming), 'The beautiful dataset', *Journal of Economic Literature* .

Panenka Magazine (2023), 'Dossier: Racismo', Panenka, Issue #126. Special dossier on racism in football.

Pathak, M. (2025), 'How fotmob app went from a honeymoon idea to 20 million monthly users', *Forbes* .
URL: *https://www.forbes.com/sites/manasipathak/2025/08/09/how-fotmob-app-went-from-a-honeymoon-idea-to-20-million-monthly-users*

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*

Price, J. and Wolfers, J. (2010), 'Racial discrimination among nba referees', *The Quarterly Journal of Economics* **125**, 1859–1887.
URL: *https://dx.doi.org/10.1162/qjec.2010.125.4.1859*

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.

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

Sarsons, H. (2022), Interpreting signals in the labor market: Evidence from medical referrals. Working paper.

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*

Szymanski, S. (2000), 'A market test for discrimination in the english professional soccer leagues', *Journal of Political Economy* **108**, 590–603.

Taylor, D., Madley, S. and Fifield, D. (2021), 'The different faces of racism', *The Athletic* . Updated November 5, 2021.
URL: *https://www.nytimes.com/athletic/2813763/2021/10/12/the-different-faces-of-racism/*

The Economist (2018), 'How an algorithm may decide your career', *The Economist* . Bartleby column.
URL: *https://www.economist.com/business/2018/06/21/how-an-algorithm-may-decide-your-career*

Webb, D. (2025), Silence to solidarity: How communication about a minority affects discrimination. Conditionally accepted at the *Journal of Political Economy.*

Wilhelmsen, P. N. (2025), 'I bergen sitter to brødre som holder en hel verden oppdatert om fotballresultater', *Nettavisen* .
URL: *https://www.nettavisen.no/okonomi/fotmob-den-norske-suksesshistorien-som-millioner-bruker/s/5-95-1410176*

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

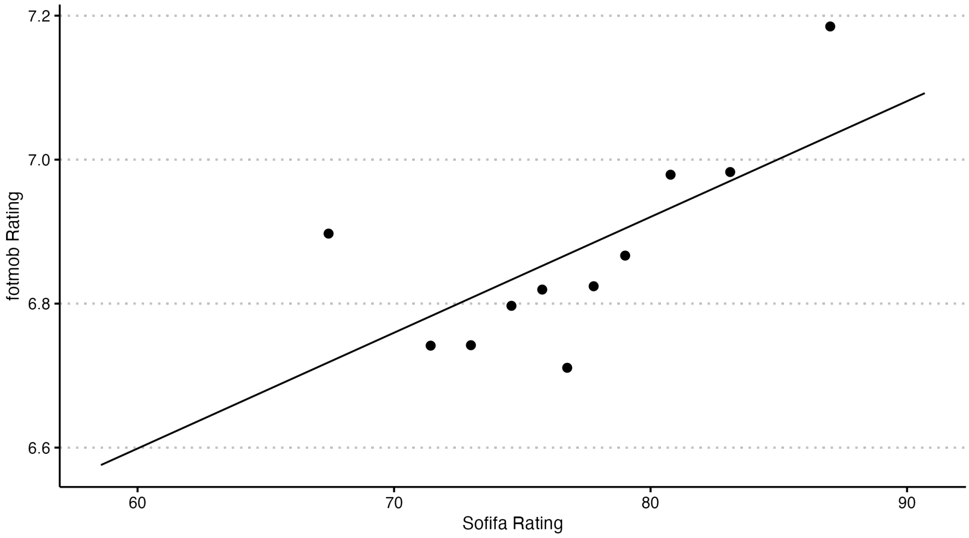
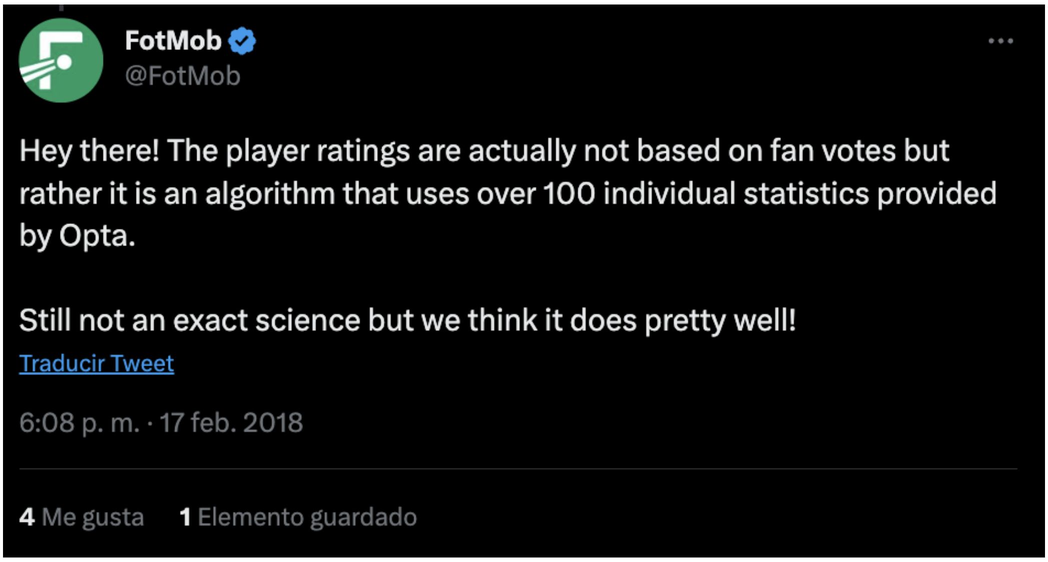


Figure A.1: Correlation Between Algorithmic (FotMob) and Fan-Based (Sofifa) Player Ratings

Notes: This figure shows a binned scatterplot of average FotMob ratings against Sofifa ratings for a sample of professional football players. Sofifa aggregates fan-based ratings used in EA Sports video games, while FotMob ratings are generated algorithmically from match-level performance data. The strong positive relationship suggests that algorithmic assessments reflect similar relative rankings to those formed by broader fan communities.



FotMob
@FotMob

Hey there! The player ratings are actually not based on fan votes but rather it is an algorithm that uses over 100 individual statistics provided by Opta.

Still not an exact science but we think it does pretty well!

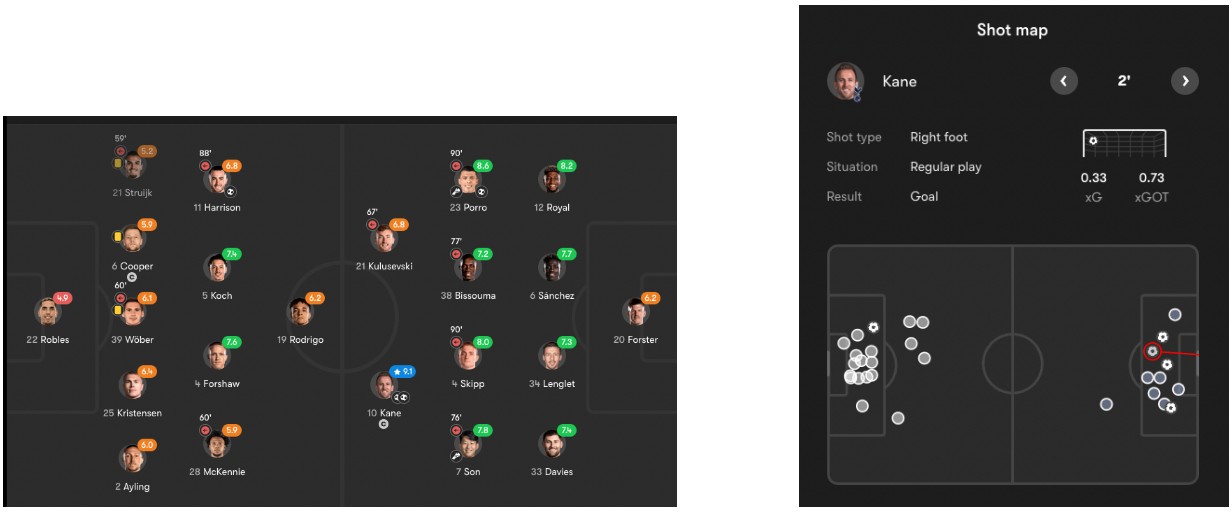
Traducir Tweet

6:08 p. m. · 17 feb. 2018

4 Me gusta 1 Elemento guardado

Figure A.2: 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.3: 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.

