Oversight and the Expression of Ethnic Bias: Evidence from Amateur Football

Abstract

Does institutional oversight suppress the expression of ethnic bias? We study this question in the context of German amateur football, where referees are routinely observed and graded by external evaluators. Drawing on a new panel of roughly 750,000 matches, 380,000 players, and 14,000 referees spanning more than a decade (2013/14–2025/26), we first document—in line with prior research—that non-white amateur players receive significantly more disciplinary cards than white players. To causally identify the effect of oversight on this ethnic card gap, we use a triple-difference design that leverages variation in referee oversight generated by club promotions and relegations across league tiers. Across a range of alternative designs and specifications, we find that external oversight substantially narrows the ethnic card gap. Our findings speak to a class of institutional remedies for discrimination that suppress biased behavior without changing underlying attitudes.

Keywords: ethnic bias; institutional oversight; discrimination; refereeing; amateur football.

Word count: 6,594 (excluding bibliography).

1 Introduction

A large and growing literature documents that ethnic discrimination persists in many social contexts despite formal commitments to equal treatment. Ethnic biases have been documented in labor markets (Bertrand and Mullainathan, 2004; Quillian et al., 2017, 2019; Kline, Rose and Walters, 2022), the criminal-justice system (Arnold, Dobbie and Yang, 2018; Goncalves and Mello, 2021; Hoekstra and Sloan, 2022; Riaz and Hamjediers, 2025), sports officiating (Price and Wolfers, 2010; Parsons et al., 2011; Magistro and Wack, 2023), healthcare (Hoffman et al., 2016; Greenwood et al., 2020), housing (Christensen and Timmins, 2022), the responsiveness of elected officials and election administrators (Butler and Broockman, 2011; White, Nathan and Faller, 2015), and everyday consumer encounters (Mujcic and Frijters, 2021; Edelman, Luca and Svirsky, 2017).

How can ethnic discrimination be reduced? Most prior research focuses on reducing prejudice directly, for example, through diversity training and implicit-bias interventions. However, the causal effects of such interventions are often weak or short-lived. Implicit-bias interventions can shift preferences in the moment, but the effects often fade within days and rarely translate into behavior (Lai et al., 2016; Forscher et al., 2019; FitzGerald et al., 2019); the effects of diversity training on attitudes decay over time (Bezrukova et al., 2016); and reviews of this literature suggest that the causal effects of the most widely deployed programs remain largely unknown, dominated by light-touch designs of uncertain durability (Paluck and Green, 2009; Paluck et al., 2021). Intergroup-contact interventions, though promising, show mixed effects in well-powered field experiments (Littman, Scacco and Weiss, 2023), and a recent review suggests that intergroup contact has, if anything, small average effects on prejudice (Lowe, 2025), partly because individuals, especially the more prejudiced, tend to avoid outgroup contact rather than engage it (Weiss, Siegel and Scacco, 2026).

If prejudice itself is hard to change, an alternative is to constrain its *expression* through institutional design. Paluck et al. (2021), for example, in a recent meta-analytic review of prejudice-reduction experiments, call for pairing efforts to change attitudes with structural and institutional

approaches. One such approach is to facilitate intergroup contact in public institutions, where routine positive interaction with minorities can decrease prejudice (Weiss, 2021, 2026). Another prominent class of such remedies centers on institutional oversight: making decisions visible, reviewable, and accountable to a third party, as in police body-worn cameras, audit and review regimes, performance grading, and observer assignments. Previous research on this class of interventions has yielded mixed findings. In some settings, oversight has been shown to suppress own-group favoritism in officiating (Parsons et al., 2011; Pope, Price and Wolfers, 2018) and to reduce police use of force (Ariel, Farrar and Sutherland, 2015); in others, including subsequent and larger evaluations of police body-worn cameras, it has not (Yokum, Ravishankar and Coppock, 2019; Ariel et al., 2016), or has worked only for some groups (Huff, 2022).[1]

In this paper, we contribute to this literature by examining whether institutional oversight can suppress the expression of ethnic bias without requiring a change in underlying attitudes. We build on a long-standing literature that distinguishes between prejudice as a disposition and its behavioral expression (LaPiere, 1934; Merton, 1949; Allport, 1954). For any given level of underlying prejudice, situational factors such as social norms govern how much of it is expressed (Crandall, Eshleman and O'Brien, 2002); accountability and observability in particular can raise the cost of acting on prejudice (Lerner and Tetlock, 1999). Institutional oversight is such a situational factor: it targets the expression of prejudice rather than the prejudice itself.

We study this question in German amateur football (soccer), specifically in Bavaria. The setting allows us to observe the expression of ethnic bias at scale: sanctioning decisions are systematically recorded, and each is a discretionary decision that referees make in applying a common rulebook. In this respect, they resemble bureaucrats whose everyday discretionary judgments determine how rules are actually applied—police officers, teachers, and judges (Lipsky, 1980)—and, like those officials, they are appointed by a principal to be impartial (Dohmen and Sauermann, 2016). Amateur football also involves a large share of the population: with millions of participants and daily in-

[1] Oversight may also shift bias to less observable margins rather than eliminate it (Mocan and Osborne-Christenson, 2024).

teraction across ethnic and social groups, it is one of Europe's largest voluntary civic institutions.[2] Because its referees and players are drawn from the general population, findings from this setting are more likely to generalize to everyday discretionary decisions than evidence from professional leagues (Price and Wolfers, 2010; Parsons et al., 2011; Gallo, Grund and Reade, 2013; Magistro and Wack, 2023).

Institutional oversight in this setting varies sharply across league tiers, while the formal rules do not: the same rules and sanctions apply at every tier of the ladder. However, matches in the top tiers are attended by an external observer whose report feeds into referee promotion decisions, whereas matches in lower tiers (*Kreisklasse* and below) are rarely observed at all (SI Section C). We leverage this variation in institutional oversight across league tiers for identification.

For our empirical analysis, we assemble a new panel data set spanning 12 seasons (2013/14 through 2025/26): roughly 750,000 matches, 380,000 unique players, and 14,000 referees. For every player-match, we observe each disciplinary card and the player's name. To approximate player ethnicity, we map each player's most likely name origin onto a binary white/non-white indicator under a fixed classification scheme (Mateos, 2007), and validate that classification against photographs coded by hand.[3]

Our empirical analysis proceeds in two steps. First, we provide descriptive evidence of ethnic bias in officiating, in line with most, though not all, prior work on football refereeing (Magistro and Wack, 2023; Gallo, Grund and Reade, 2013; Pope and Pope, 2015)[4] and the broader sports-officiating literature (Price and Wolfers, 2010; Parsons et al., 2011); for a survey, see Dohmen and Sauermann (2016). Comparing players on the same team in the same league, facing the same opponent under the same referee, we find that non-white players receive more yellow cards, more

[2] Previous research has demonstrated ethnic discrimination at the point of entry into amateur clubs: in a correspondence experiment covering roughly 23,000 clubs in 22 European countries, applicants with foreign-sounding names received about ten percent fewer positive responses (Gomez-Gonzalez, Nesseler and Dietl, 2021). For German amateur sports clubs, see Nesseler et al. (2026); for women's amateur football in Scandinavia, see Storm et al. (2023).

[3] *White* and *non-white* are shorthand here for whether a player is likely to be seen as belonging to a visible ethnic or racial minority in Germany. It should be noted that the terms do not carry the same meaning that they have in the United States, where race is a social and administrative category; in Germany, it is neither. We use them simply to capture visible outgroup status.

[4] An exception is Reilly and Witt (2011), who find no ethnic gap in disciplinary sanctions in the English Premier League.

straight reds, and more total cards. The gap is most pronounced for Turkish-origin players. This baseline estimate is descriptive, since unobserved player and referee differences may account for part of it. Our goal is to estimate the effect of oversight on the gap rather than its absolute level in any given tier.

Second, we estimate the causal effect of increased oversight on this gap. This requires separating oversight in higher leagues from at least three potential confounds. First, more talented, less foul-prone non-white players might sort into higher tiers (player selection). Second, officials in higher tiers may be less biased (referee selection). Third, league rank may itself widen or narrow the ethnic card gap through channels other than oversight, e.g., the larger crowds that higher tiers draw (league-rank effects).[5]

Our main identification strategy, a within-player triple-difference design, holds the player fixed and thus accounts for player selection by design. We test whether the white/non-white card gap shifts more when a club's promotion or relegation changes the oversight regime than when a one-league move inside the never-observed bottom divisions changes league rank alone. In other words, we compare across league-tier changes to net out general league-rank effects.[6] We address remaining identification concerns through two supplementary strategies: to rule out that higher leagues are simply staffed by less biased officials, we estimate a within-referee specification that tests whether the same referee's card gap is smaller under closer oversight (SI Section F). In the second, we hold league rank fixed rather than netting it out: within a single league tier, promotion-cadre referees are routinely observed while their peers are not, and we ask whether the observed referees' card gap is smaller once the cadre's positive selection is accounted for (SI Section D.2). Across these different specifications and sources of identifying variation, we find consistent evidence that external oversight suppresses the expression of ethnic bias in refereeing decisions.

Whether institutional oversight suppresses the expression of ethnic bias is a question on which prior research has yielded mixed findings: studies have found both meaningful suppression (Par-

[5]Match-day attendance is higher in higher league tiers (SI Table A.4).

[6]If oversight also raises sanctioning levels overall (Hediger, Meier and Franck, 2026), proportional scaling alone would widen the absolute card gap under closer oversight; the narrowing we estimate is, on that account, conservative.

sons et al., 2011; Pope, Price and Wolfers, 2018) and precise nulls or high heterogeneity across subgroups (Yokum, Ravishankar and Coppock, 2019; Huff, 2022). Much of this evidence comes from discrete, one-off interventions: the rollout of body-worn cameras in particular police departments (Ariel, Farrar and Sutherland, 2015; Yokum, Ravishankar and Coppock, 2019) or a publicized bias finding that made NBA referees aware they were being studied (Pope, Price and Wolfers, 2018). A complementary literature instead exploits variation in institutional context (Parsons et al., 2011). Callback discrimination among large U.S. employers, for example, is lower at firms where hiring is centralized and regulatory scrutiny greater, which Kline, Rose and Walters (2022) suggest reflects rules replacing the discretionary judgments of individual decision-makers. Closest to our setting, Hediger, Meier and Franck (2026) show that Swiss amateur referees impose fewer sanctions when no inspector attends the match.

Our study advances research on ethnic bias and institutional oversight in two ways. First, we provide new causal evidence on this question. Our study differs from prior work in three respects: (i) the oversight we study is not a one-off intervention but similar to the routine infrastructure through which large organizations monitor their decision-makers—structured observation by trained graders that feeds formal evaluation, promotion, and demotion; (ii) we identify the effect of oversight from within-player and within-referee changes in the oversight regime itself, not from comparisons across differently monitored environments (Kline, Rose and Walters, 2022); and (iii) unlike Hediger, Meier and Franck (2026), we ask whether oversight changes *whom* officials sanction, not how much. More generally, our setting is one instance of a broader class of accountability mechanisms that constrain behavior by making decisions observable and reviewable rather than by reforming the agent (Fearon, 1999).[7]

Second, we provide field evidence at scale for a long-standing distinction in research on prejudice: that prejudice and its expression can move on different margins, with situational constraints regulating how much is expressed (LaPiere, 1934; Allport, 1954; Merton, 1949). We ask whether

[7]Related research documents how audits curb corruption in local government (Olken, 2007; Avis, Ferraz and Finan, 2018), elections discipline incumbents (Besley and Burgess, 2002), and legislative oversight checks delegated discretion in the bureaucracy (McCubbins and Schwartz, 1984).

closer observation reduces the bias that officials express and find evidence that it does – in line with the idea that accountability and observability can work as situational determinants of expression (Crandall, Eshleman and O'Brien, 2002; Lerner and Tetlock, 1999).[8] We also interpret our findings as suggestive evidence against a purely implicit account of refereeing bias (Gallo, Grund and Reade, 2013): bias that decreases when referees know they are being graded is plausibly, at least in part, deliberate.

2 Data and empirical strategy

2.1 Data

Our empirical analysis draws on match records from the Bavarian Football Association (*Bayerischer Fußball-Verband*, BFV) covering all senior men's matches played from the 2013/14 through the 2025/26 seasons.[9] The sample spans the entire Bavarian senior men's league system, from the *Regionalliga Bayern* at the top down through the *Bayernliga*, *Landesliga*, *Bezirksliga*, *Kreisliga*, *Kreisklasse*, and the regional A-, B-, and C-Klasse divisions at the bottom.[10] Our sample comprises roughly 750,000 matches, 22.4 million player-matches, 380,000 unique players, and 14,000 referees. For each match, we observe the home and away teams, the referee, the player lineup (including substitutions), and all disciplinary cards. We do not observe player positions. SI Section A reports additional detail on data sources, sample construction, and outcome variables, and SI Table A.2 and SI Figure A.1 report the composition of the panel by league and season.

Yellow cards are cautions, issued for offenses such as unsporting conduct, dissent, or persis-

[8] A similar logic motivates a growing literature in political behavior, where social norms are the informal counterpart to the institutional oversight we study: individuals hold a fixed radical-right preference and express it only to the degree the social environment permits, so that expression rises as those norms weaken (Valentim, 2024; Bursztyn, Egorov and Fiorin, 2020; Álvarez-Benjumea and Winter, 2020).

[9] We exclude the women's leagues, which account for 6.3 percent of player-match observations in the data, since the smaller subsample is not powered for the within-actor designs. We also drop the 2020/21 season because of the pandemic-related curtailment of amateur football.

[10] We focus on Bavaria because the BFV's match records link each disciplinary card to a player. The German Football Association's national amateur football portal (`https://www.fussball.de`), which covers the rest of Germany, does not allow us to link cards and player names.

tent fouling. Red cards are reserved for more serious offenses, among them violent conduct and denying an obvious goal-scoring opportunity. Two yellows in the same match also amount to a red, and the player must leave the pitch. We use total cards (yellows, second yellows, and reds) at the player-match level as our main outcome variable. SI Table A.1 reports full summary statistics for the outcomes, treatment, and control variables.

Our measure of player ethnicity is a binary white/non-white indicator derived from a name-based classifier (Mateos, 2007). The classifier predicts each player's most likely country of origin from their first and last names. We code players whose predicted origin is a European country, the United States, Canada, Australia, or New Zealand as white and all other players as non-white. We note that our measure proxies perceived, not self-reported, group membership. Referees can perceive ethnicity through two channels: (i) the player's physical appearance, and (ii) the surname itself, read on the lineup before kickoff and, where the club prints it, on the back of the jersey throughout the match.[11]

To verify that this coding rule captures perceived ethnicity, we hand-code 1,736 player photographs; the name-based classification agrees with the photo-based coding in 90 percent of cases. SI Section B describes the classification procedure and the validation exercise in detail. SI Table B.2 reports worked examples mapping names to classifications, and SI Figure B.1 plots the distribution of the non-white share across team-matches in the analysis sample.[12]

2.2 Institutional variation across league tiers

Referee oversight rises step by step up the league ladder. Nearly every match in the *Regionalliga Bayern* is attended by a graded observer whose report feeds formal evaluation and promotion. The same corps observes the *Bayernliga* and, less often, the *Landesliga*. *Bezirksliga* coverage is

[11] Under BFV rules, only a back number is required on the jersey, and it must match the electronic match report (*Spielbericht*); the surname is not required and, where used, is printed above or below the number. The referee's pre-kickoff source is therefore the match report, which is mandatory in all BFV leagues and lists each player's number, surname, and first name. BFV-*Spielordnung* §26, §28.

[12] We observe some classification error for our white/non-white indicator. However, if anything, this attenuates our results: dropping countries where the photo codes most strongly disagree with the rule increases the estimated ethnic card gap (SI Figure B.2).

partial and not guaranteed. By the *Kreisliga*, routine observation is confined to a small promotion cadre of referees. Observation is not covert. Where a match is observed, the assigned observer introduces themselves to the referee before kickoff and reviews the match with them in person afterward. The referee thus knows during the match that they are being graded. Referees also have financial incentives to receive good evaluations: they affect chances of promotion, which in turn affects per-match compensation. For Bavarian referees, the per-match fee rises more than sixfold from the bottom of the ladder to the *Regionalliga Bayern*.[13] Only the referee is graded. Players are not evaluated and have no stake in whether a match is watched. Below the *Kreisliga* lie the never-observed bottom divisions: the *Kreisklasse* and the regional A-, B-, and C-Klasse, where no match has a routine observer regardless of its rank. Oversight rises across the ladder except in these bottom divisions, where league rank still varies but oversight stays at zero. We leverage this variation across league ranks without changing oversight to net out league-rank effects (see Section 2.3). Table 1 summarizes the oversight regime for each league. SI Section C provides additional details on observer assignment, grading, and promotion.

