SCIENTIFIC JOURNAL OF SPORT AND PERFORMANCE

ORIGINAL ARTICLE

The captain as a performance signal: Associations between captain ratings, teammates' ratings, and match outcomes at the 2026 FIFA World Cup

Fernando Adrian Mora-Davila . *Department of Humanistic Studies. Technological Institute of Monterrey, Guadalajara Campus. Zapopan, Jalisco, Mexico.*

Natalia Rocha. *Department of Humanistic Studies. Technological Institute of Monterrey, Guadalajara Campus. Zapopan, Jalisco, Mexico.*

ABSTRACT

This study examined whether the match performance of formally appointed captains, understood as a leadership-relevant performance signal rather than a measure of leadership, is associated with teammates' collective performance and match outcomes at the 2026 FIFA World Cup. Sofascore ratings were collected for every player in all 104 matches, yielding 208 captain–team–match observations from 66 starting captains of 48 national teams; teammates' performance was the minutes-weighted mean rating excluding the captain. Analyses comprised linear mixed models with crossed random effects, a paired match-level ordinal regression adjusting for teammates' performance, FIFA ranking, and stage, and permutation and quality-matched counterfactual analyses with non-captain starters. Captain ratings correlated with teammates' ratings (r = .39), but no more than those of randomly selected starters. What distinguished captains was their coupling with match outcome: the difference between the two captains' ratings predicted the result (OR = 5.78 per rating point), and the captains' rating–outcome correlation (ρ = .54) exceeded that of random, positionally matched, and quality-matched teammates. The coupling was asymmetric: captains stood out from their teammates in wins (+1.18 teammate *SD*s) but not in losses (+0.31), when captain and team declined together. Because ratings are assigned after the match and are sensitive to the result, the direction of influence cannot be established. The findings portray the captain as one consequential node within a distributed leadership network and suggest that the credibility of formal leadership rests partly on the leader's own visible performance.

Keywords: performance analysis; athlete leadership; association football; player ratings; formal leadership roles.

Cite this article as:
Mora-Davila, F. A., & Rocha, N. (2027). The captain as a performance signal: Associations between captain ratings, teammates' ratings, and match outcomes at the 2026 FIFA World Cup. *Scientific Journal of Sport and Performance*, *6*(1), 118-134. https://doi.org/10.55860/YLPR9523

Corresponding author. *Department of Humanistic Studies. Tecnologico de Monterrey, Campus Guadalajara. Av. Gral. Ramon Corona 2514, Colonia Nuevo Mexico, 45201 Zapopan, Jalisco, Mexico.*
E-mail: fmora@tec.mx
Submitted for publication September 11, 2026.
Accepted for publication November 18, 2026.
Published November 25, 2026.
Scientific Journal of Sport and Performance. ISSN 2794-0586.
©Asociación Española de Análisis del Rendimiento Deportivo. Alicante. Spain.
doi: https://doi.org/10.55860/YLPR9523

© 2027 ARD Asociación Española

INTRODUCTION

Football is a field-invasion sport in which success depends on continuous coordination and mutual adaptation among interdependent players. Teams behave as complex adaptive systems: collective patterns emerge from local interactions among teammates, opponents, and changing match conditions rather than from centralised control, so leadership can guide collective behaviour without determining it (Passos et al., 2013).

Classical sport-leadership research located leadership primarily in the coach, assessing coaching behaviours through athletes' perceptions (Chelladurai & Saleh, 1980). The later turn to athlete leadership defined the athlete leader as a player in a formal or informal role who influences teammates towards a common goal (Loughead et al., 2006), and linked such leadership to cohesion, confidence, identification, motivational climate, and team functioning (Cotterill & Fransen, 2016). Within this literature, the captain occupies an ambiguous position: the armband confers formal recognition and on-field visibility, while the coach retains authority over major tactical decisions (Cotterill et al., 2022). However, formal recognition does not guarantee effective influence. Only 1% of 4,451 players and coaches regarded the captain as the best leader across task, motivational, social, and external roles, and 44% did not see the captain as the principal leader in any role (Fransen et al., 2014); leadership also tends to become more distributed over a season (Mertens et al., 2021), and teams whose passing is less centralised around a few players score more goals (Grund, 2012).

The first proposition of this study follows from this evidence: formal captaincy and effective leadership are related but distinct constructs, and the captain is best understood not as the team's automatic leader but as one consequential node within a distributed leadership network.

Rejecting the heroic captain does not make the role irrelevant. The armband makes one player's decisions, effort, and execution especially salient, raising the question of whether the captain sustains formal status through observable contribution. Of the four athlete-leadership roles distinguished by Fransen et al. (2014), task leadership, which keeps teammates focused on collective objectives and guides in-game adjustments, is the one performed most directly during competition. Task-oriented peer leadership shows modest positive associations with cohesion and collective efficacy (Piasecki & Filho, 2025), and higher perceived athlete-leadership quality accompanies stronger team confidence and objective performance (Fransen et al., 2017).

Two mechanisms suggest why visible performance may matter. In transformational terms, leading by example requires embodying expected standards, and perceived competence is among the attributes that distinguish effective peer leaders (Price & Weiss, 2011): strong performance can lend credibility to the armband, whereas weak performance can open a gap between symbolic authority and enacted contribution. In social-identity terms, leaders are influential as far as they represent a shared sense of “we”; captains contribute distinctly to team identification, although informal leaders carry the strongest effects on performance (Fransen et al., 2020). Neither mechanism is measured here; both are plausible explanations, rather than tested hypotheses.