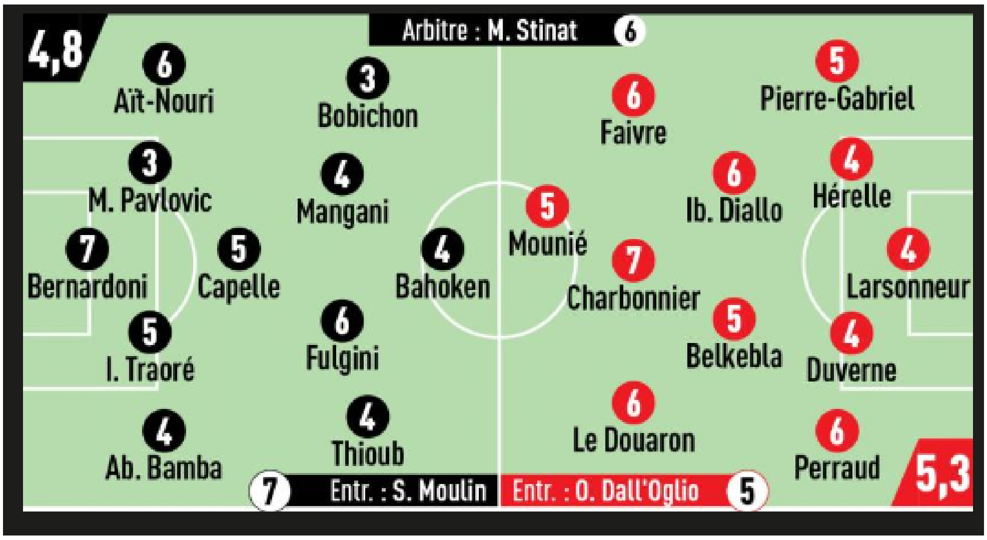


Figure A.4: Example of Post-Match Journalist Ratings published by L'Équipe

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.

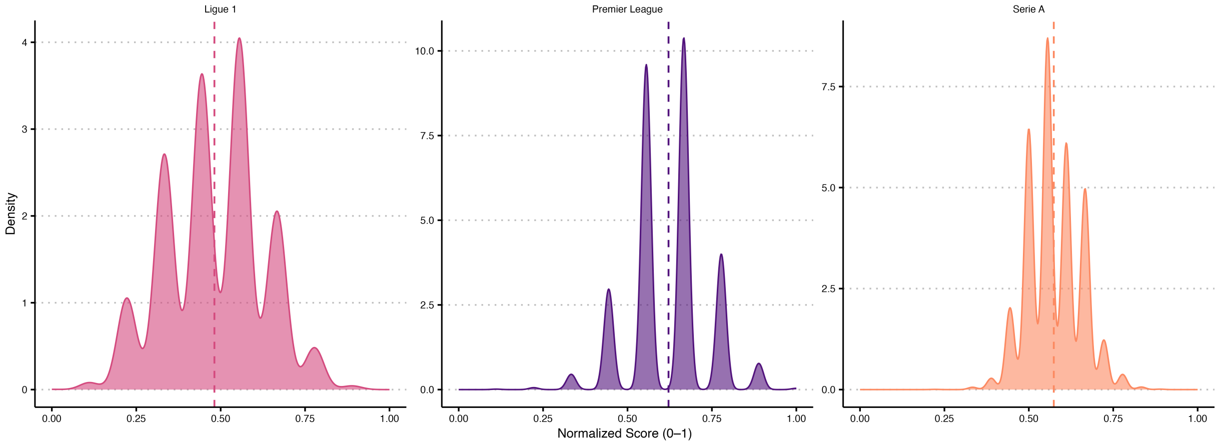


Figure A.5: Distributions of Journalist Ratings by League

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.

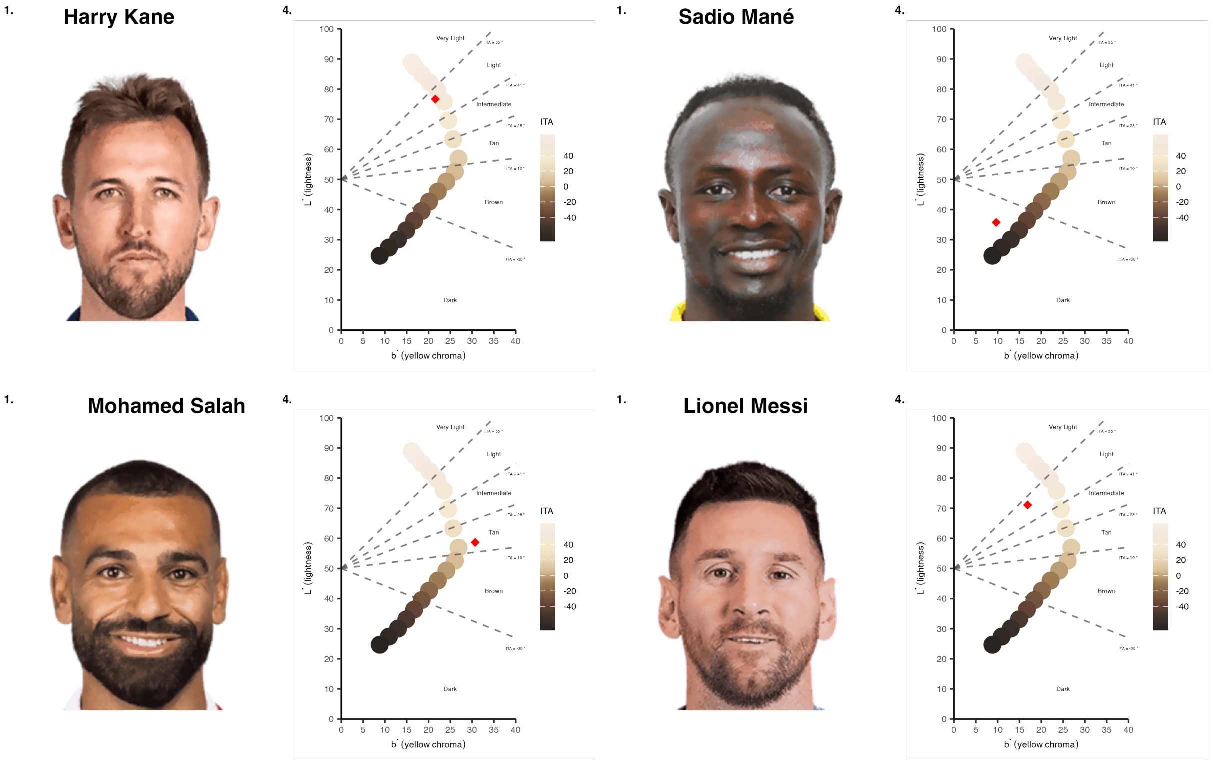
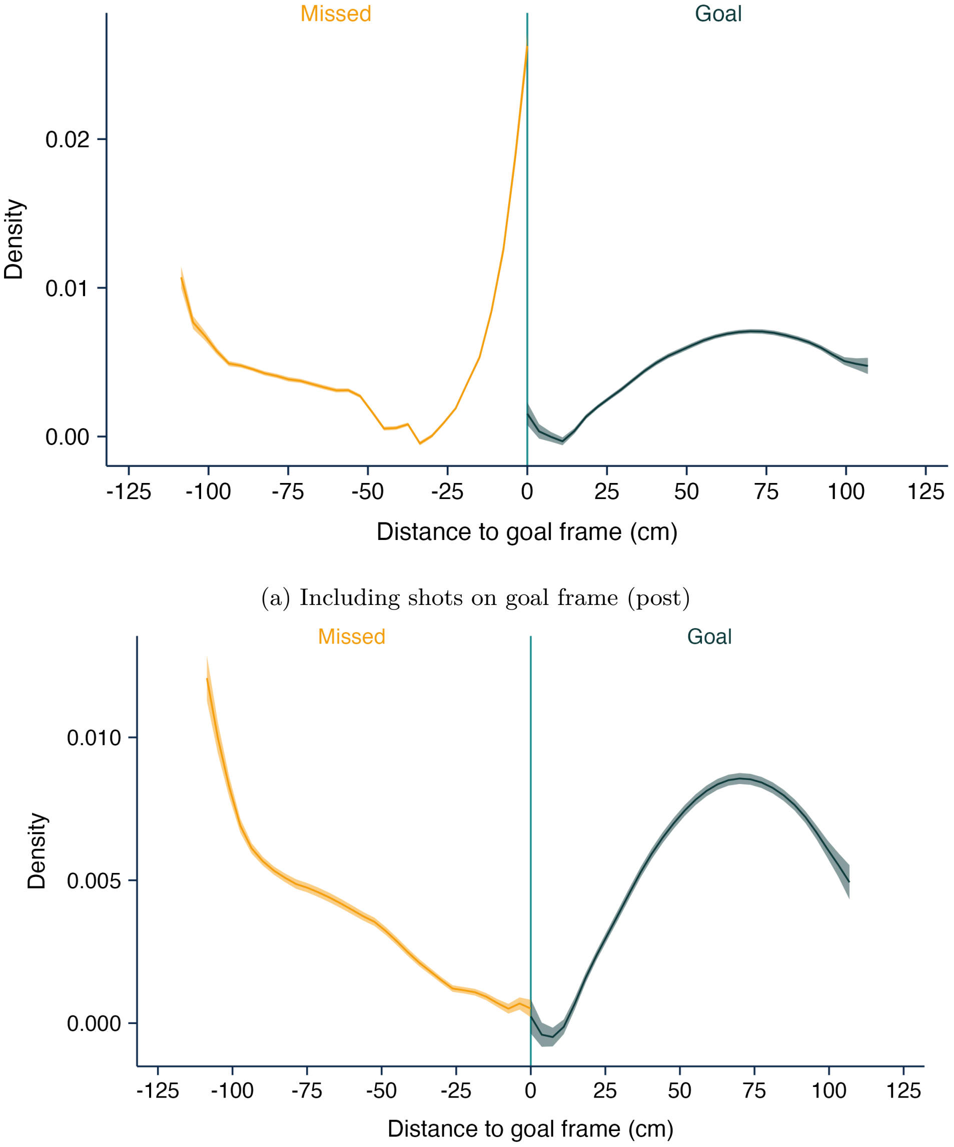