2.3 Empirical strategy

We first estimate a pooled OLS regression to describe the baseline white/non-white disciplinary card gap. We then identify the causal effect of institutional oversight on this gap. All our card-gap specifications adjust for the same vector of player-match covariates X_{icjr}: the opponent team's non-white share, a home indicator, and a sub-in indicator equal to one if the player entered as a substitute.

Descriptive baseline. We estimate a standard OLS specification of the following form:

$$y_{icjr} = \alpha_{c\ell s} + \alpha_j + \alpha_r + \beta\,\mathrm{NW}_i + \gamma' X_{icjr} + \varepsilon_{icjr} \tag{1}$$

[13] BFV *Schiedsrichter-Spesenordnung* 2025/26 (in force from 1 July 2025): the per-match referee fee runs from 41 euros in the B- and C-Klasse and 46 euros in the *Kreisklasse* through 52 euros (*Kreisliga*), 62 euros (*Bezirksliga*), 72 euros (*Landesliga*), and 98 euros (*Bayernliga*) to 260 euros in the *Regionalliga Bayern*, plus a per-kilometer travel reimbursement.

Table 1: Referee oversight across the Bavarian league system

League	Oversight regime	Pl.-matches (%)
Regionalliga Bayern	Graded observer at (nearly) every match; feeds formal evaluation	0.6
Bayernliga	Same observer corps; no guaranteed video coverage	1.1
Landesliga	Same corps; less frequent, no per-match observation target	2.6
Bezirksliga	Partial, less formal coverage; not guaranteed	5.9
Kreisliga	Sparse; only the promotion cadre observed routinely	13.2
Kreisklasse	**Never observed:** no routine observer at any rank	24.4
A-Klasse		29.8
B-Klasse		18.8
C-Klasse		3.7
Total		**100.0**

Notes: Leagues ordered from the top of the ladder down to the never-observed bottom divisions; the final column gives each league's share of the 22.4M player-match analysis sample, in percent (rounded; shares may not sum to exactly 100). SI Section C gives the rung-by-rung regime.

where y_{icjr} is a card outcome (yellow, straight red, or total cards) for player i on club c in season s, facing opponent j under referee r. NW_i is a binary indicator for non-white player ethnicity, while $\alpha_{c\ell s}$, α_j, and α_r denote club×league×season, opponent, and referee fixed effects. Our coefficient of interest is β, the white/non-white card gap from a comparison of players in the same club-league-season cell, after adjusting for opponent and referee differences. We estimate Equation 1 on the full pooled sample in a single OLS fit, with standard errors clustered at the player level. This gap is descriptive, not causal: unobservable player characteristics (for example, playing position) may still confound it. Consistent with our argument, however, the gap is concentrated in the lower tiers (SI Figure E.2).

Within-player triple-difference. To estimate the causal effect of oversight on the white/non-white card gap, we use a triple-difference design that leverages within-player variation across the two

consecutive seasons bracketing a one-league promotion or relegation of a player's club into a different oversight regime. For each such move we measure how a player's card rate changes from the season before to the season after, and ask whether this within-player change differs for non-white and white players. Holding the player fixed rules out that the shift reflects player selection rather than refereeing—for example, clubs entering a higher-oversight league fielding systematically different players. However, league moves often change a club's rank and its oversight regime at once, so the change in the card gap could reflect league rank—for instance, the larger and more partisan crowds that higher tiers draw—rather than oversight. The third difference nets out this channel leveraging league moves across the never-observed bottom divisions: a one-league promotion or relegation there changes a club's rank but *not* the oversight regime. We test whether the white/non-white card gap shifts more when a move crosses an oversight boundary than when it changes rank alone (Table 2).

Table 2: Within-player triple-difference design

	$\Delta = \bar{y}_{\mathrm{NW}} - \bar{y}_{\mathrm{W}}$		
	post	**pre**	*post* − *pre*
Oversight-crossing	$b+s+\rho+\beta$	$b+s$	$\rho+\beta$
Pure-rank	$b+\rho$	b	ρ
Oversight − *rank*	$s+\beta$	s	$\boxed{\beta}$

Notes: Each cell is the white/non-white card gap $\Delta = \bar{y}_{\mathrm{NW}} - \bar{y}_{\mathrm{W}}$ for that group of player-matches, decomposed into additive components. Here b is the baseline gap shared by every cell, and s the fixed difference in that gap between oversight-crossing and pure-rank movers. The column term ρ is the league-rank effect: the shift in the gap that one league-step of rank produces through channels other than oversight, and the confound the design nets out. Only the shaded treated cell carries the oversight effect β. The cells are written for a step up the ladder; for a step down, ρ and β enter with the opposite sign.

A player can also cross tiers by transferring to another club. Such moves are the player's own choice, however, and would reintroduce the player selection that the design is meant to rule out. We therefore leverage only club promotions and relegations for identification. This leaves four kinds of moves: (i) an oversight-crossing step up (e.g., from the *Kreisliga* to the *Bezirksliga*); (ii) an oversight-crossing step down (e.g., from the *Bezirksliga* to the *Kreisliga*); (iii) a pure-rank step

up within the bottom divisions (e.g., from the A-Klasse to the *Kreisklasse*); and (iv) a pure-rank step down within the bottom divisions (e.g., from the *Kreisklasse* to the A-Klasse). We estimate a signed pooled version (SI Section D.1) of the fully saturated triple-difference

$$
\begin{aligned}
y_{icjr} = \alpha_i + \alpha_s + \delta\,\text{post} + \sum_m \lambda_m \mathbf{1}[M=m] \\
+ \sum_m \pi_m\,(\text{post} \times \mathbf{1}[M=m]) + \sum_m \varphi_m\,(\text{NW}_i \times \mathbf{1}[M=m]) \\
+ \underbrace{\sum_{m\in\text{Ovr}} \beta_m\,(\text{post} \times \text{NW}_i \times \mathbf{1}[M=m])}_{\text{oversight-crossing}} \\
+ \underbrace{\sum_{m\in\text{Rnk}} \beta_m\,(\text{post} \times \text{NW}_i \times \mathbf{1}[M=m])}_{\text{pure rank}} + \gamma' X_{icjr} + \varepsilon_{icjr},
\end{aligned}
\tag{2}
$$

with the four move classes $m \in \{\text{Rnk}_\uparrow, \text{Rnk}_\downarrow, \text{Ovr}_\uparrow, \text{Ovr}_\downarrow\}$, grouped into the oversight-crossing moves (Ovr) and the pure-rank moves (Rnk) internal to the bottom divisions. Here α_i and α_s are player and season fixed effects, post marks the post-move season of the bracketing pair, and $\mathbf{1}[M=m]$ is the move class.[14] Standard errors are clustered at the team level, the team being the side whose promotion or relegation defines a move.

Our estimands are the two third differences in Table 2: OVR-UP $= \beta_{\text{Ovr}_\uparrow} - \beta_{\text{Rnk}_\uparrow}$ for promotions and OVR-DN $= \beta_{\text{Ovr}_\downarrow} - \beta_{\text{Rnk}_\downarrow}$ for relegations. Each compares the change in the card gap around an oversight-crossing move with the change around a pure-rank move in the same direction, so that the rank component of the move is netted out and only its oversight component remains. The two are mirror images of one another: promotions increase oversight and relegations reduce it.[15] We therefore pool them into a single signed estimate. To do so, we code every league move as one signed rank step ($+1$ for a promotion, -1 for a relegation) and add a second regressor,

[14] Because each observation falls in exactly one move class, the four indicators sum to one, so Equation 2 carries one redundant parameter in each of the λ_m, π_m, and φ_m blocks; in estimation, $\text{Rnk}_\downarrow$ is the omitted reference class. The NW_i main effect is absorbed by α_i; so are λ_m and φ_m for players observed in a single move window, which leaves them identified only from players who appear in more than one window. The four β_m are identified because there is no stand-alone post $\times$ NW_i term.

[15] Not imposing symmetric upward and downward effects yields consistent point estimates with larger standard errors (SI Table D.1).

the oversight increment, coded the same way but nonzero only for moves that cross an oversight boundary. The pure-rank moves in the bottom divisions identify the rank effect, and the coefficient on the oversight increment is the additional shift in the ethnic card gap that an oversight-crossing move produces, net of rank (SI Section D.1). Our key identification assumption is that the pure rank effect is similar in magnitude in the never-observed bottom divisions and in the higher divisions. We complement the within-player estimate with two complementary designs that rely on different identifying variation: one holds league rank fixed (SI Section D.2), the other holds the officiating referee fixed (SI Section F).

3 Results

3.1 Baseline card gap

We present descriptive estimates of the baseline white/non-white card gap in Table 3. We find a positive, statistically significant gap in card receipts between non-white and otherwise comparable white players, in line with prior research on ethnic and in-group bias in football refereeing (Magistro and Wack, 2023; Gallo, Grund and Reade, 2013; Pope and Pope, 2015).[16] The gap is statistically significant at conventional levels for all three card outcomes: yellow cards, straight reds, and pooled total cards. This ethnic card gap is concentrated in the lower, minimally overseen tiers; in the higher, more closely overseen tiers it is statistically indistinguishable from zero (SI Figure E.2).[17]

Testing for heterogeneity within the group of ethnic minority players, we find that ethnic bias is concentrated among players of Turkish origin, whose gap is roughly four times the pooled estimate. For players with Latin American names we measure no gap, and for players with Asian names we

[16] Estimates in this literature are sensitive to analytic choices: 29 teams analyzing a single soccer data set on whether referees red-card dark-skin-toned players more often returned odds ratios from 0.89 to 2.93, with 20 of the 29 finding a significant positive effect (Silberzahn et al., 2018). We therefore report the gap across a wide set of alternative cuts rather than a single preferred estimate (SI Section E).

[17] This gradient mirrors survey evidence from German amateur football: Kalter (2003) finds that migrant players report significantly less perceived discrimination in higher leagues.

Table 3: Baseline ethnic card gap

	Player-match card outcomes		
	Yellow	Red	Total
	(1)	(2)	(3)
Non-white	0.00549***	0.00041***	0.00690***
	(0.00071)	(0.00004)	(0.00079)
Mean of DV	0.124	0.002	0.133
Player-matches	22,411,568	22,411,568	22,411,568
% of DV mean	+4.4%	+18.9%	+5.2%
Controls	✓	✓	✓
Club×league×season FE	✓	✓	✓
Opponent FE	✓	✓	✓
Referee FE	✓	✓	✓

Notes: OLS estimated on the full pooled sample in a single fit, absorbing club×league×season, opponent, and referee fixed effects (2013/14–2025/26; 2020/21 dropped). Yellow and Red are binary indicators for receiving any yellow / straight-red card; Total counts yellow + second-yellow + red. Controls: opponent non-white share, home indicator, sub-in indicator. SI Table K.1 reports the full estimates, including the coefficients on the controls. The *% of DV mean* row reports each gap as a share of the dependent-variable mean, computed from unrounded values. Standard errors clustered at the player level. Signif. Codes: ***: 0.01, **: 0.05, *: 0.1.

estimate a small negative gap (-0.0045, $p < 0.01$). We find little evidence, in contrast, that the gap reflects the hostility of the crowd or the local political climate. Our two proxies for an audience channel are the local vote share of the far-right Alternative für Deutschland (AfD) and whether the player's team plays away. The gap is statistically indistinguishable across municipalities with high and low AfD support, and only slightly larger for away than for home players. Nor does it depend on the referee's own ethnicity, with white and non-white referees exhibiting similar gaps. The gap is likewise stable across the stages of the season, with the early, middle, and late thirds of each season's schedule yielding statistically indistinguishable gaps (SI Table E.1). SI Section E reports the full set of subsample estimates.

We caution that this baseline gap is descriptive rather than causal. Non-white and white players can differ in playing position, style of play, on-pitch role, or fitness, and any of these factors may contribute to the observed gap. One version of this concern is that non-white players commit

more dangerous fouls. We find no evidence that they do. Our proxy for whether a foul injured an opponent is whether that opponent is substituted off in the minutes that follow. Conditional on a straight-red foul, such a substitution is no more likely when the dismissed player is non-white than when he is white.[18] If anything, the estimate runs the other way, with non-white players appearing to play *less* aggressively (SI Section G, Table G.1). The same caveat applies to how the gap varies across tiers. The confounds we flag in the introduction, in particular player and referee selection, themselves move with league tier. The gap's decline toward the top may therefore reflect the selection of less foul-prone minority players and less biased referees into higher leagues rather than oversight itself. Next, we turn to how institutional oversight shifts this gap.

3.2 Effect of institutional oversight

Table 4 reports our main within-player estimate of the effect of oversight on the expression of ethnic bias. We find that a one-league move that adds oversight narrows the white/non-white total-card gap, net of one league-step of rank, while the netted pure-rank slope, identified from equivalent moves inside the never-observed bottom divisions, is small and statistically insignificant. The effect is statistically significant and in the predicted direction: closer oversight suppresses the expression of ethnic bias in refereeing decisions.

We estimate that a one-league step into closer oversight narrows the white/non-white total-card gap by roughly 1.6 percent of a standard deviation of total cards in the estimation sample, or about 4.1 percent of its mean (0.147 cards per player-match). Clubs, and the players who move with them, rarely shift more than a rung or two across neighboring seasons. We therefore caution against extrapolating this per-step estimate linearly across the full league ladder.

[18] Because carding is itself the outcome under study, an unconditional rate of injury events would conflate how dangerously non-white players foul with how often referees card them. We therefore condition on a dismissal, which is a post-treatment referee decision. The comparison is distorted if referees apply a lower bar to non-white players (our account of the carding gap). However, if carding decisions are unbiased and minorities play more rough (the alternative channel), injuries should be more frequent among non-white dismissals.

Table 4: Within-player triple-difference

	Total cards			
	Player FE		Window × player FE	
	Main estimate (1)	Pre-trend placebo (2)	Main estimate (3)	Pre-trend placebo (4)
NW × Oversight	−0.0060**	−0.0004	−0.0063**	−0.0006
	(0.0028)	(0.0031)	(0.0028)	(0.0030)
NW × Pure rank	0.0007	0.0026	0.0006	0.0024
	(0.0017)	(0.0020)	(0.0018)	(0.0020)
Mean of DV	0.147	0.146	0.147	0.146
Player-matches	5,434,230	4,457,065	5,434,230	4,457,065
Controls	✓	✓	✓	✓
Season FE	✓	✓	✓	✓
Player FE	✓	✓		
Window × player FE			✓	✓

Notes: Outcome: total cards (yellow + second-yellow + red). NW × Oversight is the per-step oversight effect on the white/non-white card gap, net of one league-step of rank (the boxed β of Table 2); NW × Pure rank is the netted league-rank slope. The estimator codes each one-league move ± 1 by direction, pooling the mirror-image moves and imposing OVR-UP $= -$OVR-DN (SI Section D.1). Columns (1) and (3) use the move window $(t, t+1)$; Columns (2) and (4) are the pre-trend placebo on $(t-1, t)$. Columns (1) and (2) include player fixed effects; Columns (3) and (4) replace them with window × player fixed effects, which confine each comparison to a single move window for players whose club moves in consecutive seasons. Controls: opponent non-white share, home indicator, sub-in indicator. SI Table K.2 reports the full estimates, including all lower-order interaction terms and the coefficients on the controls. Standard errors clustered at the team level. Signif. Codes: ***: 0.01, **: 0.05, *: 0.1.

3.3 Alternative estimation strategies and robustness

We probe the within-player result with two alternative identification strategies—a within-referee design that holds the officiating referee fixed (SI Section F) and a within-league design that holds league rank fixed (SI Section D.2)—and a battery of additional specifications.

A first concern is referee selection. Promotion raises oversight, but it also moves a club's players under a more selected pool of referees, so the narrowing we attribute to oversight could instead reflect the selection of less biased officials into higher leagues. This is the alternative a market-competition account of ethnic bias would emphasize, in which the institutionalized competition of league football tends to select bias out of the system, increasingly so in higher tiers (Kalter,

2003).[19] To address this, we hold the officiating referee fixed and ask whether the *same* referee's white/non-white card gap is smaller when they officiate under closer oversight (SI Section F). We find that the same referee's white/non-white card gap is smaller in higher tiers, but only when those higher tiers carry more oversight. In the never-observed bottom divisions, where rank varies but no observer attends, the same placebo returns a precise null (SI Table F.1). This speaks against referee selection driving our result.