Figure A.6: Skin tone classification through ITA groupings for selected players

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.

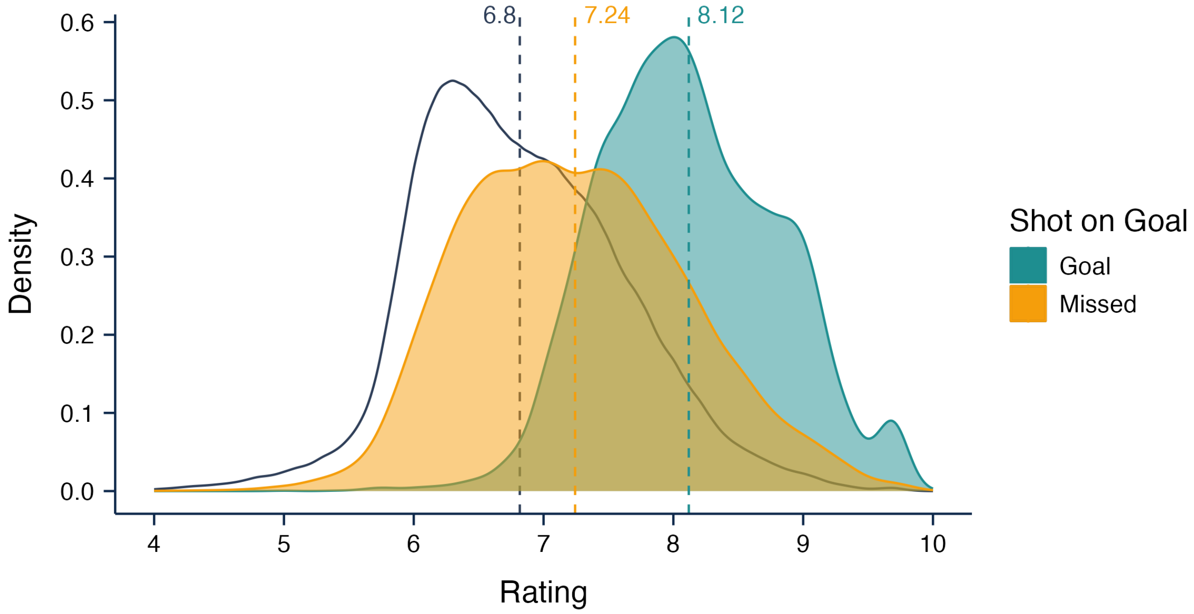


Figure B.2: Post-match Ratings Distribution: Scoring Goal Premium

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 took a shot but did not score (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 did not score or "missed" group includes only players whose shot missed the goal frame (e.g., off-target or hit the post) and excludes saved shots. Ratings for non-scoring players closely resemble the overall distribution, while those for goal scorers are shifted to the right, with a higher mean (8.12 vs. 7.24). This visual evidence suggest 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	0.515	0.064	0.388	0.641	0.000	1	68.2	114.601	8839	16987	Triangular	HC3
Log points	0.069	0.008	0.053	0.085	0.000	1	68.2	114.601	8839	16987	Triangular	HC3
Z-score	0.576	0.072	0.434	0.717	0.000	1	68.2	114.601	8839	16987	Triangular	HC3
Z-score + Clean controls	0.554	0.067	0.422	0.685	0.000	1	68.2	114.601	8839	16987	Triangular	HC3
Z-score + DLASSO controls	0.590	0.069	0.455	0.725	0.000	1	68.2	114.601	8839	16987	Triangular	HC3
Z-score + DLASSO controls	0.423	0.126	0.176	0.670	0.001	2	68.2	114.601	8839	16987	Triangular	HC3
Z-score + DLASSO controls	0.590	0.069	0.455	0.725	0.000	1	68.2	114.601	8839	16987	Triangular	HC3
Z-score + DLASSO controls	0.603	0.067	0.471	0.734	0.000	1	68.2	114.601	8839	16987	Epanechnikov	HC3
Z-score + DLASSO controls	0.684	0.061	0.565	0.802	0.000	1	68.2	114.601	8839	16987	Uniform	HC3
Z-score + DLASSO controls	0.467	0.152	0.169	0.765	0.002	1	34.1	57.301	3834	6808	Triangular	HC3
Z-score + DLASSO controls	0.677	0.057	0.566	0.788	0.000	1	136.4	229.202	16987	16987	Triangular	HC3
Z-score + DLASSO + Match FE	0.470	0.085	0.303	0.636	0.000	1	68.2	114.601	8839	16987	Triangular	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.

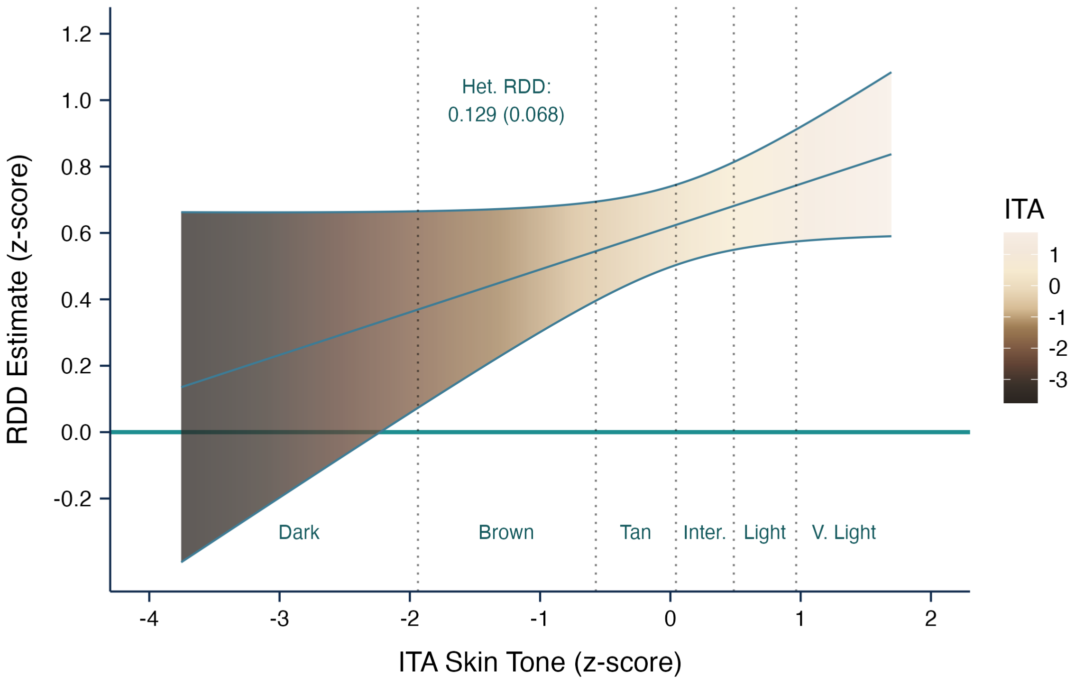


Figure B.3: Continuous Heterogeneous RDD: Goal Premium by ITA Skin Tone (z-score)

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 triangular kernel. The specification includes Double LASSO-selected covariates and are weighted by the inverse number of within-bandwidth shots per player-match.