A second concern is that our main design nets out general league-rank effects only under the assumption that rank effects are of similar magnitude in the never-observed bottom divisions and in the observed tiers above. We therefore complement it with a within-league design that holds league rank fixed instead of netting it out. Within the *Kreisliga*, the BFV routinely observes only a small promotion cadre (*Förderkader*) of referees and leaves the rest largely unobserved. This lets us contrast the white/non-white card gap of observed cadre referees with that of otherwise comparable referees at the same rank, net of the same promotion-track contrast among the never-observed referees below (SI Section D.2, Table D.7). The results are consistent with our main design: being observed suppresses the card gap. We also estimate a placebo treatment test: we estimate a specification where we shift the observation boundary one rung down, so that rank still changes but no match is watched. The estimates are smaller and statistically indistinguishable from zero (Columns 3–4 of Table D.7).

We further verify that the within-player result is not an artifact of pre-existing trends, the match-day audience, referee experience, or differential player behavior. First, the card gap might already have been moving for players whose club is about to move, so what we read as an oversight effect could be a pre-existing trend; re-estimating the design on the two seasons before the move yields a point estimate that is close to zero and statistically insignificant (Column 2 of Table 4). Second, higher tiers draw larger crowds, so the effect could reflect a watching crowd rather than a graded evaluator; adding match-day audience and its interaction with player ethnicity to the con-

[19] In that account the biased actors are clubs, trainers, and teammates, whose biased choices carry a sporting cost and are therefore competed away more strongly in higher, more competitive tiers. The analogous concern in our setting is referee selection: referees in higher tiers may be less biased. We address this by estimating a specification including referee fixed effects, comparing within the same referee (SI Section F).

trols leaves the estimate essentially unchanged (SI Section H), pointing to the cost of a graded, consequential assessment rather than mere observation. Third, referees in higher tiers have typically officiated for longer, so the effect could reflect what referees learn over their careers rather than being observed; we therefore estimate a specification that includes referee experience and its interaction with player ethnicity in the controls. More experienced referees do hand out cards to white and non-white players somewhat more evenly, but the estimated effect of oversight on the card gap does not move (SI Section I). Fourth, the result might reflect a change in how non-white players play under oversight rather than in how referees judge them; we therefore re-estimate the design on whether a player is substituted off, an outcome that coaches rather than referees decide. This placebo returns a null (SI Section J), which we interpret as suggestive evidence that our results are driven by referee rather than player behavior. Fifth, six in ten oversight-crossing moves run between the *Kreisliga* and the *Kreisklasse*, the least closely observed boundary; dropping that rung leaves the point estimate unchanged (-0.0070, standard error 0.0037; SI Table D.4). Finally, the pooled estimator imposes mirror-image symmetry between promotions and relegations for power; estimating the two directions separately yields consistent point estimates and does not reject the pooling restriction (SI Table D.1). Together, these alternative designs and specifications increase our confidence that the narrowing of the ethnic card gap reflects institutional oversight rather than selection, league rank, audience, referee experience, or player behavior.

4 Discussion

Our findings show that institutional oversight can suppress the expression of ethnic bias. Across multiple identification strategies, we find that oversight substantially narrows the ethnic disciplinary-card gap, consistent with the idea that accountability changes how underlying biases translate into behavior. These results speak against a purely implicit account of ethnic bias in our setting (Gallo, Grund and Reade, 2013; Price and Wolfers, 2010). If biased carding were entirely implicit and subconscious, we would not expect a narrowing of the ethnic card gap when referees

know that they are being graded. We interpret the finding that bias is moderated by incentives as suggestive evidence that it is at least in part driven by deliberate decision-making.

Our findings point to a strategy for reducing discrimination that complements efforts targeting attitudes directly, such as diversity training, implicit-bias interventions, and intergroup contact (Paluck and Green, 2009; Lai et al., 2016; Paluck et al., 2021). Rather than eliminating prejudice itself, institutional design operates on a different margin: it can limit the extent to which existing prejudice translates into consequential decisions that disadvantage minorities. Institutional design thus deserves a central place in debates about reducing discrimination (Weiss, 2026).

At the same time, monitoring involves trade-offs: it typically requires organizational resources and trained personnel. In our setting, each observation requires an experienced referee, who must attend the match and produce an evaluation report. The German Football Association thus concentrates systematic oversight in the higher leagues, where the stakes of officiating decisions are greatest. Whether monitoring is worth the costs will depend on the specific setting, how common and how consequential discrimination is, weighed against the costs of implementing oversight at scale. Advances in artificial intelligence may shift this balance: automated tools could lower the cost of evaluating individual decisions at scale, making systematic oversight feasible even in settings where human observers are too costly.

Data Availability Statement

All data and code will be made publicly available upon conditional acceptance.

Ethics Declaration

This study does not contain any original human subjects research.

Statement of AI Use

Generative artificial intelligence (Claude Code) was used to assist with coding, copyediting, and the write-up of the Supporting Information. All instances of AI use were supervised by the authors, and AI outputs were manually verified for their quality and completeness.

References

Aigner, Dennis J. 1973. "Regression with a Binary Independent Variable Subject to Errors of Observation." *Journal of Econometrics* 1(1):49–59.

Allport, Gordon W. 1954. *The Nature of Prejudice*. Reading, MA: Addison-Wesley Publishing Company.

Álvarez-Benjumea, Amalia and Fabian Winter. 2020. "The Breakdown of Antiracist Norms: A Natural Experiment on Hate Speech after Terrorist Attacks." *Proceedings of the National Academy of Sciences* 117(37):22800–22804.

Ariel, Barak, Alex Sutherland, Darren Henstock, Josh Young, Paul Drover, Jayne Sykes, Simon Megicks and Ryan Henderson. 2016. "Wearing Body Cameras Increases Assaults Against Officers and Does Not Reduce Police Use of Force: Results from a Global Multi-Site Experiment." *European Journal of Criminology* 13(6):744–755.

Ariel, Barak, William A. Farrar and Alex Sutherland. 2015. "The Effect of Police Body-Worn Cameras on Use of Force and Citizens' Complaints Against the Police: A Randomized Controlled Trial." *Journal of Quantitative Criminology* 31(3):509–535.

Arnold, David, Will Dobbie and Crystal S. Yang. 2018. "Racial Bias in Bail Decisions." *The Quarterly Journal of Economics* 133(4):1885–1932.

Avis, Eric, Claudio Ferraz and Frederico Finan. 2018. "Do Government Audits Reduce Corruption? Estimating the Impacts of Exposing Corrupt Politicians." *Journal of Political Economy* 126(5):1912–1964.

Bertrand, Marianne and Sendhil Mullainathan. 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.

Besley, Timothy and Robin Burgess. 2002. "The Political Economy of Government Responsiveness: Theory and Evidence from India." *The Quarterly Journal of Economics* 117(4):1415–1451.

Bezrukova, Katerina, Chester S. Spell, Jamie L. Perry and Karen A. Jehn. 2016. "A Meta-Analytical Integration of Over 40 Years of Research on Diversity Training Evaluation." *Psychological Bulletin* 142(11):1227–1274.

Bursztyn, Leonardo, Georgy Egorov and Stefano Fiorin. 2020. "From Extreme to Mainstream: The Erosion of Social Norms." *American Economic Review* 110(11):3522–3548.

Butler, Daniel M. and David E. Broockman. 2011. "Do Politicians Racially Discriminate Against Constituents? A Field Experiment on State Legislators." *American Journal of Political Science* 55(3):463–477.

Christensen, Peter and Christopher Timmins. 2022. "Sorting or Steering: The Effects of Housing Discrimination on Neighborhood Choice." *Journal of Political Economy* 130(8):2110–2163.

Crandall, Christian S., Amy Eshleman and Laurie O'Brien. 2002. "Social Norms and the Expression and Suppression of Prejudice: The Struggle for Internalization." *Journal of Personality and Social Psychology* 82(3):359–378.

Dohmen, Thomas and Jan Sauermann. 2016. "Referee Bias." *Journal of Economic Surveys* 30(4):679–695.

Edelman, Benjamin, Michael Luca and Dan Svirsky. 2017. "Racial Discrimination in the Sharing Economy: Evidence from a Field Experiment." *American Economic Journal: Applied Economics* 9(2):1–22.

Fearon, James D. 1999. Electoral Accountability and the Control of Politicians: Selecting Good Types versus Sanctioning Poor Performance. In *Democracy, Accountability, and Representation*,

ed. Adam Przeworski, Susan C. Stokes and Bernard Manin. Cambridge: Cambridge University Press pp. 55–97.

FitzGerald, Chloë, Angela Martin, Delphine Berner and Samia Hurst. 2019. "Interventions Designed to Reduce Implicit Prejudices and Implicit Stereotypes in Real World Contexts: A Systematic Review." *BMC Psychology* 7(1):29.

Forscher, Patrick S., Calvin K. Lai, Jordan R. Axt, Charles R. Ebersole, Michelle Herman, Patricia G. Devine and Brian A. Nosek. 2019. "A Meta-Analysis of Procedures to Change Implicit Measures." *Journal of Personality and Social Psychology* 117(3):522–559.

Gaddis, S. Michael. 2017. "How Black Are Lakisha and Jamal? Racial Perceptions from Names Used in Correspondence Audit Studies." *Sociological Science* 4:469–489.

Gallo, Edoardo, Thomas Grund and J. James Reade. 2013. "Punishing the Foreigner: Implicit Discrimination in the Premier League Based on Oppositional Identity." *Oxford Bulletin of Economics and Statistics* 75(1):136–156.

Gomez-Gonzalez, Carlos, Cornel Nesseler and Helmut M. Dietl. 2021. "Mapping Discrimination in Europe through a Field Experiment in Amateur Sport." *Humanities and Social Sciences Communications* 8(1):95.

Goncalves, Felipe and Steven Mello. 2021. "A Few Bad Apples? Racial Bias in Policing." *American Economic Review* 111(5):1406–1441.

Greenwood, Brad N., Rachel R. Hardeman, Laura Huang and Aaron Sojourner. 2020. "Physician–Patient Racial Concordance and Disparities in Birthing Mortality for Newborns." *Proceedings of the National Academy of Sciences* 117(35):21194–21200.

Hediger, Andrin, Pascal Flurin Meier and Egon Franck. 2026. "The Effect of Inspector Absence on Referees' Sanctioning Behavior: Evidence from Amateur Soccer." *Journal of Sports Economics*

.

Hoekstra, Mark and CarlyWill Sloan. 2022. "Does Race Matter for Police Use of Force? Evidence from 911 Calls." *American Economic Review* 112(3):827–860.

Hoffman, Kelly M., Sophie Trawalter, Jordan R. Axt and M. Norman Oliver. 2016. "Racial Bias in Pain Assessment and Treatment Recommendations, and False Beliefs About Biological Differences Between Blacks and Whites." *Proceedings of the National Academy of Sciences* 113(16):4296–4301.

Huff, Jessica. 2022. "Do Body-Worn Cameras Reduce Disparities in Police Behavior in Minority Communities? Evidence of Nuanced Influences Across Black and Hispanic Neighborhoods." *Criminology & Public Policy* 21(3):671–711.

Kalter, Frank. 2003. *Chancen, Fouls und Abseitsfallen: Migranten im deutschen Ligenfußball*. Wiesbaden: Westdeutscher Verlag.

Kline, Patrick, Evan K. Rose and Christopher R. Walters. 2022. "Systemic Discrimination Among Large U.S. Employers." *The Quarterly Journal of Economics* 137(4):1963–2036.

Lai, Calvin K., Allison L. Skinner, Erin Cooley, Sohad Murrar, Markus Brauer, Thierry Devos, Jimmy Calanchini et al. 2016. "Reducing Implicit Racial Preferences: II. Intervention Effectiveness Across Time." *Journal of Experimental Psychology: General* 145(8):1001–1016.

LaPiere, Richard T. 1934. "Attitudes vs. Actions." *Social Forces* 13(2):230–237.

Lerner, Jennifer S. and Philip E. Tetlock. 1999. "Accounting for the Effects of Accountability." *Psychological Bulletin* 125(2):255–275.

Lipsky, Michael. 1980. *Street-Level Bureaucracy: Dilemmas of the Individual in Public Services*. New York: Russell Sage Foundation.

Littman, Rebecca, Alexandra Scacco and Chagai M. Weiss. 2023. Reducing Prejudice through Intergroup Contact Interventions. In *Psychological Intergroup Interventions*, ed. Eran Halperin, Boaz Hameiri and Rebecca Littman. London: Routledge pp. 3–16.

Lowe, Matt. 2025. "Has Intergroup Contact Delivered?" *Annual Review of Economics* 17(1):321–344.

Magistro, Beatrice and Morgan Wack. 2023. "Racial Bias in Fans and Officials: Evidence from the Italian Serie A." *Sociology* 57(6):1302–1321.

Mateos, Pablo. 2007. "A Review of Name-Based Ethnicity Classification Methods and Their Potential in Population Studies." *Population, Space and Place* 13(4):243–263.

McCubbins, Mathew D. and Thomas Schwartz. 1984. "Congressional Oversight Overlooked: Police Patrols versus Fire Alarms." *American Journal of Political Science* 28(1):165–179.

Merton, Robert K. 1949. Discrimination and the American Creed. In *Discrimination and National Welfare*, ed. R. M. MacIver. New York: Harper & Brothers pp. 99–126.

Mocan, Naci and Eric Osborne-Christenson. 2024. "In-Group Favoritism and Peer Effects in Wrongful Acquittals: National Basketball Association Referees as Judges." *The Journal of Law and Economics* 67(4):731–766.

Mujcic, Redzo and Paul Frijters. 2021. "The Colour of a Free Ride." *The Economic Journal* 131(634):970–999.

Nesseler, Cornel, Carlos Gomez-Gonzalez, Petr Parshakov, Thadeu Gasparetto and Helmut Dietl. 2026. "Knocking on Sports Clubs' Doors: A Field Experiment on Ethnic Discrimination in Germany." *PNAS Nexus* 5(7):pgag216.

Olken, Benjamin A. 2007. "Monitoring Corruption: Evidence from a Field Experiment in Indonesia." *Journal of Political Economy* 115(2):200–249.

Paluck, Elizabeth Levy and Donald P. Green. 2009. "Prejudice Reduction: What Works? A Review and Assessment of Research and Practice." *Annual Review of Psychology* 60(1):339–367.

Paluck, Elizabeth Levy, Roni Porat, Chelsey S. Clark and Donald P. Green. 2021. "Prejudice Reduction: Progress and Challenges." *Annual Review of Psychology* 72(1):533–560.

Parsons, Christopher A., Johan Sulaeman, Michael C. Yates and Daniel S. Hamermesh. 2011. "Strike Three: Discrimination, Incentives, and Evaluation." *American Economic Review* 101(4):1410–1435.

Pope, Bryson R. and Nolan G. Pope. 2015. "Own-Nationality Bias: Evidence from UEFA Champions League Football Referees." *Economic Inquiry* 53(2):1292–1304.

Pope, Devin G., Joseph Price and Justin Wolfers. 2018. "Awareness Reduces Racial Bias." *Management Science* 64(11):4988–4995.

Price, Joseph and Justin Wolfers. 2010. "Racial Discrimination Among NBA Referees." *The Quarterly Journal of Economics* 125(4):1859–1887.

Quillian, Lincoln, Anthony Heath, Devah Pager, Arnfinn H. Midtbøen, Fenella Fleischmann and Ole Hexel. 2019. "Do Some Countries Discriminate More than Others? Evidence from 97 Field Experiments of Racial Discrimination in Hiring." *Sociological Science* 6:467–496.

Quillian, Lincoln, Devah Pager, Ole Hexel and Arnfinn H. Midtbøen. 2017. "Meta-Analysis of Field Experiments Shows No Change in Racial Discrimination in Hiring Over Time." *Proceedings of the National Academy of Sciences* 114(41):10870–10875.

Reilly, Barry and Robert Witt. 2011. "Disciplinary Sanctions in English Premiership Football: Is There a Racial Dimension?" *Labour Economics* 18(3):360–370.

Riaz, Sascha and Maik Hamjediers. 2025. "Anti-Immigrant Bias in the Choice Between Punitive and Rehabilitative Justice." *British Journal of Political Science* 55:e147.

Silberzahn, Raphael, Eric Luis Uhlmann, Daniel P. Martin, Pasquale Anselmi, Frederik Aust, Eli Awtrey, Štěpán Bahník et al. 2018. "Many Analysts, One Data Set: Making Transparent How Variations in Analytic Choices Affect Results." *Advances in Methods and Practices in Psychological Science* 1(3):337–356.

Storm, Rasmus K., Cornel Nesseler, Marthe Holum, Andreas Nygaard and Tor Georg Jakobsen. 2023. "Ethnic Discrimination in Scandinavia: Evidence from a Field Experiment in Women's Amateur Soccer." *Humanities and Social Sciences Communications* 10(1):230.

Valentim, Vicente. 2024. *The Normalization of the Radical Right: A Norms Theory of Political Supply and Demand*. Oxford: Oxford University Press.

Weiss, Chagai M. 2021. "Diversity in Health Care Institutions Reduces Israeli Patients' Prejudice Toward Arabs." *Proceedings of the National Academy of Sciences* 118(14):e2022634118.

Weiss, Chagai M. 2026. *Delivering Tolerance: How Unintended Institutional Inclusion Reduces Prejudice*. Princeton Studies in Political Behavior Princeton: Princeton University Press.

Weiss, Chagai M., Alexandra Siegel and Alexandra Scacco. 2026. "Out-Group Avoidance." *The Journal of Politics* 88(4):1318–1335.

White, Ariel R., Noah L. Nathan and Julie K. Faller. 2015. "What Do I Need to Vote? Bureaucratic Discretion and Discrimination by Local Election Officials." *American Political Science Review* 109(1):129–142.

Yokum, David, Anita Ravishankar and Alexander Coppock. 2019. "A Randomized Control Trial Evaluating the Effects of Police Body-Worn Cameras." *Proceedings of the National Academy of Sciences* 116(21):10329–10332.

Zschirnt, Eva and Didier Ruedin. 2016. "Ethnic Discrimination in Hiring Decisions: A Meta-Analysis of Correspondence Tests 1990–2015." *Journal of Ethnic and Migration Studies* 42(7):1115–1134.

Oversight and the Expression of Ethnic Bias: Evidence from Amateur Football

Supporting Information

Contents

A Data **A–2**

A.1 Sources, sample, and variables A–2

A.2 Descriptive statistics A–2

B Ethnicity classification **A–6**

C Institutional oversight across league tiers **A–12**

C.1 League system and observer assignment A–12

C.2 Referee grading, promotion, and the *Förderkader* A–14

D Empirical strategy details **A–16**

D.1 Within-player design details A–16

D.2 Within-league (promotion-cadre) design A–20

E Heterogeneity of the baseline gap **A–25**

E.1 Stability across the stages of a season A–26

F Within-referee analysis **A–29**

G Testing for differential dangerous play **A–31**

H Controlling for match-day audience **A–32**

I Controlling for referee experience **A–34**

J Placebo outcome outside the referee's control **A–36**

K Full estimates for the main-text tables **A–37**

A Data

A.1 Sources, sample, and variables

We retain only the senior men's league competitions, which excludes cup matches, friendlies, indoor and futsal competitions, regional tournaments, reserve and junior teams, women's competitions, and recreational divisions (*Ü32* and *Freizeit*). With the 2011/12 season, the *Bezirksoberliga* was discontinued. This was a league tier specific to Bavaria, situated between the *Landesliga* and the *Bezirksliga*. It was absorbed into an expanded *Landesliga*. However, as this predates our data set, it does not contribute to our observations. Therefore, the analysis sample spans exactly nine league competitions (Section C.1 details the eight-rank index, in which the B- and C-Klasse are pooled).

The opponent non-white share is calculated at the match-by-team level, among starters with a valid name-based classification, requiring a minimum of seven classified starters per team per match. Card-gap specifications are estimated based on complete cases: a player-match enters only when its player's ethnicity classification, the opponent's non-white share, and the officiating referee are all recorded. The player's own classification is available for almost all player-matches (Section B).

A.2 Descriptive statistics

Table A.1 reports summary statistics for the treatment, outcome, and control variables for the analysis sample.

Table A.2 reports the distribution of player-matches across all leagues in the analysis sample, broken out by season. Figure A.1 plots the same as a stacked bar chart across seasons. The panel is unbalanced by construction (lower tiers carry an order of magnitude more player-matches than the upper tiers), but each league is observed in every season. Table A.3 shows the unconditional rates of yellow, second-yellow, and straight-red cards by league, ordered from the top of the league

Table A.1: Summary statistics

	Mean	SD	Min	Max
Outcome variables				
Yellow card (any)	0.124	0.330	0	1
Second-yellow card (any)	0.006	0.079	0	1
Straight red card (any)	0.002	0.047	0	1
Total cards (count)	0.133	0.357	0	2
Treatment and controls				
Non-white player	0.104	0.306	0	1
Opponent non-white share	0.102	0.154	0.000	1.000
Home indicator	0.506	0.500	0	1
Sub-in indicator	0.153	0.360	0	1
Player-matches	22,411,568			
Matches	752,354			
Players	377,823			
Referees	14,058			

Notes: Sample: Bavarian senior men's leagues, 2013/14–2025/26 (2020/21 dropped). Yellow, second-yellow, and straight-red are binary indicators for receiving any such card; total cards is their count.

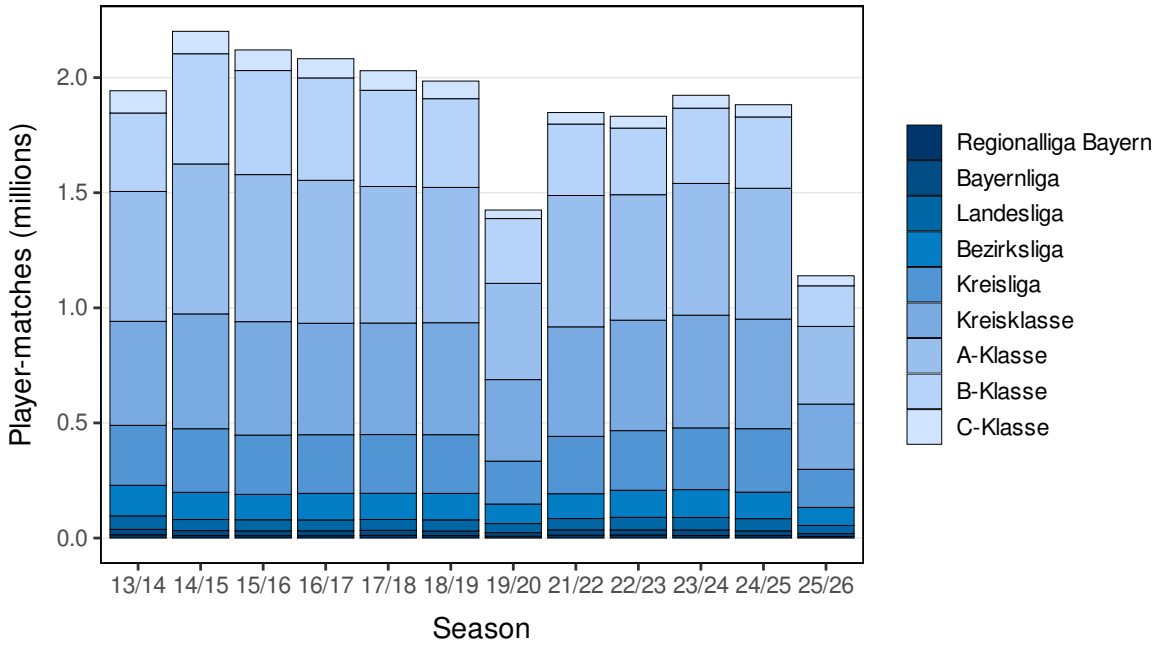
system down to the bottom divisions. Table A.4 reports match-day attendance by league, with the number of matches, the percentage of matches without attendance figures, and the median and mean recorded number of spectators.

Table A.2: Player-matches by league and season

	Player-matches (in thousands)											
League (× 1,000)	13/14	14/15	15/16	16/17	17/18	18/19	19/20	21/22	22/23	23/24	24/25	25/26
Regionalliga Bayern	14	11	10	10	12	11	8	13	13	11	11	7
Bayernliga	23	22	21	21	22	20	15	23	22	24	21	12
Landesliga	59	48	48	47	47	49	40	49	54	54	52	35
Bezirksliga	133	118	110	115	114	115	84	107	117	121	115	78
Kreisliga	261	276	258	254	255	255	187	250	259	269	276	165
Kreisklasse	451	499	493	485	485	486	354	476	480	490	476	284
A-Klasse	564	651	639	621	593	588	418	570	545	572	568	337
B-Klasse	341	479	452	445	418	385	281	310	289	327	310	177
C-Klasse	97	97	90	83	85	77	37	50	51	56	53	44
Total	1,943	2,201	2,120	2,082	2,030	1,985	1,425	1,848	1,832	1,923	1,882	1,139

Notes: Cells are player-match counts *in thousands* (1,943 = 1.94 million player-matches); 2020/21 dropped. B- and C-Klasse are shown separately but pooled into one tier rank in the analysis (Section C.1). The legacy *Bezirksoberliga* contributes no player-matches and is omitted (Section A.1).

Figure A.1: Player-match composition by league and season



Notes: Stacked bar chart of player-match counts by season, with bars colored by league and shaded from the top of the ladder (darkest) down to the bottom divisions (lightest). The 2020/21 season is dropped.

Table A.3: Unconditional card rates by league

League	Player-matches (millions)	Yellow (%)	Second yellow (%)	Straight red (%)
Regionalliga Bayern	0.13	11.7	0.33	0.29
Bayernliga	0.25	13.5	0.54	0.37
Landesliga	0.58	13.3	0.56	0.34
Bezirksliga	1.33	14.0	0.64	0.34
Kreisliga	2.96	14.2	0.69	0.29
Kreisklasse	5.46	13.6	0.69	0.22
A-Klasse	6.67	12.1	0.63	0.19
B-Klasse	4.21	9.9	0.50	0.15
C-Klasse	0.82	10.2	0.54	0.17
Pooled	22.41	12.4	0.62	0.22

Notes: Mean card rates per player-match by league, ordered from the *Regionalliga Bayern* (top tier) at the top of the table down to the *C-Klasse* at the bottom of the table. Yellow, second-yellow, and straight-red columns report means of the respective binary card indicators, as percentages. The pooled row shows the player-match-weighted mean across leagues. The 2020/21 season is dropped.

Table A.4: Match-day attendance by league

League	Matches	Missing (%)	Median (spectators)	Mean (spectators)
Regionalliga Bayern	3,760	2.5	600	832
Bayernliga	7,176	2.0	250	304
Landesliga	17,479	2.1	160	201
Bezirksliga	41,654	3.0	125	151
Kreisliga	96,344	3.6	100	119
Kreisklasse	180,459	4.4	80	92
A-Klasse	226,530	5.1	50	62
B-Klasse	149,501	6.7	32	38
C-Klasse	29,451	11.4	30	33
Pooled	752,354	5.1	60	83

Notes: Match-day attendance for the Bavarian senior men's league matches in the analysis sample (2013/14–2025/26; 2020/21 dropped), ordered from the top of the league system (*Regionalliga Bayern*) down to the bottom divisions (*C-Klasse*). Median and mean are over matches with a recorded figure; the mean is upper-winsorized at the 99th percentile within league to bound implausible entries, while the median is not. Attendance is not recorded for the professional tiers above the *Regionalliga Bayern*, as the Bavarian association does not administer those.

B Ethnicity classification

Onomastic procedure. For each player, first and last names are looked up in the Forebears database,[20] a cross-national name-frequency reference covering more than 200 countries and territories. The database returns a probability vector over potential jurisdictions of origin. We then select the one with the highest probability as the player's (most likely) origin. This is in line with the onomastic approach for identifying ethnicities reviewed by Mateos (2007).

Classification rule. We classify a player as white if their most likely jurisdiction of origin lies in geographic Europe, the United States, Canada, Australia, or New Zealand, and as non-white otherwise. The construct this rule targets is the outgroup status a referee can read off a player's name, not ancestry. Names carry racial information that observers act on (Gaddis, 2017), and correspondence audits of ethnic discrimination use them as the treatment for the same reason (Zschirnt and Ruedin, 2016).[21]

Validation against human-coded photographs. We validate the rule against manual codes of 1,736 player photographs.[22] The photographs were coded in two rounds. In the first, two coders independently coded the same random sample of 500 photographs; we use the 246 male players among them, since that round also sampled women's teams while the analysis sample is male. In the second, three coders divided a sample of 1,500 photographs, 1,490 of which carry valid codes; 196 were coded by all three to measure reliability. No coder took part in both rounds. In total, five coders looked at the photographs. The two first-round coders agreed on 94 percent of the photographs; in the second round, pairwise agreement ranges from 85 to 90 percent. A photograph's

[20] https://forebears.io

[21] The line runs through the settler societies: Argentina, Chile, Uruguay, South Africa, and Brazil do not enter the white set despite their European-descended populations. The photo codes mark some of these origins as genuinely ambiguous. Where the rule errs on this margin, it labels white-perceived players non-white, which attenuates the estimated gap; the trimming exercise below removes these ambiguous origins and the gap widens. The rule covers every origin in the name database's vocabulary; the full list is part of the replication materials.

[22] The validation photographs come from the DFB national platform, where player photographs are available; BFV records do not include photographs. Because the classification depends only on names, and Bavarian amateur players are drawn from the same broader population of German residents as the DFB-side player pool, the DFB validation is informative about classification quality on the BFV sample.

groundtruth is defined as the modal value of its valid codes, counting the 19 ties as white. The rule has no fitted parameter, so the photo codes are used as an out-of-sample benchmark. Table B.1 reports the agreement. Weighted by each origin's player-match count across all classified player-matches (24.2 million, before the complete-case restrictions of Section A.1), the rule classifies 90 percent of player-matches in line with the coders, with a sensitivity of 0.95 and a specificity of 0.53. Errors of this kind attenuate rather than inflate our estimates as misclassification in a binary regressor biases its coefficient toward zero provided the classification error is unrelated to the outcome and to the remaining regressors (Aigner, 1973). The trimming exercise below is consistent with that condition as removing the origins whose photo codes most conflict with the rule widens the estimated gap rather than shrinking it (Figure B.2). The card gaps we report are therefore, if anything, understated. The exercise targets the binary white-versus-non-white distinction the regressions rely on rather than finer ethnic categories.

Coder instructions. In both rounds, the coders were given the same short written brief. We reproduce it below in our own translation from the German, but with coder names, the access link, and scheduling details removed. Each coder independently viewed the photographs and recorded the ethnicity they perceived:

> *View the photograph with the corresponding ID and record the ethnicity you, as the viewer, would ascribe to the person. (1) Enter a 1 under* `white` *or* `non-white`*. (2) If you enter a 1 under* `non-white`*, also enter a 1 under one of* `asian`*,* `black`*,* `indian`*,* `latino hispanic`*, or* `middle eastern`*—whichever you would most likely ascribe. Enter exactly one 1 (*`white`*) or exactly two (*`non-white` *and one subcategory) per ID, no more and no fewer; record your most plausible guess; and work independently of one another, without discussing individual images.*

We use only the binary white/non-white judgment. Finer non-white categories were recorded but do not enter the analysis. Although the images are publicly availiable, coders were directed to work within secure institutional storage. We acknowledge that classifying people as white or non-white

can be discomforting. Coders were therefore instructed to take breaks at any time and were told that the task could be declined altogether beforehand or at any time whilst working on it.