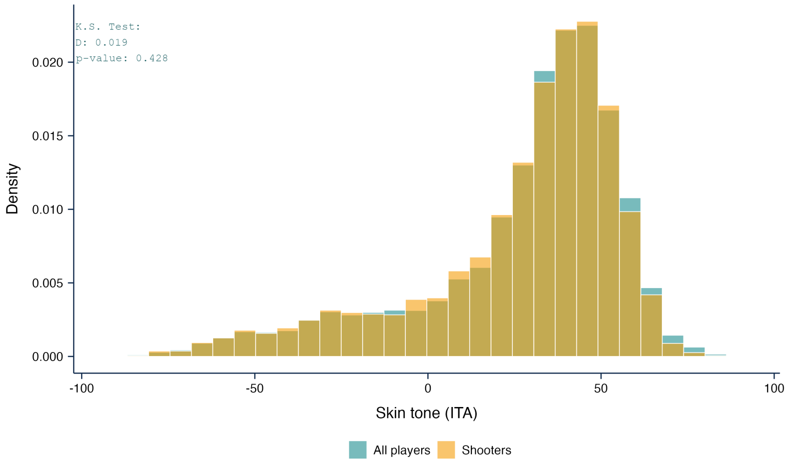


Figure B.4: Skin Tone Distribution: All Players vs. Shooters Sample

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.

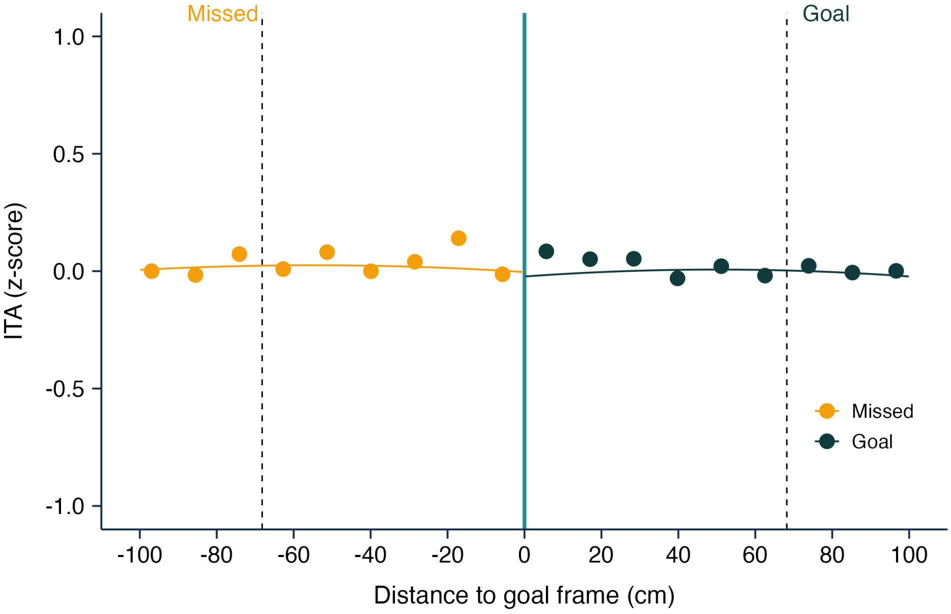


Figure B.5: Regression Discontinuity Test: Skin Tone at the Scoring Threshold

Notes: This figure plots a regression discontinuity design with standardized skin tone (ITA, z-score) as the outcome and Euclidean distance to the goal frame (in centimeters) as the running variable. Zero denotes the scoring threshold, with negative values corresponding to narrowly missed shots and positive values to narrowly scored goals. Dots represent evenly spaced mean bins constructed using evenly spaced bin selection, and lines show local quadratic fits estimated separately on each side of the cutoff using the MSE-optimal bandwidth. Vertical dashed lines indicate the main estimation bandwidth. Observations are weighted by the inverse number of within-bandwidth shots per player–match. The absence of any visible discontinuity indicates that skin tone evolves smoothly at the scoring threshold, supporting the assumption that players are not sorted by skin tone around marginal goals.

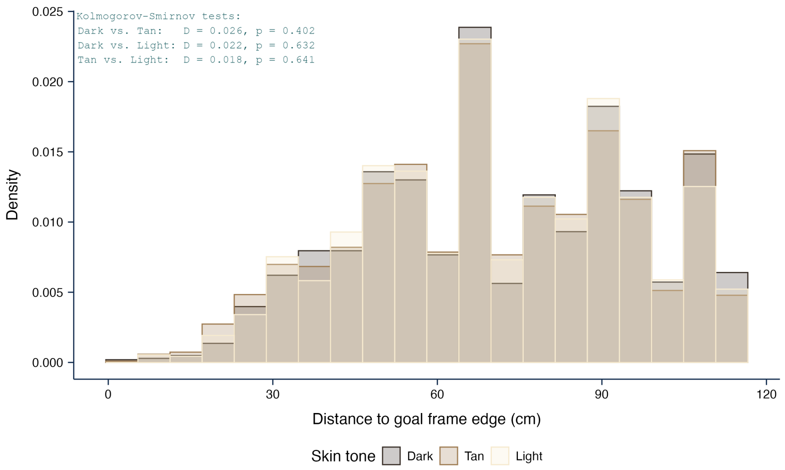


Figure B.6: Distribution of Distance to Goal Frame Edge by Skin Tone

Notes: This figure shows the distribution of distance to the nearest goal frame edge (left post, right post, or crossbar) for successfully scored goals within the RDD bandwidth (±114 cm), by skin tone group. Sample: N = 8,370 goals. Kolmogorov-Smirnov tests fail to reject equality of distributions across all pairwise comparisons, indicating that players of different skin tones aim at statistically identical locations within the goal frame. The overlapping distributions confirm that differential ratings cannot be attributed to systematic differences in shot difficulty.

Table B.2: Heterogeneous Treatment Effects by Skin Tone: Robustness Checks

Specification	Dark (95% CI)	Tan (95% CI)	Light (95% CI)	Dark - Light	Dark - Tan	Tan - Light
Match FE	0.223 [-0.116, 0.562]	0.521 [0.300, 0.743]	0.553 [0.290, 0.815]	-0.330 (p=0.066)	-0.298 (p=0.074)	-0.031 (p=0.429)
Player FE	0.377 [0.079, 0.675]	0.471 [0.273, 0.669]	0.736 [0.477, 0.995]	-0.360 (p=0.037)	-0.095 (p=0.302)	-0.265 (p=0.056)
Epa kernel	0.459 [0.153, 0.765]	0.525 [0.354, 0.696]	0.827 [0.618, 1.036]	-0.369 (p=0.026)	-0.066 (p=0.356)	-0.302 (p=0.014)
Uniform kernel	0.528 [0.253, 0.804]	0.567 [0.398, 0.737]	0.911 [0.699, 1.123]	-0.382 (p=0.016)	-0.039 (p=0.407)	-0.344 (p=0.007)
No posts	1.068 [0.637, 1.498]	1.063 [0.800, 1.325]	1.308 [1.005, 1.611]	-0.240 (p=0.185)	0.005 (p=0.508)	-0.246 (p=0.115)
Aggregated	0.389 [0.035, 0.742]	0.490 [0.293, 0.687]	0.829 [0.565, 1.093]	-0.440 (p=0.025)	-0.101 (p=0.312)	-0.339 (p=0.022)

Notes: This table reports RDD estimates of the goal premium by skin tone group across robustness specifications. Entries show point estimates with 95% confidence intervals in brackets. Final three columns show pairwise differences in recognition (Light vs. Dark, etc.), with one-sided p-values in parentheses.

Table B.3: Balance Tests: Pre-Determined Covariates by Skin Tone

	Dark		Tan		Light	
Covariate	Coef.	q-value	Coef.	q-value	Coef.	q-value
Match minute	0.085	[1.000]	0.090	[0.628]	0.361	[0.249]
Shot Y-coordinate	-0.527	[0.072]	-0.463	[0.003]	-0.265	[1.000]
Shot X-coordinate	0.014	[1.000]	0.186	[0.553]	0.208	[1.000]
Starting XI	-0.040	[1.000]	0.021	[0.628]	-0.024	[1.000]
Home team	-0.008	[1.000]	-0.027	[0.668]	-0.002	[1.000]
Team captain	-0.043	[1.000]	-0.027	[0.628]	-0.032	[1.000]
Stadium attendance	0.283	[1.000]	-0.191	[0.553]	0.064	[1.000]
Penalty kick	0.014	[1.000]	0.058	[0.553]	0.010	[1.000]
Defender	0.012	[1.000]	-0.052	[0.628]	-0.058	[1.000]
Header	-0.041	[1.000]	-0.005	[1.000]	-0.062	[1.000]
Shirt number 10	-0.032	[1.000]	0.069	[0.553]	-0.043	[1.000]
Shirt number 16	-0.008	[1.000]	0.014	[0.628]	-0.013	[1.000]
Free kick	0.007	[1.000]	-0.040	[0.553]	-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). Standard errors are robust and clustered at the player-match level. Bandwidth: 114 cm.

Table B.4: 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	0.453	[0.045]	0.501	[0.000]	0.808	[0.000]
Minutes played	0.113	[1.000]	0.028	[0.786]	-0.027	[1.000]
Accurate passes	0.060	[1.000]	-0.036	[0.882]	0.050	[1.000]
Accurate long balls	-0.487	[0.042]	0.105	[0.672]	-0.258	[0.450]
Recoveries	0.084	[1.000]	0.101	[0.672]	-0.031	[1.000]
Touches	0.187	[1.000]	-0.005	[1.000]	-0.008	[1.000]
Goals	1.254	[0.000]	1.262	[0.000]	1.312	[0.000]
Assists	-0.127	[1.000]	-0.124	[0.533]	0.064	[1.000]
Total shots	-0.029	[1.000]	-0.262	[0.121]	-0.166	[0.836]
Chances created	-0.102	[1.000]	-0.145	[0.454]	0.216	[0.836]
Shot accuracy (%)	0.776	[0.010]	0.991	[0.000]	1.221	[0.000]
Passes into final third	0.021	[1.000]	-0.118	[0.533]	0.058	[1.000]
Accurate crosses	-0.073	[1.000]	0.070	[0.786]	-0.046	[1.000]
Offsides	-0.150	[1.000]	0.083	[0.711]	-0.100	[1.000]
Dispossessed	-0.284	[1.000]	0.069	[0.786]	-0.342	[0.088]
Tackles won	0.230	[1.000]	0.196	[0.358]	0.166	[1.000]
Clearances	0.055	[1.000]	0.226	[0.247]	0.156	[1.000]
Headed clearances	0.022	[1.000]	0.282	[0.143]	0.151	[1.000]
Interceptions	-0.179	[1.000]	0.137	[0.454]	0.035	[1.000]
Ground duels won	-0.043	[1.000]	-0.087	[0.749]	0.031	[1.000]
Aerial duels won	0.500	[0.162]	0.210	[0.290]	-0.069	[1.000]
Was fouled	0.008	[1.000]	0.067	[0.786]	-0.187	[0.912]
Fouls committed	0.163	[1.000]	0.258	[0.203]	0.006	[1.000]
Clearances off line	0.011	[1.000]	-0.004	[0.672]	0.004	[1.000]
Blocks	0.210	[1.000]	0.235	[0.454]	-0.096	[1.000]
Dribbled past	0.045	[1.000]	-0.045	[0.882]	0.225	[0.836]
Corners	-0.104	[1.000]	0.053	[0.876]	0.069	[1.000]
Blocked shots	0.202	[1.000]	-0.048	[0.882]	-0.150	[1.000]
Successful dribbles	-0.031	[1.000]	-0.235	[0.261]	0.134	[1.000]
Big chances missed	-0.894	[0.000]	-0.704	[0.000]	-0.393	[0.030]
Error led to goal	0.037	[1.000]	-0.006	[0.247]	-0.008	[0.308]
Conceded penalty	-0.010	[1.000]	-0.015	[0.121]	0.004	[1.000]
Penalties won	-0.076	[1.000]	-0.125	[0.454]	0.132	[1.000]
Last man tackle	-0.110	[1.000]	-0.018	[0.882]	-0.064	[1.000]
Missed penalty	-0.028	[0.126]	-0.041	[0.021]	-0.029	[0.016]
Own goal	0.005	[1.000]	-0.007	[0.121]	-0.003	[1.000]

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. Robust standard errors clustered at player level. Bandwidth: 114 cm.

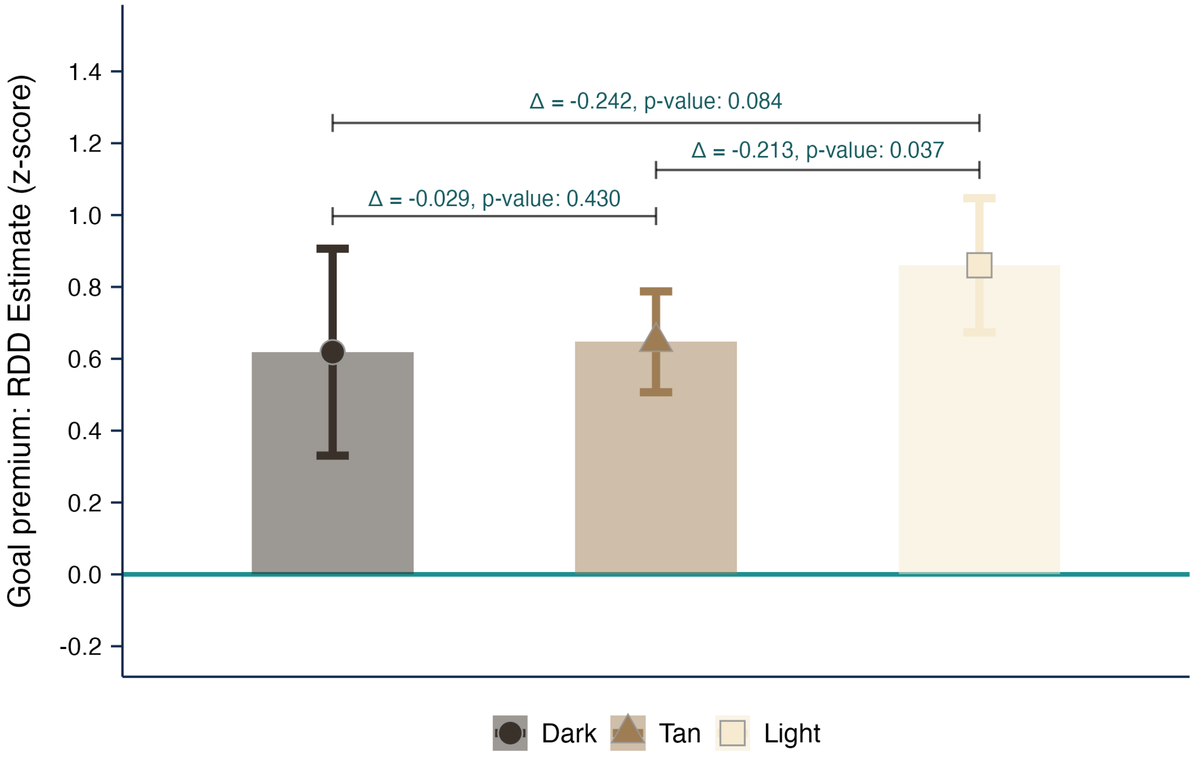


Figure B.7: Difference-in-Discontinuities Controlling for Post-Treatment Performance Metrics

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 total goals scored. Error bars represent 95% confidence intervals, clustered at the player-match level. Brackets indicate pairwise differences (Δ) with one-sided p-values. Estimates are obtained using the `rdhte` package (Calonico et al., 2025*b*), with MSE-optimal bandwidth ($h = 114$ cm), a triangular kernel, and Double LASSO-selected covariates. All models weight observations by the number of shots per player-match.