Table B.1: Out-of-sample validation of the origin rule against the manual photo codes

	Unweighted (1)	Player-match weighted (2)
Accuracy	0.649	0.903
Sensitivity (white)	0.474	0.950
Specificity (non-white)	0.908	0.534
F1 (white)	0.617	0.945
Photos	1,736	1,736
coded white	1,038	—
coded non-white	698	—

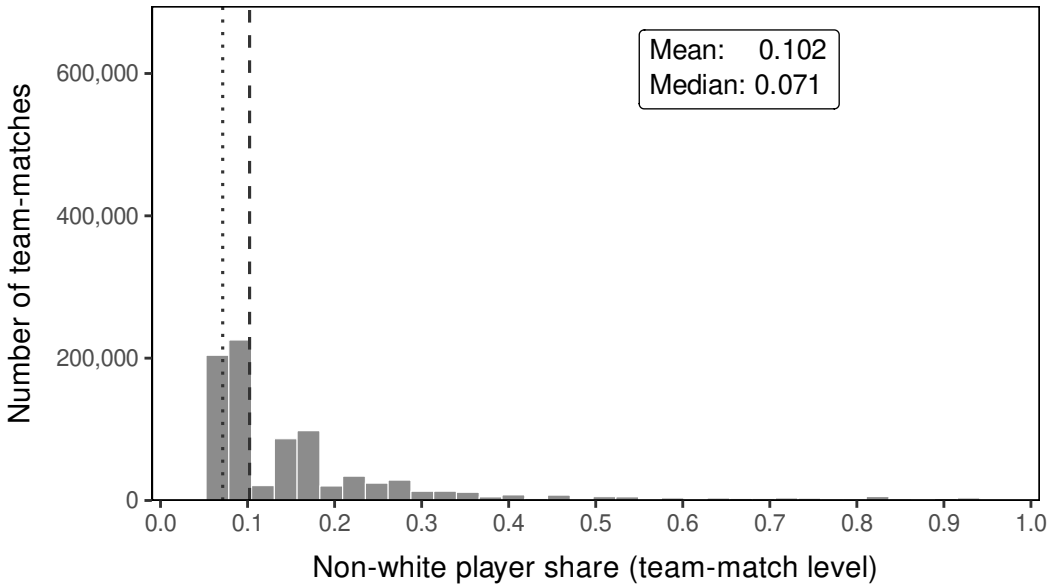
Notes: Agreement between the a-priori origin rule (geographic Europe and the four European-settler states → white) and the manual photo codes. Ground truth is the photo-level majority across the valid codes of its round (round 1: two coders, male photos only; round 2: three coders); ties count as white. No photo enters the classification, so both columns are fully out-of-sample. Column (1) weights every photograph equally; column (2) weights each photograph by its origin's player-match count across all classified player-matches, split evenly across the origin's photos. White is the positive class: sensitivity is the share of coder-white photos the rule classifies white, specificity the share of coder-non-white photos it classifies non-white.

Distribution of the non-white share. Figure B.1 reports the distribution of the player-level non-white share at the team-match level across the analysis sample.

Worked examples. The classification is based on the joint information in the player's first and last name. The procedure assigns to each name a vector of jurisdictional probabilities; the player's most likely origin is the highest-weight entry, and the binary indicator applies the origin rule described above. Table B.2 reports illustrative examples spanning common origins in the analysis sample, together with the share of each origin's validation photographs that the coders judged white.

Sensitivity to ambiguous origins. Each origin's validation photographs yield a continuous companion to the binary rule: the share of its photographs the coders judged white, reported in Table B.2. Misclassification risk concentrates where this photo evidence sits against the rule's label—rule-white origins whose photographs were often coded non-white, and rule-non-white origins

Figure B.1: Distribution of the team-match non-white share



Notes: Histogram of the team-match-level non-white share (mean of the binary non-white indicator across starters with a valid classification) across team-matches in the analysis sample with at least five classified starters. This descriptive cut is looser than the one behind the opponent non-white share covariate, which requires seven (Section A.1). Dashed: mean. Dotted: median.

whose photographs were often coded white, such as the Sudanese-origin players in Table B.2.

To assess the resulting measurement error, we re-estimate the baseline total-card specification (Equation 1) on a series of samples that successively remove conflicted origins, each step dropping the lowest-scored remaining origins on the rule-white side and the highest-scored remaining origins on the rule-non-white side, until the white cut lies above the non-white cut, so that every retained trim-eligible white origin scores above every retained trim-eligible non-white one. At no point we remove origins with fewer than five validation photographs (including those with none). German- and Turkish-origin players (shares 0.95 and 0.24) anchor the two groups throughout. The final, fully separated sample retains 96.7 percent of player-matches.

Under classical misclassification, the measured gap is attenuated toward zero, so removing conflicted origins should move the estimate moderately away from zero. A shrinking gap would instead indicate that the baseline result is driven by the ambiguous origins rather than attenuated by them. Figure B.2 shows the former: the gap widens steadily, from 0.69 to 0.86 percentage points,

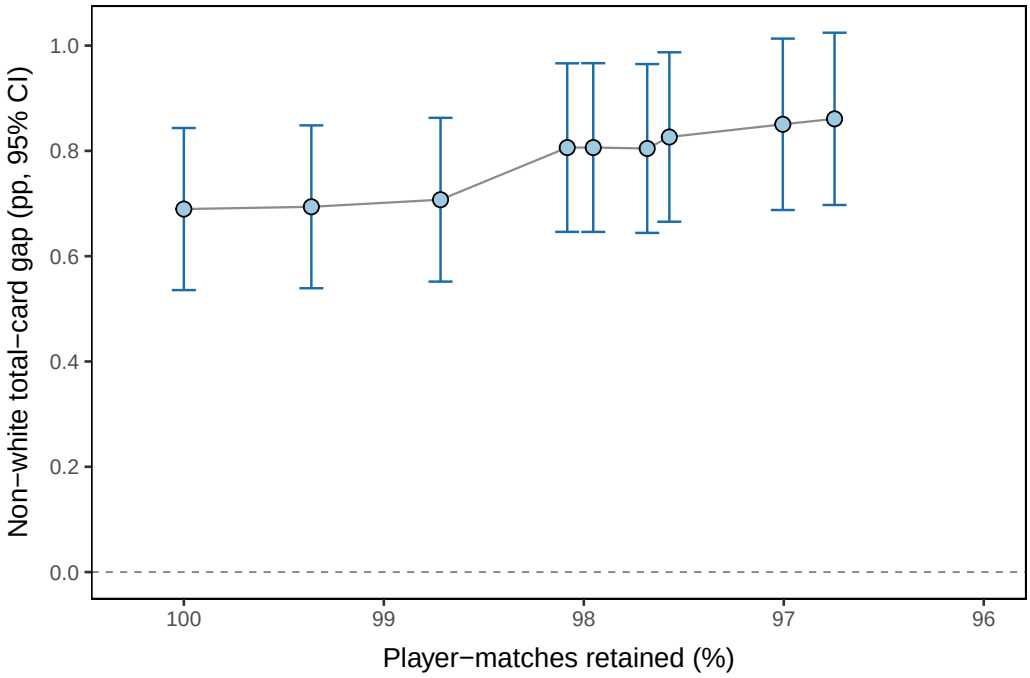
Table B.2: Worked examples of the name classification

Player name	**Dominant jurisdiction**	**Share coded white**	**Classification**
Dominik Bauer	Poland	1.00 (5)	white
Alexander Huber	Russia	1.00 (6)	white
Wolfgang Scholz	Germany	0.95 (102)	white
Maximilian Müller	Germany	0.95 (102)	white
Michael Aigner	Austria	0.90 (10)	white
Altin Gashi	Kosovo	0.89 (9)	white
Antonio Coppola	Italy	0.86 (7)	white
Endrit Hoxha	Albania	0.78 (9)	white
Mohamed Osman	Sudan	0.75 (8)	non-white
Mohamed Mohamed	Egypt	0.33 (9)	non-white
Mohamed Ali	Egypt	0.33 (9)	non-white
Mehmet Demir	Turkey	0.24 (21)	non-white
Emre Yilmaz	Turkey	0.24 (21)	non-white
Hoang Nguyen	Vietnam	0.22 (9)	non-white
Ali Hussaini	Nigeria	0.00 (6)	non-white
Ibrahim Ibrahim	Nigeria	0.00 (6)	non-white

Examples from the analysis sample. Dominant jurisdiction is the highest-weight entry of the name's jurisdictional-probability vector; classification follows the origin rule. Share coded white is the share of the origin's validation photographs whose subject the coders judged white (number of photographs in parentheses); it plays no part in the classification itself. Upper block: origins the rule classifies white; lower block: non-white. Common names are used to avoid identifying individuals.

as the conflicted origins are removed.

Figure B.2: The baseline non-white total-card gap as ambiguous origins are removed



Notes: Outcome: total cards (yellow + second-yellow + red). Each marker is the non-white total-card gap, in percentage points, from the baseline specification (Equation 1) re-estimated on the indicated subsample. Moving from left to right, successive points remove the origins whose photo share most conflicts with their rule classification—the lowest-scored rule-white and the highest-scored rule-non-white origins—until the retained trim-eligible groups are fully separated on the photo scale. Origins with fewer than five validation photographs are never removed; German- and Turkish-origin players are retained throughout as anchors. The most aggressive trim removes 3.3 percent of player-matches. The leftmost point (100 percent retained) is the untrimmed baseline; the dashed line marks a zero gap. Bars: 95 percent confidence intervals; standard errors clustered at the player level.

C Institutional oversight across league tiers

C.1 League system and observer assignment

Across our 2013/14–2025/26 panel the senior men's league system administered by the *Bayerischer Fußball-Verband* (BFV) comprises nine league competitions, which we index by eight ranks: the B- and C-Klasse, the two lowest competitions, are pooled into one rank, and the remaining seven each occupy their own.[23] At the top of the Bavarian league system is the *Regionalliga Bayern*, the fourth tier of the broader German football league system; below it sit the *Bayernliga* (split into Nord and Süd sub-divisions), the *Landesliga* (Nordwest, Nordost, Mitte, Südwest, Südost), the *Bezirksliga*, the *Kreisliga*, the *Kreisklasse*, and the regional A-, B-, and C-Klasse competitions at the lowest tiers. Geographic and tier structures interact: higher-tier divisions (*Regionalliga* through *Landesliga*) draw clubs from across Bavaria, while lower-tier divisions are organized within seven *Bezirke* (Oberbayern, Niederbayern, Oberpfalz, Oberfranken, Mittelfranken, Unterfranken, and Schwaben).

The institutional environment in which referees officate matches varies across league tiers. Every fixture from the *Regionalliga Bayern* down to the C-Klasse is administered by the same federation under the same disciplinary code and the same set of rules. However, the oversight regime under which referees actually officiate differs along three dimensions: whether a graded observer is present at the match, how often referees are formally evaluated per season, and the consequences of poor evaluations regarding promotion or relegation.

The federation partitions appointment, observation, and grading authority by tier. The *Verbands-Schiedsrichter-Ausschuss* (VSA) oversees the *Verbandsebene* (the *Regionalliga Bayern*, the *Bayernliga*, and the *Landesliga*). The *Bezirks-Schiedsrichter-Ausschuss* (BSA) is responsible for the *Bezirksliga*; and the *Kreis-Schiedsrichter-Ausschuss* (KSA) for the *Kreisliga*. The *Gruppen-Schiedsrichter-Ausschuss* (GSA) oversees the *Kreisklasse* and the regional A-, B-,

[23] A tenth competition, the *Bezirksoberliga*, historically sat between the *Landesliga* and the *Bezirksliga*. However, it was discontinued before our panel began, so the *Landesliga* and *Bezirksliga* are adjacent rungs in our panel.

and C-Klasse. A referee's performance class determines the highest tier to which they may be assigned, but any of them might be subject to officiating in lower tiers—and therefore to more than one oversight regime—as they might get assigned by their local GSA and the committee responsible for the higher tier. Observers are recruited from the referee community itself, and observe two tiers below the one they are officiating. Rormer or retired referees are drafted as well. According to the BSA Oberbayern *Beobachtungsrichtlinien*, a new candidate for the *Bezirksliga* observer pool must not have reached the age of 65.

In the *Regionalliga Bayern*, the VSA assigns a grading observer (*Schiedsrichterbeobachter*) to every match, either on site or via the *Home-Office-Verfahren* by video. A referee's seasonal mean grade determines retention, relegation, and nomination to the DFB list. Referees in the *Bayernliga* are observed and graded by the same VSA corps and on the same basis. The *Landesliga*, also administered by the VSA, is observed and graded by the same corps; however, public materials do not specify a per-match observation target for it. In the *Bezirksliga*, the Oberpfalz *Qualifikationsbestimmungen* call for observer attendance at every match of a *Bezirksliste* referee only up to a season cut-off date but set no mandatory number of observations. The Oberbayern guidelines state a non-binding target of five to six observations per referee per season. Availability is the binding constraint. The resulting partial coverage informs tier movement at season's end.

Below the *Bezirksliga*, the observation regime changes in kind. Routine observer attendance starts at the *Kreisliga*, where the KSA appoints referees, but observation is sparse: non-cadre referees are observed only a few times per season. The only source of systematic repeated observation is the BFV's small promotion cadre. In the *Kreisklasse*, observation is even less frequent as it is essentially limited to promotion candidates. In the A-, B-, and C-Klasse observation is minimal to absent, occurring mostly as an ad hoc accompaniment of newcomers.[24]

Although realized coverage varies from rung to rung, the oversight *regime* changes discretely.

[24] Sources: BFV-*Schiedsrichterordnung* (Stand 25.11.2024, §2, §4, §5, §10, §11); BFV-*Regionalligaordnung* (Stand 01.07.2025, §29, §36, §40; these three provisions are verbatim unchanged in every version from the league's founding in 2012/13, where the numbering differs, through 2026); BSA Oberbayern *Beobachtungsrichtlinien* (2018) and *Durchführungsbestimmungen Beobachtungen* 2025/26; BSA Oberpfalz *Qualifikationsbestimmungen* (2023, §2.3, §2.7); BFV/schiedsrichter.bayern, "Der Beobachter" (07.11.2022).

This change is most pronounced at two points. The upper one lies between the *Bayernliga* and the *Landesliga*, where observation is no longer guaranteed per match. The lower point falls between the *Bezirksliga* and the *Kreisliga*, where no longer the BSA, but the KSA, appoints observers. Additionally, routine list-wide observation becomes sparse, cadre-only observation. The regime is also not uniform from one rung to the next with the *Regionalliga Bayern* being covered more fully and involving higher grading stakes than the *Bayernliga*, one rung below. The observed *Kreisliga* is covered more closely than the unobserved *Kreisklasse* and the A-, B-, and C-Klasse beneath it. Our main within-player design leverages this rank-level variation directly, whereas our supplementary within-league specification (Section D.2) contrasts the systematically observed promotion cadre with other *Kreisliga* referees at the same rank, and the unobserved bottom divisions net out promotion-track selection.

C.2 Referee grading, promotion, and the *Förderkader*

The observer report concludes with a single, numerical assessment. The association centrally sets the assessment instrument, which the districts adopt, so that referees are graded on the same criteria throughout. The criteria consist of match-relevant individual decisions, the quality of match control, disciplinary control, personal presence, fitness and positioning, and teamwork with the assistants. The most consequential decisions have the highest impact on the final grade because individual criteria are differently weighted. Within this framework, districts set their own roster sizes, promotion quotas, and observation targets. Referees are evaluated based on the mean of their assessments, irrespective of the total number of observations. How often a referee is observed depends on availability and the referee's performance.[25]

The Bavarian referee corps is organized in performance classes (*Leistungsklassen*) that mirror the league tier structure. Promotion and relegation between officiating tiers are governed by the

[25] Sources: BFV *Schiedsrichterordnung* (25.11.2024, §§4, 5, 11); BFV/VSA *Durchführungsbestimmungen SR-Qualifikation* 2023/24 and 2025/26 (§3.3); BSA Oberbayern *Beobachtungsrichtlinien* (2018) and *Qualifikationsrichtlinien* (2022, §3.1); BSA Oberpfalz *Qualifikationsbestimmungen* (2023, §§1.2, 2.3) and *Durchführungsbestimmungen Beobachtungen* (2022, Ziff. 14); BSA Unterfranken *Erläuterungen zum Beobachtungs- und Coachingbogen* (2026).

seasonal mean grade, supplemented by a qualification course (*Qualifikationslehrgang*), an annual rules test (*Regeltest*; 15 questions, 25-point minimum), and an annual fitness test (*Leistungsprüfung*; 40-meter sprints and an interval test of alternating 75-meter runs and 25-meter walks, with volumes graded by class). At the *Bezirksliga* level the BSA decides promotions and relegations; the VSA decides at the *Verbandsligen*. The size of each performance class is fixed (*Sollzahl*), and it is maintained by sliding relegation (*gleitender Abstieg*): once promotions and voluntary departures are accounted for, a sufficient number of referees are relegated from the bottom of the seasonal ranking to restore the target count. Promotion from the *Kreisliga* to the *Bezirksliga* is rationed by quotas that vary across *Bezirke*: in Oberpfalz, as a rule, three referees per *Bezirk* per season, with each *Kreis* entitled to nominate one, while Oberbayern instead allocates promotions *Kreis* by *Kreis* (e.g., three from München, two from Zugspitze). The *Kreisliga Förderkader* (discussed below) is the standard channel in either case. Decisions are made under *pflichtgemäßes Ermessen* (sound discretion); no referee has a legal claim to a particular performance class. The *Beobachtungsrichtlinien* also cap observers at 70 years old at the start of a season. Experienced referees have an exit of their own: the BSA may nominate *Bezirksliga* referees with at least five years at that level or above and past age 30 (or entering the coaching track) outside the regular qualification, and on application they then officiate without observation for three years, renewable annually.[26]