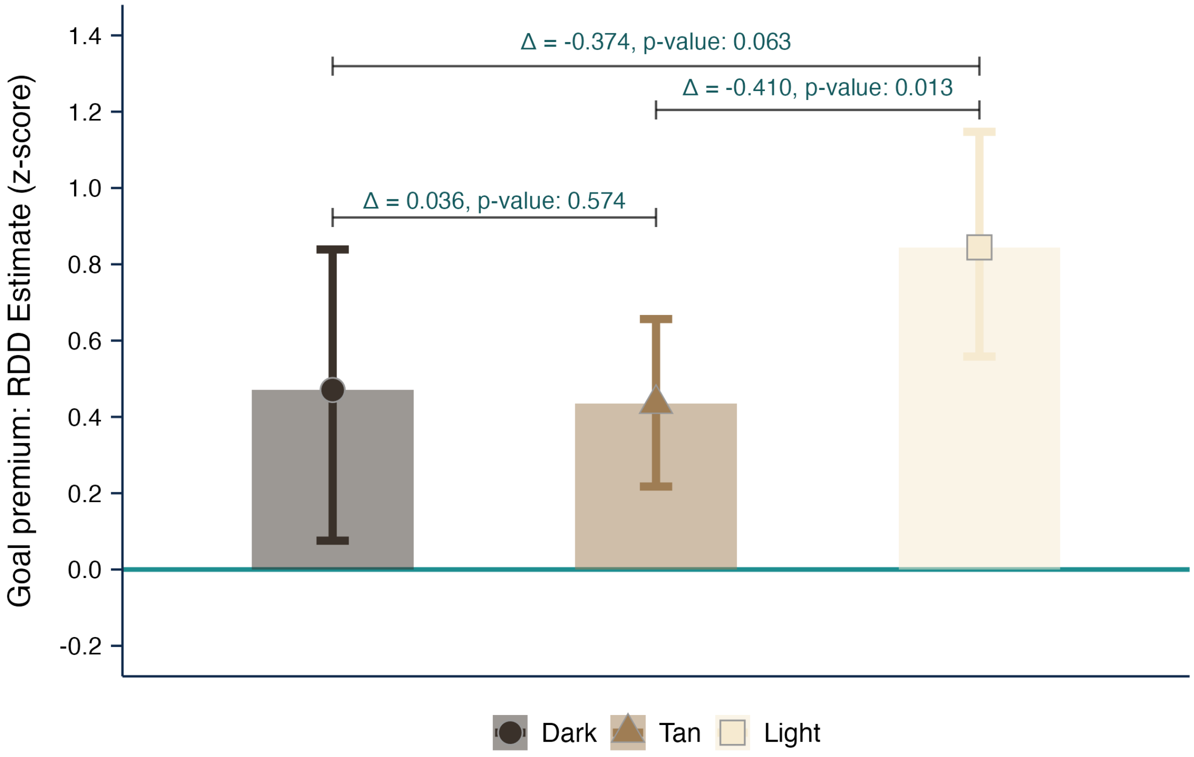
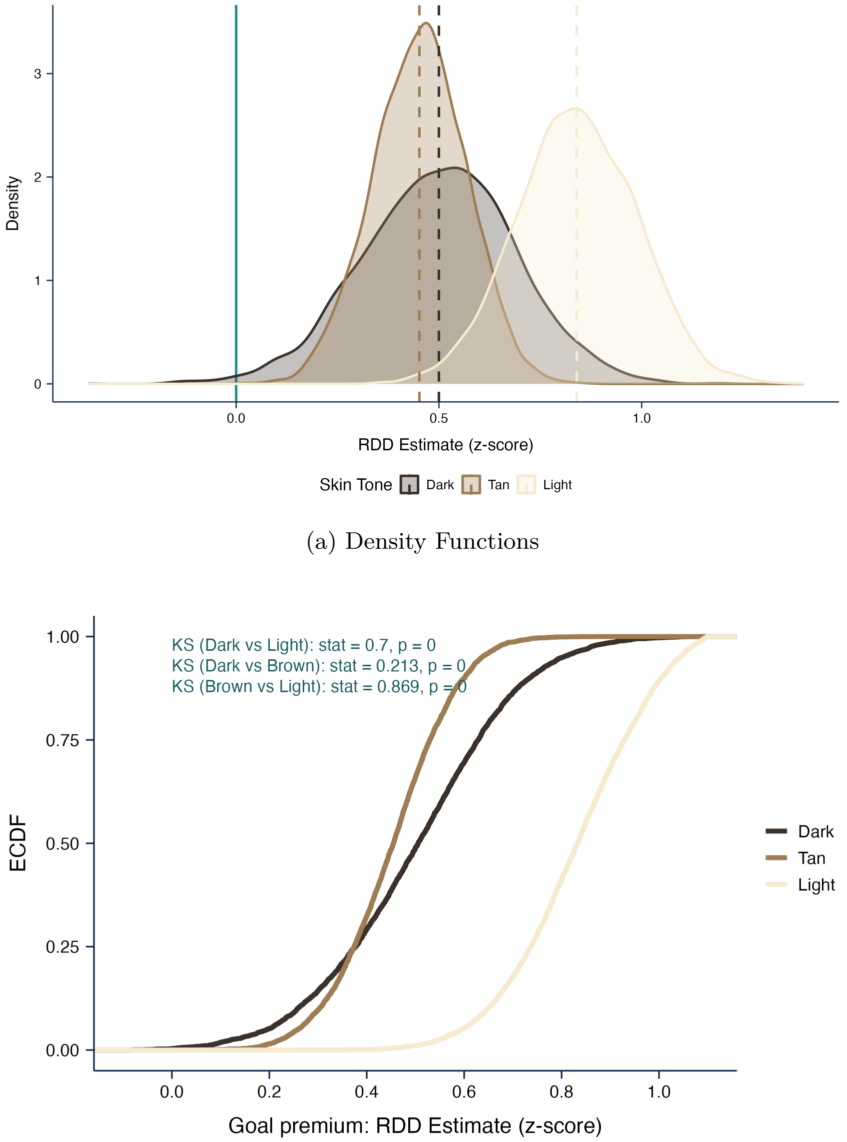


Figure B.8: Bayesian Bootstrap Estimates: Goal Premium by Skin Tone

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 3. All bootstrap samples use the same bandwidth, kernel (triangular), controls (Double LASSO-selected covariates), and weighting scheme (inverse number of within-bandwidth shots per player-match) as the main specification.



(a) Density Functions

(b) Empirical Cumulative Distribution Functions

Figure B.9: 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), triangular 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.

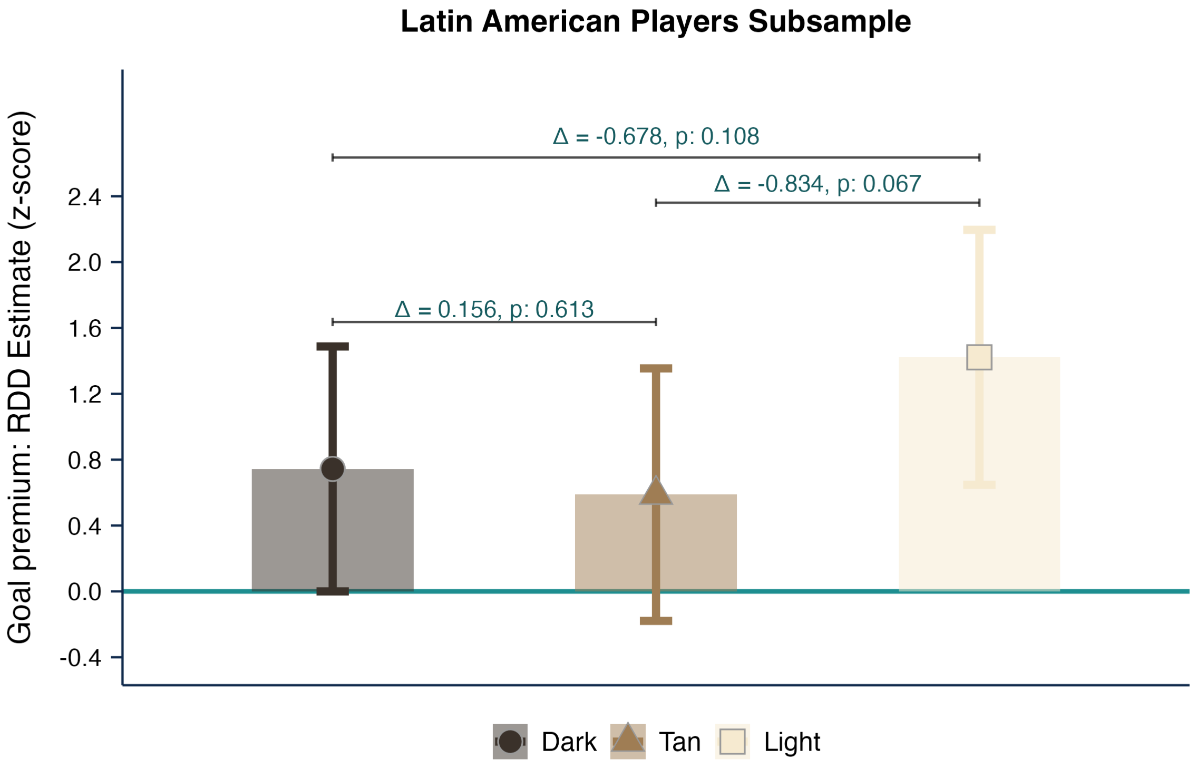


Figure B.10: RDD and DiDC Estimates: Latin America Subsample

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.

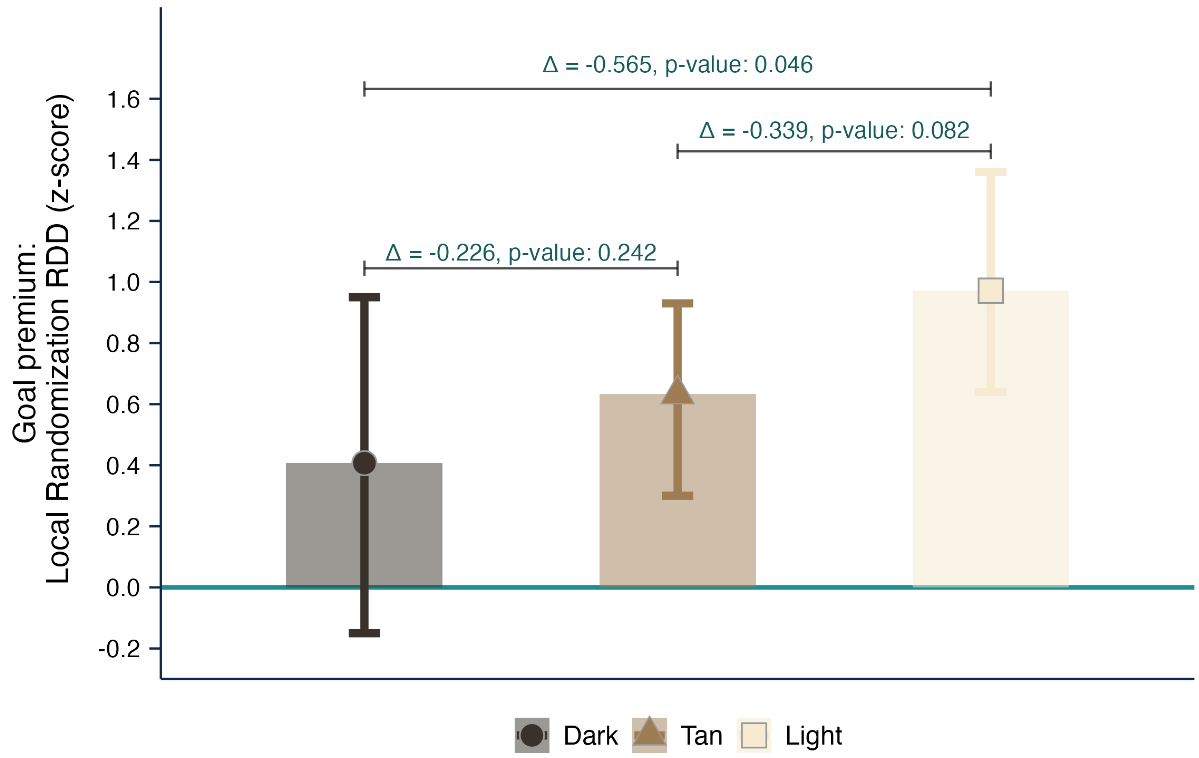


Figure B.11: Local Randomization RDD: Goal Premium by Skin Tone

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 3, 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 ±15 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 randomization distribution of the difference under the null hypothesis of no discrimination (H$_0$: $\Delta \geq 0$ vs. H$_1$: $\Delta < 0$). Standard errors for visualizing these differences 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: $SE \approx (\text{UCI} - \text{LCI})/(2 \times 1.96)$. These SEs approximate $\text{Var}(A - B)$ under the assumption of independence across group estimates and may thus be interpreted as a weak upper bound when positive covariance is present. Importantly, all inference on Δ is based on the exact permutation distribution under the Local Randomization design and is not affected by this approximation.

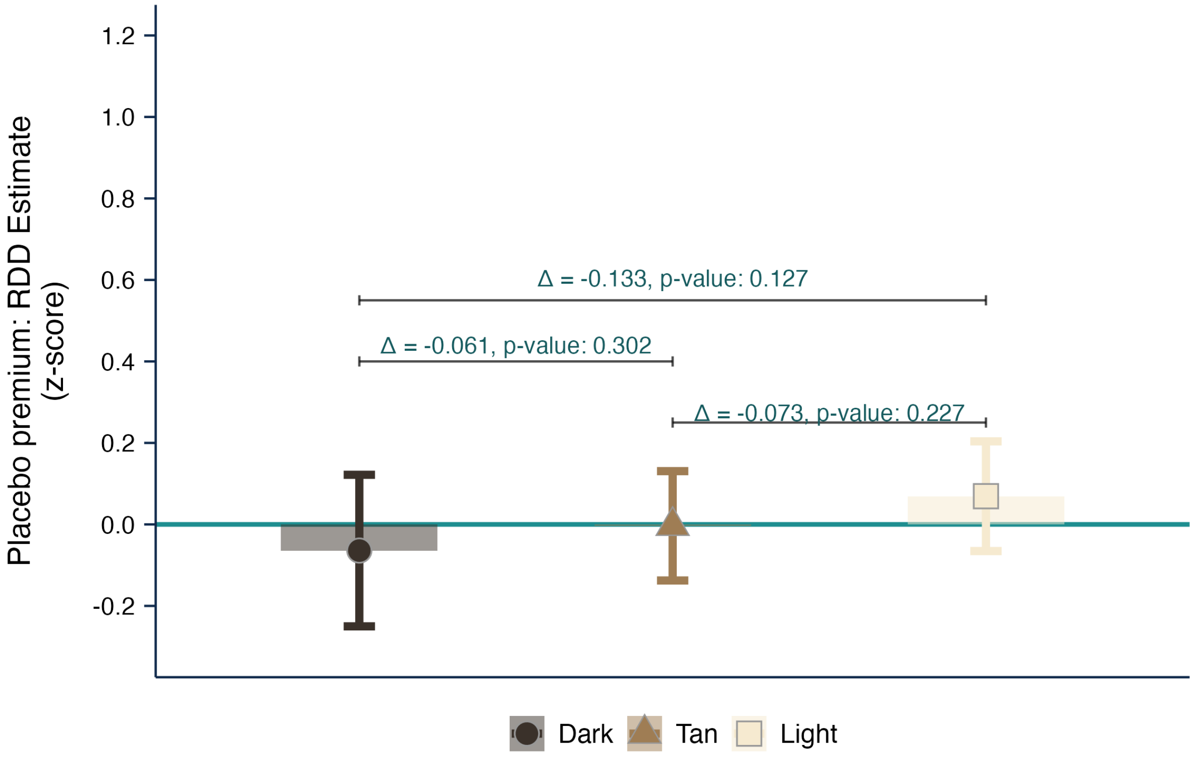


Figure B.12: Placebo Test: Artificial Threshold Among Scored Goals

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 70 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.