The *Förderkader* is the BFV's talent-development pipeline for promising referees officiating in the *Kreisliga* and below. Cadre members are observed repeatedly over the season and ranked for promotion to the *Bezirksliga* at year-end. The cadre is organized by *Kreis* (district group) and by *Bezirk*, and its size varies markedly across districts: the Zugspitze district of Oberbayern runs a *Kreisliga Förderkader* of 12 referees (three from each of four subgroups), of whom the top two are promoted to the *Bezirksliga* at the end of the season. Cadre members are nominated by the local *Schiedsrichter-Ausschuss* on the basis of performance in the prior season. This produces observed matches at *Kreisliga* matches which would otherwise carry almost none: cadre referees

[26]Sources: BSA Oberpfalz *Qualifikationsbestimmungen* (2023, §§1.1–1.2, 2.2, 2.4, 2.7, 3.2, 4, 5.1); BSA Oberbayern *Qualifikationsrichtlinien* (2022, §3.2, §5.1, §6.3—*Sonderliste C*, a comparable observation-free list) and *Beobachtungsrichtlinien* (2018, Ziff. 6 j); BFV "Leistungsprüfung – Übersicht" (Zugspitze district).

are observed repeatedly, so a small share of *Kreisliga* matches in our sample is in fact observed. Most *Kreisliga* matches nonetheless go unobserved, and the *Kreisklasse*, A-, B-, and C-Klasse competitions are not part of the routine observer schedule at all.[27]

D Empirical strategy details

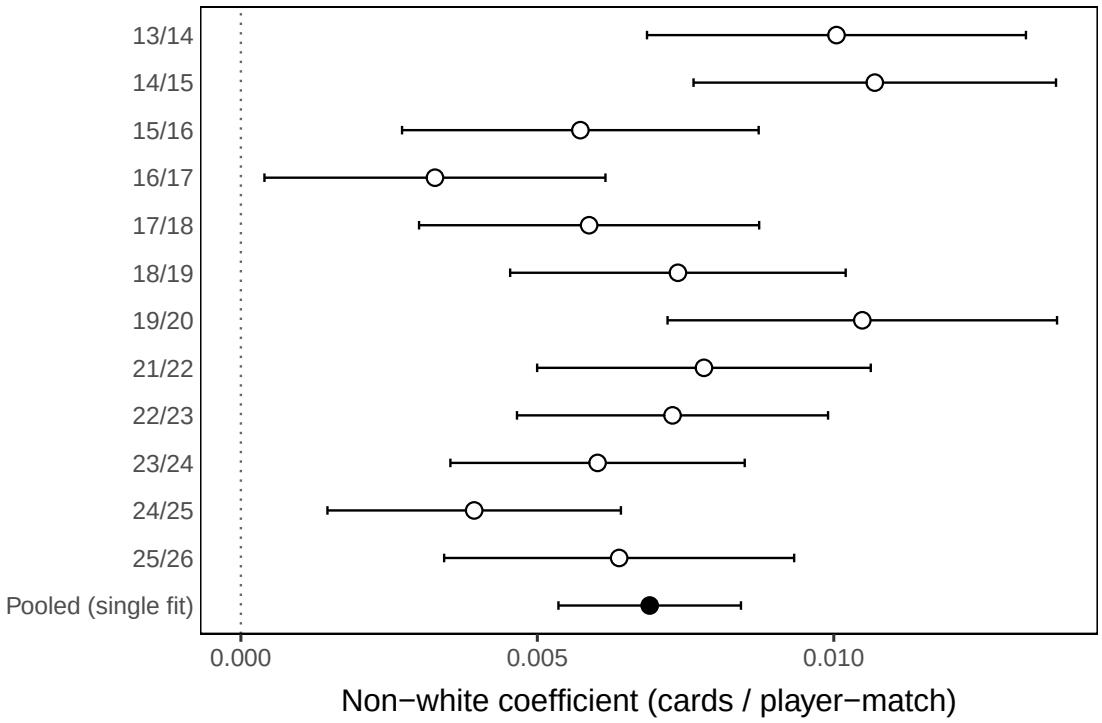
We estimate Equation 1 on the full pooled sample in a single OLS fit, absorbing the club×league×season, opponent, and referee fixed effects jointly, with standard errors clustered at the player level. As a robustness check, we also fit the specification season by season: Figure D.1 plots the per-season white/non-white total-card gap with 95 percent confidence intervals alongside the single-fit pooled estimate. The per-season gaps are stable across seasons, and the pooled estimate is representative of them. It is not driven by a single season.

D.1 Within-player design details

The signed pooled estimate reported in Section 3.2 replaces the four-class move variable in Equation 2 with two signed regressors that together impose the mirror-image symmetry OVR-UP $=$ $-$OVR-DN: a league-rank step, coded $+1$ for a one-league promotion and -1 for a relegation, entered for *every* move, and an oversight increment coded the same way but entered only for moves that also cross an oversight boundary. The model retains all lower-order interactions, including a common post $\times$ NW$_i$ term, so the symmetry is the only restriction it adds to Equation 2; the post-move shifts in the gap for pure-rank promotions and relegations remain unrestricted. Because an oversight-crossing move changes rank as well, it loads on both regressors; the rank step then absorbs the one-league rank slope common to all moves, and the coefficient on the oversight increment is the per-step oversight effect net of that rank slope (the third difference of Table 2), recovered directly with its own clustered standard error. Column (1) of Table 4 reports this esti-

[27] Sources: BFV Schongau / Zugspitze, "Aufstieg und Beobachtung – Übersicht"; BFV Ammersee/FFB, "Spitzen-Schiris halten ihre Positionen"; BFV Bezirk Unterfranken, "BSA Stützpunkt Mai 2022" (*3-Säulen-Förderkonzept*); BSA Oberpfalz *Qualifikationsbestimmungen* (2023, §2.1, §2.3).

Figure D.1: Per-season white/non-white total-card gap



Notes: Outcome: total cards (yellow + second-yellow + red). Per-season fits of the baseline specification (Equation 1) across the 12 seasons (2013/14–2025/26; 2020/21 dropped), shown alongside the single-fit pooled estimate (Equation 1 estimated on the full pooled sample). Bars: 95 percent confidence intervals; standard errors clustered at the player level. Pooled $N = 22.4$ million player-matches.

mate on the move window $(t, t+1)$; Column (2) repeats it on the two pre-move seasons $(t-1, t)$ as a pre-trend placebo.

Because the signed estimator imposes the mirror-image symmetry for power, we check that it is consistent with the data rather than merely assumed. Estimating the four move classes of Equation 2 without the restriction recovers the two oversight contrasts separately—OVR-UP for promotions, which add oversight, and OVR-DN for relegations, which remove it. Table D.1 reports both directions on the move window and the pre-trend placebo, together with a Wald test of the pooling restriction $\text{OVR-UP} + \text{OVR-DN} = 0$.

We assign each player in each season to the club for which they make the most league appearances, and set the tier of that club-season to the median rank of those appearances. The within-player design (Section 2.3) then draws on players whose club is promoted or relegated by one

Table D.1: Within-player triple-difference: oversight effect by direction

	Total cards	
	Move window	Pre-trend placebo
	(1)	(2)
NW × Oversight (pooled)	−0.0060**	−0.0004
	(0.0028)	(0.0031)
promotion (OVR-UP)	−0.0063*	−0.0017
	(0.0034)	(0.0038)
relegation (OVR-DN)	0.0059	−0.0012
	(0.0042)	(0.0045)
Symmetry test (p)	0.94	0.61
NW × Pure rank	0.0007	0.0026
	(0.0017)	(0.0020)
Mean of DV	0.147	0.146
Player-matches	5,434,230	4,457,065
Controls	✓	✓
Season FE	✓	✓
Player FE	✓	✓

Notes: Outcome: total cards (yellow + second-yellow + red). The within-player per-step oversight estimator of Section D.1, net of one league-step of rank, with the mirror-image restriction relaxed. NW × Oversight (pooled) imposes OVR-UP = −OVR-DN and reproduces Column (1) of Table 4; the promotion (OVR-UP) and relegation (OVR-DN) rows instead estimate the two directions freely from the four move classes of Equation 2. Symmetry test reports the p-value for OVR-UP + OVR-DN = 0. Column (1) uses the move window $(t, t+1)$; column (2) the pre-trend placebo $(t-1, t)$. All within player. Standard errors clustered at the team level. Signif. Codes: ***: 0.01, **: 0.05, *: 0.1.

league between two consecutive seasons, who stay at that club across the move, and who make at least three league appearances for it in each of the two bracketing seasons; all of a qualifying player's appearances in the two seasons enter the estimation, with the later season marked as post-move. About 7 percent of these appearances are for the club's other senior side, its first team or its reserves, between which players rotate within a season. Every such move falls into one of four classes (an oversight-crossing step up or down, or a step up or down within the never-observed bottom divisions, where rank changes but oversight does not) and spans one rung of the league ladder. Tables D.2 and D.3 report how the identifying variation is distributed across these classes,

rungs, directions, and seasons, counting for each cell both the distinct club promotion and relegation events and the player-match appearances they contribute to the estimation window.

Table D.2: Within-player identifying moves, by rung and direction

	Promotion (up)		Relegation (down)	
Move (league rungs)	Club-moves	Pl.-matches	Club-moves	Pl.-matches
Oversight-crossing				
Regionalliga–Bayernliga (1–2)	29	20,093	29	14,395
Bayernliga–Landesliga (2–3)	67	47,590	67	34,649
Landesliga–Bezirksliga (3–5)	193	138,319	174	95,391
Bezirksliga–Kreisliga (5–6)	585	367,450	563	295,041
Kreisliga–Kreisklasse (6–7)	1,318	766,936	1,219	612,315
Subtotal	2,192	1,340,388	2,052	1,051,791
Pure-rank (bottom divisions)				
Kreisklasse–A-Klasse (7–8)	2,028	1,070,355	1,795	815,884
A-Klasse–B/C-Klasse (8–9)	1,683	706,810	1,214	449,002
Subtotal	3,711	1,777,165	3,009	1,264,886
Total	5,903	3,117,553	5,061	2,316,677

Notes: The identifying moves of the within-player design (Section 2.3), by move class, the league rung the move spans, and direction. A club-move is one club promotion or relegation event; player-matches are appearances in the move window $(t, t+1)$ and sum to the estimation sample of Table 4. A club fields one league per season, so each club-move is assigned the rung and direction of its squad's predominant tier change. The pure-rank moves run within the never-observed bottom divisions (*Kreisklasse* and below); the *Landesliga–Bezirksliga* step is a single rung in our panel, the historical tier between them being empty (Section C.1).

The lowest oversight boundary. The *Kreisliga–Kreisklasse* rung supplies six in ten oversight-crossing club-moves (Table D.2), and it is the boundary at which the change in oversight is smallest: below it no match carries a routine observer, above it only the promotion cadre is observed routinely (Section C.1). Table D.4 re-estimates the signed per-step estimator with that rung dropped, so that the oversight effect is identified from the boundaries between ranks 1 and 6 only, from 1,707 rather than 4,244 club-moves. The point estimate is unchanged (-0.0070 against -0.0060), with a larger standard error (0.0037) given the smaller set of identifying moves.

Table D.3: Within-player identifying moves, by season

Season	Ovr↑	Ovr↓	Rnk↑	Rnk↓	Total
Panel A: distinct club-moves					
13/14	197	196	348	309	1,050
14/15	210	215	385	345	1,155
15/16	214	208	394	359	1,175
16/17	209	195	379	312	1,095
17/18	208	186	364	287	1,045
18/19	206	194	336	261	997
19/20	145	134	245	216	740
21/22	213	175	347	233	968
22/23	200	192	321	240	953
23/24	204	178	299	231	912
24/25	186	179	293	216	874
Total	2,192	2,052	3,711	3,009	10,964
Panel B: player-matches (in thousands)					
13/14	121	103	172	132	528
14/15	132	116	194	155	597
15/16	134	107	197	156	594
16/17	133	103	195	136	567
17/18	131	98	183	130	542
18/19	115	90	143	97	445
19/20	75	57	93	70	295
21/22	136	93	177	101	507
22/23	132	103	160	108	503
23/24	131	102	150	104	487
24/25	101	79	115	75	370
Total	1,340	1,052	1,777	1,265	5,434

Notes: The same identifying moves by the pre-move season t of the bracketing pair $(t, t+1)$. Panel A counts distinct club promotion and relegation events; Panel B counts player-match appearances in the move window, in thousands. Columns are the four move classes of Equation 2: oversight-crossing (Ovr) and pure-rank (Rnk, within the bottom divisions), each up and down. The 2020/21 season is absent from the panel.

D.2 Within-league (promotion-cadre) design

Our supplementary within-league design holds league rank fixed instead of netting it out: we compare the white/non-white card gap among referees who all officiate at the same rank, the *Kreisliga*,

Table D.4: Within-player triple-difference: dropping the lowest oversight boundary

	Total cards	
	Baseline	Drop 6–7 rung
	(1)	(2)
NW × Oversight	−0.0060**	−0.0070*
	(0.0028)	(0.0037)
NW × Pure rank	0.0007	0.0007
	(0.0017)	(0.0018)
Mean of DV	0.147	0.143
Player-matches	5,434,230	4,057,792
Oversight-crossing club-moves	4,244	1,707
Pure-rank club-moves	6,720	6,720
Controls	✓	✓
Season FE	✓	✓
Player FE	✓	✓

Notes: Outcome: total cards (yellow + second-yellow + red). The within-player per-step oversight estimator of Section D.1, net of one league-step of rank, on the move window $(t, t+1)$. Column (1) reproduces Column (1) of Table 4. Column (2) drops the *Kreisliga–Kreisklasse* rung (6–7), so oversight-crossing moves are those between ranks 1 and 6; the pure-rank moves are unchanged. Club-moves count distinct club promotion and relegation events, assigned to a rung as in Table D.2. All within player. Standard errors clustered at the team level. Signif. Codes: ***: 0.01, **: 0.05, *: 0.1.

but differ in whether they are observed. The BFV's small *Förderkader* (promotion cadre) is observed repeatedly across the season, while other *Kreisliga* referees are observed only rarely if at all. The *Kreisliga* is the only rank that offers this contrast: higher up the ladder observation is uniform across referees, and below it no match carries a routine observer.

Promotion-track referees are selected for advancement, and so may be more professional or less biased to begin with; a bare cadre-versus-other referees comparison would confound oversight with this positive selection. Three differences identify the observation effect (see Table D.5). First, the white/non-white card gap. Second, how that gap differs between promotion-track and stayer referees. Third, the same promotion-track-versus-stayer contrast in the never-observed bottom divisions, where the promotion track still marks the positively selected referees but no match is observed. The design's key identification assumption is that promotion-track selection shifts the

gap equally in the *Kreisliga* and the bottom divisions, so that differencing the two track-versus-stayer gaps removes it and leaves the effect of oversight.

Table D.5: Within-league triple-difference design

	$\Delta = \bar{y}_{\mathrm{NW}} - \bar{y}_{\mathrm{W}}$		
	Kreisliga	**bottom divisions**	*Kreisliga* − bottom
Promotion-track	$b+s+\ell+\beta$	$b+s$	$\ell+\beta$
Stayer	$b+\ell$	b	ℓ
Track − stayer	$s+\beta$	s	$\boxed{\beta}$

Notes: Each cell is the white/non-white card gap $\Delta = \bar{y}_{\mathrm{NW}} - \bar{y}_{\mathrm{W}}$ for that group of player-matches, decomposed into additive components. The baseline b is the gap shared by every cell. The row term s is the referee-selection confound: promotion-track referees are positively selected, so their gap differs from stayers' regardless of league. The column term ℓ is the level difference in the gap between the *Kreisliga* and the bottom divisions. The oversight effect β is present only in the shaded treated cell.

We build the sample at the referee-season level. We do not observe cadre membership directly, so we proxy it by the promotion track.[28] For each referee in a season, we set their focal tier to the modal rank of the matches they officiate, mark them as on the promotion track if that focal tier rises by one rank the next season and as a stayer if it holds, and keep only the two types whose focal tier is the *Kreisliga* or a bottom division. Keeping each referee's matches at their focal tier then sorts every player-match into one of the four groups—about nine million player-matches in all. On these we estimate the fully saturated triple-difference

$$\begin{aligned} y_{icjr} = {} & \alpha_s + \alpha_\ell + \alpha_c + \beta_1 \,\mathrm{NW}_i + \beta_2 \,\mathrm{Track}_r + \beta_3 \,\mathrm{Obs}_r \\ & + \beta_4 \,(\mathrm{NW}_i \times \mathrm{Track}_r) + \beta_5 \,(\mathrm{NW}_i \times \mathrm{Obs}_r) + \beta_6 \,(\mathrm{Track}_r \times \mathrm{Obs}_r) \\ & + \underbrace{\beta}_{\text{observation effect}} (\mathrm{NW}_i \times \mathrm{Track}_r \times \mathrm{Obs}_r) + \gamma' X_{icjr} + \varepsilon_{icjr}, \end{aligned} \tag{3}$$

where α_s, α_ℓ, and α_c are season, league, and club fixed effects; $\mathrm{Track}_r = 1$ for a promotion-track referee and 0 for a same-tier stayer; and $\mathrm{Obs}_r = 1$ in the *Kreisliga* and 0 in the bottom

[28] Because we proxy cadre membership by realized promotion, a referee who is genuinely in the promotion cadre but is not promoted the following season is coded as a stayer. Such referees behave like cadre members but sit in the comparison group, so each such case shrinks the estimated contrast. The error works against the effect we report.

divisions. Because Obs_r is a function of the league, the league fixed effects absorb its main effect β_3; the interactions that carry it remain identified because non-white status and promotion track both vary within league. The triple interaction β is the estimated effect of observation. Standard errors are clustered at the referee level.

A referee-season enters the within-league design only when it is followed by a consecutive season, against which the promotion-track and stayer labels are defined. The within-league design's key identification assumption is plausible on institutional grounds because the two tiers promote referees through different channels. Below the *Kreisliga* no match is observed, so advancement cannot run on graded officiating performance; the *Kreis-Schiedsrichter-Ausschuss* instead moves referees up on administrative criteria: availability and reliability, tenure and age progression, attendance at monthly training evenings, the annual rules and fitness test, and the assignor's discretion. None of these measures officiating quality or impartiality. Graded observation of officiating enters only at the *Kreisliga*, through the cadre. The bottom-division promotion-track contrast therefore captures the administrative selection common to both tiers, and the triple isolates what graded observation adds on top.[29]

Table D.6 reports the distinct-referee counts in each focal-tier and promotion-track cell of the design.

[29]The BFV does not codify a promotion rule for the bottom divisions; the regional rulebooks specify grading-based promotion at the *Kreisliga* (through the cadre) and above, but not for the *Kreisklasse* and the A-, B-, and C-Klasse below it (Section C.2). This characterization of advancement in the bottom divisions therefore follows from the DFB *Schiedsrichterordnung* (§4: promotion depends on performance, with no entitlement, though what counts as performance differs by tier) and the DFB's own description of lower-class advancement through club and coach feedback rather than graded observation; we treat it as institutional inference, not a documented rule.

Table D.6: The promotion cadre by focal tier and promotion track

Focal officiating tier	Stayer	Promotion-track
Kreisliga	1,894	762
Kreisklasse	2,522	1,365
A-Klasse	3,903	2,649
B/C-Klasse	3,648	2,608

Notes: Distinct-referee counts in the real design (the focal officiating tier is inferred from match assignments, not observed directly; see text). Stayer = modal tier unchanged the next season; Promotion-track = modal tier one rank higher. A referee can enter more than one cell across seasons, so cells need not sum to a fixed total. *Kreisliga* is observed; the *Kreisklasse* and the A/B/C-Klasse are the unobserved bottom divisions. The Table D.7 placebo drops the *Kreisliga* and reuses the bottom divisions only.

Table D.7: Within-league triple-difference

	Total cards			
	Promotion cadre		Placebo	
	(1)	(2)	(3)	(4)
NW × Track × Observed	−0.0060*	−0.0068*	−0.0043	−0.0031
	(0.0035)	(0.0035)	(0.0031)	(0.0030)
Mean of DV	0.128	0.128	0.124	0.124
Player-matches	8,958,762	8,958,762	7,716,685	7,716,685
Controls	✓	✓	✓	✓
Season FE	✓	✓	✓	✓
League FE	✓	✓	✓	✓
Club FE	✓	✓	✓	✓
Referee FE		✓		✓

Notes: Outcome: total cards (yellow + second-yellow + red). Columns (1)–(2) are the promotion cadre triple, contrasting the observed *Kreisliga* with the unobserved bottom divisions (*Kreisklasse* and below); columns (3)–(4) repeat the bottom-division placebo, with a fake observation boundary imposed one rung below the real one. Columns (2) and (4) add referee fixed effects, which absorb referee-level card propensity. Standard errors clustered at the referee; controls: opponent non-white share, home, sub-in. A referee's tier is inferred from the matches they officiate, not observed directly. See SI Section D.2 for the design and identifying assumption. Signif. Codes: ***: 0.01, **: 0.05, *: 0.1.

E Heterogeneity of the baseline gap

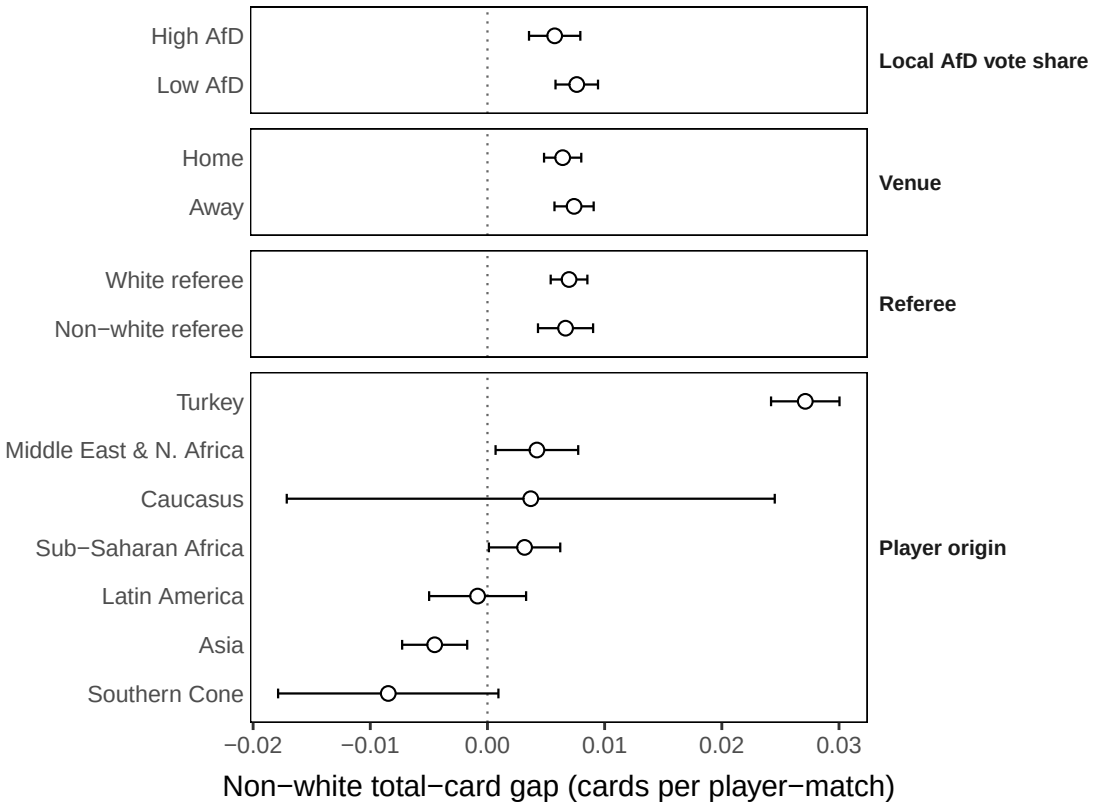
We subset the pooled baseline gap of Section 3.1 along several dimensions: the non-white players' origin region, the officiating referees' ethnicity, whether it is a home or away match, the local far-right vote share, and the rung of the league ladder at which the match is played. The analysis sample is split on the moderator and we re-fit the baseline total-card specification (Equation 1) within each cell in a single OLS fit as in the pooled baseline (Section D); standard errors are clustered at the player level. Every cell re-fits the full specification, so the club×league×season, opponent, and referee fixed effects remain. For player origin, we group the non-white origin jurisdictions (Section B) into world regions and pool each region in turn with the common white baseline, so the estimated coefficient is that region's total-card gap against white players.[30] The far-right rows attach to each match its home-venue municipality's AfD vote share at the most recent federal election and split matches at the median. Figure E.1 reports the estimates for player origin, referee ethnicity, and match context; we turn to the league-rank cut, and to the gap's stability across the season, below. These subgroup cuts are descriptive and exploratory; we apply no multiple-comparison adjustment and draw no confirmatory inference from any individual subgroup.

The remaining cut is by league rank, the dimension along which oversight varies. The cell here is a single rung of the ladder: we restrict the sample to one rung at a time and re-fit the gap within it. Within a single rung the club×league×season fixed effects reduce to club×season fixed effects, so the coefficient is the within-club-season, within-opponent, within-referee white/non-white total-card gap among players competing at that rung. The rungs are the eight tier ranks used throughout (Section C.1), with the B- and C-Klasse pooled into the single bottom rank and the legacy *Bezirksoberliga* contributing no player-matches. Figure E.2 plots the per-rung gap with 95 percent confidence intervals, ordered from the top of the ladder down to the bottom divisions.

Like the pooled baseline (Section 3.1), these per-rung gaps are descriptive rather than causal. Within a rung the comparison still rests on players who may differ in unobservable characteristics,

[30] A small residual group of non-white players from scattered origins outside these regions (Pacific islands and some European-administered overseas territories outside geographic Europe) is estimated but not shown in the figure.

Figure E.1: Heterogeneity of the non-white total-card gap



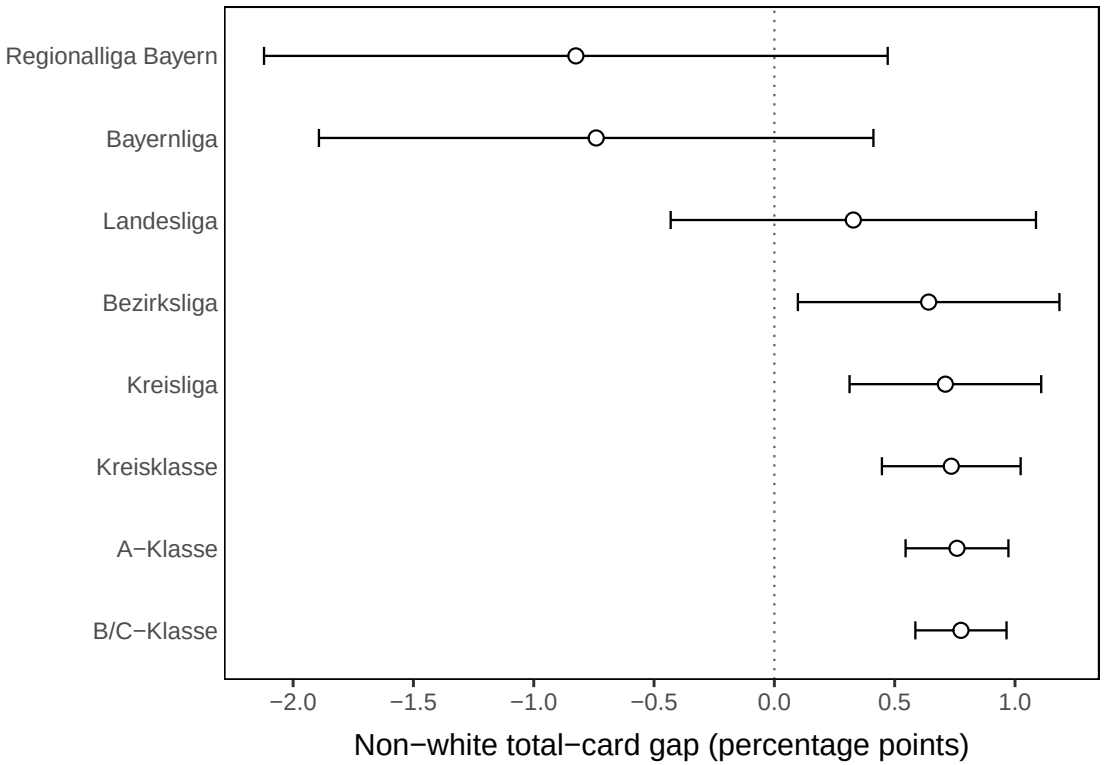
Notes: Outcome: total cards (yellow + second-yellow + red). Markers are point estimates of the non-white total-card gap within each subsample, with 95 percent confidence-interval bars and a dotted line at zero. Origin rows compare each region against the common white baseline; the AfD rows split matches at the median of the home-venue municipality's AfD vote share at the most recent federal election. Standard errors clustered at the player level.

and the comparison across rungs additionally bundles league rank with oversight and with the selection of players and referees into each tier. The figure therefore describes where the gap sits along the ladder; the effect of changing a player's oversight regime is what our designs (Section 3.2) are built to isolate.

E.1 Stability across the stages of a season

We also examine whether the gap shifts over the course of a season. We split each season's matches into three stages of roughly equal size and re-fit the baseline specification (Equation 1) within each stage, separately for each card outcome, in a single OLS fit as in the pooled baseline (Section 3.1), with standard errors clustered at the player level. Stages are assigned by within-season terciles

Figure E.2: The baseline non-white total-card gap by league rank



Notes: Outcome: total cards (yellow + second-yellow + red). Each marker is the non-white total-card gap estimated within one league rung by re-fitting the baseline specification (Equation 1) on that rung's player-matches, expressed in percentage points (the coefficient $\times 100$); bars are 95 percent confidence intervals. Rungs are ordered from the top of the ladder (*Regionalliga Bayern*) down to the bottom divisions (B- and C-Klasse, pooled; Section C.1). Dotted line at zero. Standard errors clustered at the player level.

of the match date, computed over distinct matches; because a division's match-days are calendar-synchronized, these thirds track progress through the schedule. The early stage covers the opening match-days, the middle stage the close of the autumn round and the turn of the year, and the late stage the spring run-in; the near-empty December–February winter break falls inside the middle stage. We drop the 2025/26 season, whose spring round is not yet observed and whose terciles would otherwise misclassify stage. Table E.1 reports the gaps for all three card outcomes.

We also test for heterogeneity by game importance. Because the stakes are highest late in the season, we flag late-season matches in which at least one side is still contesting promotion or relegation, from the standings entering the match, and ask whether the gap shifts with these stakes and whether any shift differs by oversight (ever-observed tiers versus the never-observed

Table E.1: Baseline ethnic card gap by stage of the season

Season stage	Player-match card outcomes		
	Yellow	Red	Total
	(1)	(2)	(3)
Early	0.00540***	0.00038***	0.00668***
	(0.00084)	(0.00007)	(0.00092)
Mid	0.00476***	0.00053***	0.00632***
	(0.00083)	(0.00008)	(0.00092)
Late	0.00629***	0.00042***	0.00777***
	(0.00080)	(0.00007)	(0.00088)
Controls	✓	✓	✓
Club×league×season FE	✓	✓	✓
Opponent FE	✓	✓	✓
Referee FE	✓	✓	✓

Notes: Columns are the three card outcomes: Yellow and Red are binary indicators for receiving any yellow / straight-red card, and Total counts yellow + second-yellow + red. Each row re-fits the baseline specification (Equation 1) on the player-matches falling in that third of the season – early, 7,074,058; middle, 7,123,838; late, 7,074,358 – in a single OLS fit as in the pooled baseline (Table 3), over the eleven complete seasons; 2025/26 is dropped because its spring round is not yet observed. The season thirds are defined in Section E.1. Standard errors clustered at the player level. Signif. Codes: ***: 0.01, **: 0.05, *: 0.1.

bottom divisions). It does not: the non-white × high-stakes × observed interaction is a precise null ($+0.0018$, $p = 0.33$; standard error ≈ 0.0018 on 17.6 million player-matches). Within the ever-observed tiers the stakes effect is null as well ($+0.0001$, $p = 0.94$); within the never-observed divisions the gap is somewhat smaller in high-stakes matches (-0.0017, $p = 0.047$), the opposite of what we would expect if stakes amplified the gap. Reading the standings entering a match requires a division whose table we observe in full, so this test runs on a restricted frame. Beyond the 2020/21 season dropped throughout, we also drop the COVID-curtailed 2019/20 season and the mid-season 2025/26 snapshot, and we keep only matches after the opening match-day in divisions that we observe to within one match-day of a complete double round, that field at least eight teams—which excludes the small promotion and relegation play-off groups—and in whose table both clubs appear. These restrictions leave 17.6 rather than 22.4 million player-matches.