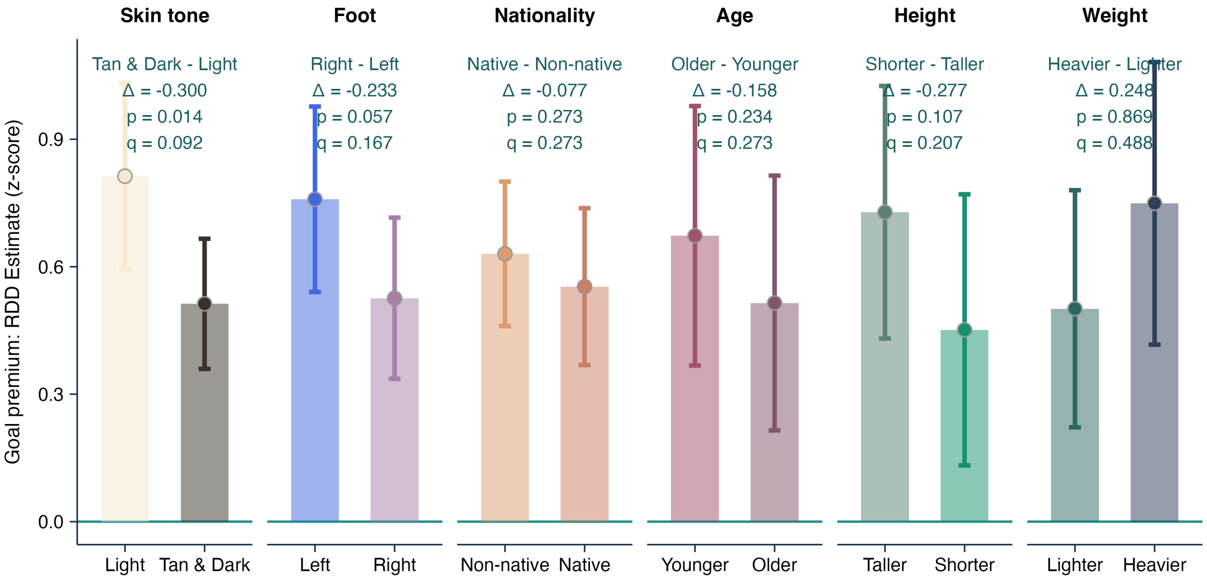


Figure B.13: Heterogeneity Analysis: Goal Premia Across Player Dimensions

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), triangular 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 (Δ) with one-sided *p*-values and FDR-corrected *q*-values computed across all six comparisons using the Storey (2002) method at the 10% level.

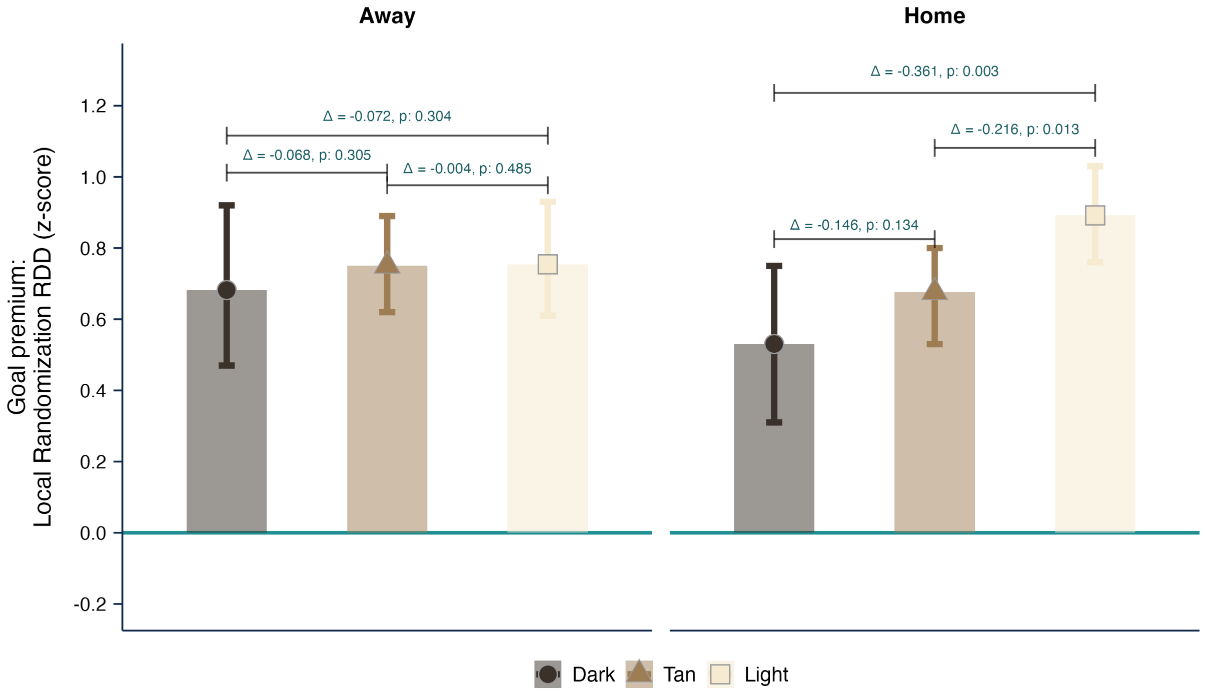


Figure B.14: RDD and Diff-in-Disc Estimates: Away vs. Home Matches.

Notes: The left (right) panel shows the estimated coefficients for the subsample of matches where the player's team played an away (home) game. Estimates are derived using the Local Randomization RDD framework.

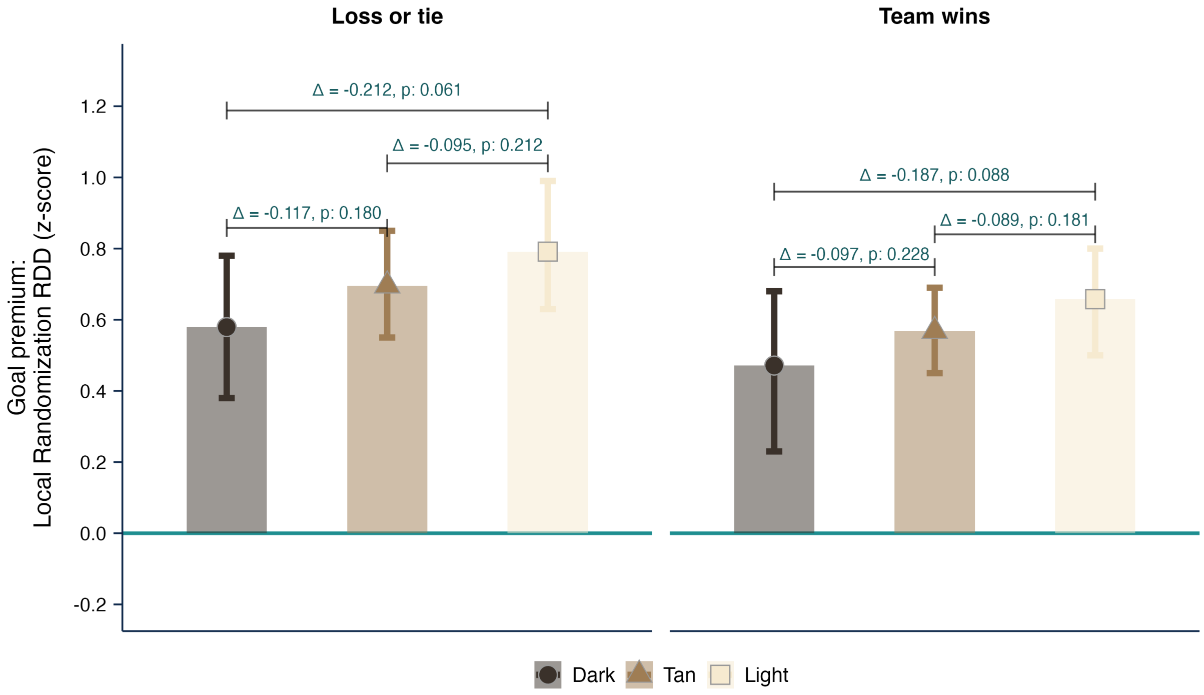


Figure B.15: RDD and Diff-in-Disc Estimates: Match Result

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.

Table B.5: Estimated Season-Level Rating and Valuation Gaps by Skin Tone

Model	ITA Coef. (SE)	Rating Gap (SD)	Valuation Gap (%)	N
(1) Minimal Controls	0.062 (0.023)	0.249	4.00	2,905
(2) Opportunity Controls	0.039 (0.016)	0.157	2.53	2,905
(3) Full Controls	0.006 (0.017)	0.026	0.42	2,231

Notes: This table reports coefficients from regressions of standardized season-average algorithmic ratings on standardized ITA skin tone (higher = lighter), with effects scaled over the full observed ITA range (from -2 to $+2$). The *Rating Gap (SD)* column shows the corresponding difference in season-average ratings in standard deviations. The *Valuation Gap (%)* multiplies this difference by the 0.161 semi-elasticity of market value with respect to ratings (Table 3, col 3). Model (1) includes team×season, position, and nationality fixed effects only. Model (2) adds opportunity controls: minutes played and matches played. Model (3) adds full performance controls: goals, assists, expected goals (xG), and total shots. Standard errors are clustered at the player level.