F Within-referee analysis

A potential concern with our main design is that our results might be driven by referee differences and selection: if less biased referees are concentrated in the more closely observed leagues, the pattern would arise from referee selection rather than from oversight. To address this, we estimate specifications in which we hold the referee fixed and ask whether the *same* official's white/non-white card gap is smaller when officiating at higher tiers. This variation arises because referees are at times assigned to officiate matches across (typically consecutive) tiers.

We estimate, on the player-match panel, OLS models of the following form:

$$y_{icjr} = \alpha_r + \alpha_c + \alpha_s + \alpha_\ell + \beta_1 \,\mathrm{NW}_i + \beta_2 \,(\mathrm{NW}_i \times \mathrm{rung}_\ell) + \gamma' X_{icjr} + \varepsilon_{icjr}, \tag{4}$$

where y_{icjr} denotes the total-card count for player i on club c facing opponent j under referee r; NW_i is the non-white indicator; rung_ℓ indexes league ℓ's tier on the oversight ladder (Section C.1), rising by one with each rung of additional oversight; and X_{icjr} is the control vector used throughout (opponent non-white share, home, sub-in). The terms α_r, α_c, α_s, and α_ℓ denote referee, club, season, and league fixed effects; the league fixed effects absorb the rung main effect. Our main quantity of interest is β_2, which identifies the change in the within-referee white/non-white card gap per rung of additional oversight. We keep only referees who officiate in at least two distinct ranks, since they alone give within-referee variation in rung_ℓ. Standard errors are clustered at the match level.

Oversight rises with league rank only over the upper part of the ladder; across the never-observed bottom divisions, league rank still varies, but no match carries a routine observer (Table 1). We leverage this for a placebo test, estimating the specification above on two subsets: first, on the oversight-varying ladder, from the *Regionalliga Bayern* down to the *Kreisliga*, where each rung adds oversight; second, on the never-observed floor (the *Kreisklasse* and the A-, B-, and C-Klasse), where rank moves by rung but oversight does not. Table F.1 reports the results.

Table F.1: Within-referee oversight gradient

	Total cards	
	Oversight-varying ladder	Placebo: floor
	(1)	(2)
Non-white × oversight rung	−0.0033***	−0.0005
	(0.0006)	(0.0004)
Mean of DV	0.148	0.129
Player-matches	3,781,246	16,313,223
Referees	1,652	9,969
Controls	✓	✓
Referee FE	✓	✓
Club FE	✓	✓
Season FE	✓	✓
League FE	✓	✓

Notes: Outcome: total cards (yellow + second-yellow + red). Each column regresses the total-card count on the non-white indicator and its interaction with an oversight rung, plus controls, with referee, club, season, and league fixed effects; standard errors clustered at the match. The oversight rung indexes a league's tier on the eight-rank oversight ladder (Section C.1), rising by one with each rung of more oversight; the same metric is used in both columns. The coefficient on the interaction is the change in the within-referee non-white card gap per rung of additional oversight. Column (1) uses the oversight-varying ladder (*Regionalliga Bayern* down to the *Kreisliga*); column (2) is the placebo on the never-observed bottom divisions (*Kreisklasse* and the A/B/C-Klasse), where league rank still varies but no match is observed. Each column keeps only referees who officiate in at least two distinct ranks of its region, so the interaction is identified within referee. *Referees* counts these across-rank referees. Controls: opponent non-white share, home, sub-in. Signif. Codes: ***: 0.01, **: 0.05, *: 0.1.

G Testing for differential dangerous play

A natural concern regarding the baseline gap (Section 3.1) is that non-white players might foul more dangerously, so that part of their higher card count reflects rougher play rather than referee bias. Injuries are not recorded in our data, but we can approximate them. A straight red card is issued to players, who commit a serious foul, and also records the minute of the foul's occurrence. If such a foul inflicts an injury to an opponent, the opposing side will typically substitute the injured player. We therefore take the substitution of an opposing player in the five minutes following a straight red as a proxy for an injury to that opponent. Because the proxy is defined only once a serious foul has been observed, the analysis is conditional on a dismissal: we restrict the sample to the roughly 49,000 player-matches in which a given player was sent off with a straight red, and regress an indicator for such a substitution on the non-white indicator, the opponent non-white share, and a home indicator. Table G.1 reports the results.

Table G.1: Foul injuriousness, by player ethnicity

	Opponent injury	
	League & season FE	Baseline FE
	(1)	(2)
Non-white	−0.0041	0.0026
	(0.0052)	(0.0107)
Mean of DV	0.196	0.201
Send-off player-matches	48,603	37,590
Controls	✓	✓
League, season FE	✓	✓
Club, opponent, referee FE		✓
Cluster	Referee	Player

Notes: Conditional on the focal player being sent off with a straight red. *Opponent injury* is approximated by an opposing player being substituted off within five minutes of the red (Section G). Controls: opponent non-white share and home (plus a sub-in indicator in column 2, as in the baseline). Column 1 absorbs league and season fixed effects; column 2 the baseline's club×league, opponent, and referee fixed effects (Equation 1), with season fixed effects and the seasons pooled. Signif. Codes: ***: 0.01, **: 0.05, *: 0.1.

H Controlling for match-day audience

Higher league tiers draw larger crowds (Table A.4), so league rank and match-day audience move together. A concern is therefore that our main within-player result might be driven by audience effects rather than external oversight. We test for this by re-estimating the within-player design (Section D.1) with match-day audience added to the controls. We construct a standardized audience measure for each match m,

$$A_m = \frac{\ln(1+\tilde{s}_m) - \mu}{\sigma}, \qquad \tilde{s}_m = \min\left(s_m, c_{\ell(m)}\right), \tag{5}$$

where s_m is the recorded number of spectators at match m, c_ℓ is the 99th percentile of spectators within league ℓ (so that $\tilde{s}_m$ winsorizes implausible entries, as in Table A.4), $\ell(m)$ is the league of match m, and μ and σ are the mean and standard deviation of $\ln(1+\tilde{s}_m)$ in the sample. By construction A_m has mean zero and unit variance, and a one-unit increase corresponds to roughly a doubling of the crowd. We add A_m and its interaction with the non-white indicator to the design's controls. We observe attendance for about 96 percent of player-matches and estimate on that subsample. Table H.1 reports the results. With the audience terms included, the oversight effect is -0.0066 (standard error 0.0028), compared with -0.0060 on the full sample; larger crowds come with more cards overall (0.0106), but the audience's interaction with the non-white indicator is small and insignificant (0.0014, standard error 0.0009).

Table H.1: Within-player design: robustness to match-day audience

	Total cards
NW × Oversight	-0.0066^{**}
	(0.0028)
Audience	0.0106^{***}
	(0.0003)
NW × Audience	0.0014
	(0.0009)
Mean of DV	0.149
Player-matches	5,225,702

Notes: Total cards. The within-player per-step oversight effect (the specification of Table 4, Column 1), adding the match-day audience measure A_m (Eq. 5) and its interaction with the non-white indicator to the controls; the "Audience" and "NW × Audience" rows report these two terms. Player and season fixed effects; standard errors clustered at the team level. See SI Section H. Signif. Codes: ***: 0.01, **: 0.05, *: 0.1.

I Controlling for referee experience

League rank and referee experience increase together, as referees in higher tiers have typically officiated for a longer time, so it could be that our main within-player result might actually reflect what referees learn over their careers. To adress this we re-estimate the within-player design (Section D.1), adding referee experience to the controls. We use tenure as a proxy measurement for experience, defining it as the number of earlier seasons in which a referee officiated at least one match in our panel. We do not know for referees already active in 2013/14, how long they had been officiating before our panel begins. Therefore, we use 2013/14 only to initiate the tenure count and exclude its matches from estimation. This ensure that every estimated player-match has at least one season of observed history. A stricter variant drops this cohort altogether, leaving only referees whose full officiating history falls inside the panel, retaining less than a third of all observations.

Table I.1 reports the results. Excluding the first season does not change the oversight effect (-0.0060, as at baseline; Column 2). Introducing referee experience and its interaction with the non-white variable essentially does not move the estimate: it shifts from -0.0060 to -0.0061, which is about four percent its standard error (Column 3). The experience terms themselves are considerable: more experienced referees issue fewer cards overall, and the white/non-white gap is smaller by 0.0007 cards per season of tenure ($p = 0.02$). The difference in cards across the interquartile range of referee tenure (one to six prior seasons) is 0.0036, which is approximately half the baseline gap of 0.0069, and three-fifths of one oversight step. This is the effect that this model is built to control for, and as such leaves the oversight estimate unchanged. We do not interpret the experience terms as a learning curve within referees. With season but no referee fixed effects, tenure mainly varies across referee entry cohorts. Thus, the gradient mixes learning with who enters, remains, and is assigned within the system. Furthermore, the tenure measure itself is noisy, which attenuates both terms. The strict variant is uninformative, not discrepant: it retains less than a third of the player–matches, and its confidence interval, -0.0090 to $+0.0099$, comfortably contains the main estimate (Column 4). Therefore, the oversight effect is thus not an

artifact of referee experience.

Table I.1: Within-player design: robustness to referee experience

	Baseline	Excl. first season	+ Ref. experience	Strict cohort
	(1)	(2)	(3)	(4)
NW × Oversight	−0.0060**	−0.0060**	−0.0061**	0.0004
	(0.0028)	(0.0029)	(0.0029)	(0.0048)
NW × Ref. experience	–	–	−0.0007**	−0.0019***
			(0.0003)	(0.0005)
Ref. experience	–	–	−0.0016***	0.0002
			(0.0001)	(0.0002)
Excludes 2013/14	No	Yes	Yes	Yes
Excludes 2013/14 ref. cohort	No	No	No	Yes
Player-matches	5,434,230	4,905,690	4,905,690	1,563,981
Player moves	151,368	138,116	138,116	122,248

Notes: Total cards. The within-player per-step oversight effect (the specification of Table 4, Column 1). Column 2 drops the 2013/14 season, which seeds the experience measure. Column 3 adds referee experience (prior active seasons) and its interaction with the non-white indicator. Column 4 additionally drops referees first observed in 2013/14, so experience is fully observed. Player and season fixed effects; standard errors clustered at the team level. See SI Section I. Signif. Codes: ***: 0.01, **: 0.05, *: 0.1.

J Placebo outcome outside the referee's control

To check that the within-player design (Section D.1) isolates referee behavior rather than differences in who plays, or how players are used, across tiers, we re-estimate it on an outcome a referee cannot set—whether the player was substituted off.[31] We drop the sub-in indicator from the control vector, since it is mechanically related to whether a player is substituted off, and otherwise change only the dependent variable. A contrast that instead reflected such differences would surface on this outcome too.

We re-fit the signed per-step oversight estimator (Section D.1) on this placebo outcome. Table J.1 reports the estimates.

Table J.1: Within-player triple-difference: placebo outcome

	Placebo outcome
	Subbed off
	(1)
NW × Oversight	0.0014
	(0.0031)
NW × Pure rank	−0.0018
	(0.0019)
Mean of DV	0.178
Player-matches	5,434,230
Season FE	✓
Player FE	✓

Notes: The per-step oversight estimator of Section D.1, net of one league-step of rank, applied to an outcome the referee does not set: whether the player was substituted off. The table reports the net oversight effect (NW × Oversight) and the netted rank slope (NW × Pure rank). Within player; no referee fixed effect, by design. Standard errors clustered at the team level. Signif. Codes: ***: 0.01, **: 0.05, *: 0.1.

[31] Substitution is not a perfect placebo outcome: a yellow card can prompt a coach to strategically withdraw the player in order to avoid a second yellow and thereby their dismissal. Non-white players are cautioned more often, and therefore may get withdrawn more frequently. This reaches few player-matches, and, if anything, pushes the placebo away from zero in the direction of the card gap, so the null we report bounds it as small.

K Full estimates for the main-text tables

Table K.1: Baseline ethnic card gap, full estimates

	Player-match card outcomes		
	Yellow	Red	Total
	(1)	(2)	(3)
Non-white	0.00549***	0.00041***	0.00690***
	(0.00071)	(0.00004)	(0.00079)
Opponent non-white share	−0.00331***	−0.00048***	−0.00427***
	(0.00095)	(0.00014)	(0.00103)
Home	−0.00714***	−0.00036***	−0.00860***
	(0.00014)	(0.00002)	(0.00015)
Sub-in	−0.05044***	−0.00075***	−0.05464***
	(0.00025)	(0.00002)	(0.00027)
Mean of DV	0.124	0.002	0.133
Player-matches	22,411,568	22,411,568	22,411,568
Clusters (players)	377,823	377,823	377,823
Club×league×season FE	✓	✓	✓
Opponent FE	✓	✓	✓
Referee FE	✓	✓	✓

Notes: Full estimates of the specifications in Table 3 (Equation 1), estimated by OLS on the full pooled sample (2013/14–2025/26; 2020/21 dropped). Yellow and Red are binary indicators for receiving any yellow / straight-red card; Total counts yellow + second-yellow + red. Opponent non-white share is the non-white share among the opposing team's classified starters (Section A.1); Home equals one for players on the home team; Sub-in equals one if the player entered as a substitute. Fixed-effect estimates are not reported. Standard errors clustered at the player level. Signif. Codes: ***: 0.01, **: 0.05, *: 0.1.

Table K.2: Within-player triple-difference, full estimates

	Total cards			
	Player FE		Window × player FE	
	Main estimate (1)	Pre-trend placebo (2)	Main estimate (3)	Pre-trend placebo (4)
Post × NW × Oversight increment	−0.0060**	−0.0004	−0.0063**	−0.0006
	(0.0028)	(0.0031)	(0.0028)	(0.0030)
Post × NW × Rank step	0.0007	0.0026	0.0006	0.0024
	(0.0017)	(0.0020)	(0.0018)	(0.0020)
Post × NW	0.0017	−0.0009	0.0020	−0.0006
	(0.0012)	(0.0014)	(0.0013)	(0.0014)
Post × Oversight increment	−0.0090***	0.0017*	−0.0091***	0.0017*
	(0.0009)	(0.0009)	(0.0009)	(0.0009)
Post × Rank step	0.0271***	−0.0197***	0.0272***	−0.0199***
	(0.0006)	(0.0006)	(0.0006)	(0.0006)
NW × Oversight increment	0.0042*	0.0026		
	(0.0022)	(0.0025)		
NW × Rank step	0.0005	−0.0007		
	(0.0014)	(0.0016)		
Oversight increment	0.0042***	0.0022***		
	(0.0007)	(0.0007)		
Rank step	−0.0193***	−0.0007		
	(0.0004)	(0.0005)		
Post	0.0060***	0.0015***		
	(0.0004)	(0.0004)		
Opponent non-white share	−0.0135***	−0.0142***	−0.0131***	−0.0134***
	(0.0014)	(0.0015)	(0.0014)	(0.0015)
Home	−0.0077***	−0.0084***	−0.0076***	−0.0084***
	(0.0004)	(0.0004)	(0.0004)	(0.0004)
Sub-in	−0.0449***	−0.0473***	−0.0407***	−0.0439***
	(0.0006)	(0.0006)	(0.0006)	(0.0006)
Mean of DV	0.147	0.146	0.147	0.146
Player-matches	5,434,230	4,457,065	5,434,230	4,457,065
Clusters (teams)	8,984	9,882	8,984	9,882
Season FE	✓	✓	✓	✓
Player FE	✓	✓		
Window × player FE			✓	✓

Notes: Full estimates of the specifications in Table 4. The Post × NW × Oversight increment and Post × NW × Rank step rows are the NW × Oversight and NW × Pure rank coefficients of that table. Rank step is coded +1 for a one-league promotion and −1 for a relegation, for every move; Oversight increment is coded the same way for moves that cross an oversight boundary and is zero otherwise (Section D.1). Post marks the later season of the bracketing pair: $t+1$ in Columns (1) and (3), t in the pre-trend placebo Columns (2) and (4). Controls as in Table K.1. The fixed effects absorb the NW main effect. Post, Rank step, Oversight increment, and the interactions of the latter two with NW are identified only from players observed in more than one move window in Columns (1) and (2); in Columns (3) and (4), the window × player and season fixed effects absorb them (empty cells). In Columns (2) and (4), Post × Rank step reflects that card rates fall in the season before a promotion and rise before a relegation; the placebo concerns the rows interacted with NW. Fixed-effect estimates are not reported. Standard errors clustered at the team level. Signif. Codes: ***: 0.01, **: 0.05, *: 0.1.