The second proposition is therefore that the captain's match performance constitutes a leadership-relevant performance signal: an observable indicator of task contribution that may sustain the credibility of the formal role, but not a measure of leadership, which also involves communication, motivation, emotional regulation, and teammate acceptance. Because a well-functioning team also creates the conditions for its captain to perform, the signal is expected to be reciprocally and loosely coupled with collective performance rather than to cause it (Orton & Weick, 1990). It is also distinct from leader emergence, being perceived as leader-like,

and from leadership effectiveness, influencing and guiding the unit towards its goals (Judge et al., 2002): the captain is appointed, and a post-match rating records task contribution, not influence on teammates.

Athlete-leadership research has relied on questionnaires, interviews, peer nominations, and social-network analysis, which capture perceptions of leaders but are rarely feasible in elite international competition. Event-based player ratings offer a complementary, observational window onto task contribution: they relate systematically to observable match actions, although platforms weight offensive and defensive contributions differently (Ball et al., 2025). Four limits are central to inference. First, the underlying event data are coded by human operators, and high inter-operator reliability has been documented for only one provider (Liu et al., 2013). Second, ratings have no objective ground truth: action-valuation models credit each action by its effect on the team's chances of scoring or conceding (Decroos et al., 2019), consumer algorithms are proprietary (García-Fernández & Quian, 2026), event statistics capture only part of a player's contribution and may miss contributions such as motivating teammates (Gelade & Hvattum, 2020), and expert human ratings are no gold standard (Bergkamp et al., 2022). Third, ratings are assigned after the match and are sensitive to the result (García-Fernández & Quian, 2026; Kausel et al., 2019). Fourth, because fielded players' inputs are complements (Franck & Nüesch, 2010), an individual–team association is expected for any player and such an association cannot by itself reveal which influences which (Manski, 1993).

The third proposition is that the captain's performance signal can nonetheless be examined observationally, provided the design separates the captain from the collective measure, respects the repeated and paired structure of match data, and tests what is specific to the captain against comparable non-captains. The 2026 FIFA World Cup offers a suitable setting: 48 national teams and 104 matches under increasing competitive pressure, with repeated observations of captains across group and knockout stages. The closest precedent, covering seven international tournaments, found that outcomes aligned with the team's mean rating and relative strength, whereas a standout individual performance had no clear independent association with outcomes (Ball et al., 2026). Whether the formally designated leader differs from other players in this respect has not been examined.

Accordingly, this study examines the relationship between the match performance of formally appointed captains, the collective performance of their teammates, and match outcomes. The research question is: To what extent does the captain's match performance, understood as a leadership-relevant performance signal, relate to teammates' collective performance and to competitive outcomes in an elite international football tournament?

Two hypotheses were formulated. H1: Higher match performance by the formally appointed captain will be associated with higher collective performance of his teammates. H2: Higher match performance by the formally appointed captain will be associated with a greater likelihood of a positive match outcome. In addition, an exploratory question asks whether these associations are specific to the captain or are shared by comparable non-captain starters.

MATERIAL AND METHODS

Participants

This study employed an observational, archival design based on match-level data from the 2026 FIFA World Cup, collected retrospectively from official tournament records and football-analytics platforms. The unit of analysis was the captain–team–match observation: the performance of the player who started a match wearing the captain's armband, and the corresponding performance and result of his team.

 © 2027 ARD Asociación Española

The analytical sample comprised 208 captain–team–match observations from all 104 matches, corresponding to 66 players who started at least one match as captain for one of the 48 national teams. All captains were men (age *M* = 32.0 years, *SD* = 3.9, range = 24–41; height *M* = 182.0 cm, *SD* = 6.0): 24 defenders, 21 midfielders, 18 forwards, and 3 goalkeepers from all six confederations (UEFA = 21, AFC = 15, CAF = 12, CONCACAF = 9, CONMEBOL = 8, OFC = 1). Captains started between one and eight matches (*M* = 3.2, *SD* = 1.8), and 51 started at least two. By observation, positions were 70 defender, 66 midfielder, 64 forward, and 8 goalkeeper. Age, height, and position were taken from the official FIFA squad lists.

Measures

Captain and team performance. Ratings were obtained from three widely used platforms that generate systematic, event-based ratings from recorded match actions rather than journalistic evaluations: Sofascore (sofascore.com), FotMob (fotmob.com), and WhoScored (whoscored.com). Sofascore processes hundreds of data points weighted by the context and impact of attacking, passing, dribbling, defensive, and goalkeeping actions; FotMob uses more than 300 data points from Opta event data; and WhoScored computes ratings from more than 200 Opta-based statistics weighted by their influence on the game.

Sofascore was the primary source because it was the only provider for which ratings could be collected for every player in every match, as the present analyses require; FotMob, WhoScored, and their equal-weighted composite were used in sensitivity analyses. The three ratings relate systematically to match actions but weight offensive and defensive contributions differently and are not interchangeable (Ball et al., 2025); because FotMob and WhoScored share Opta event data, their convergence does not imply independence. Inter-provider reliability for captain ratings was high (ICC(2,1) = .85; ICC(2,3) = .94; pairwise *r* = .83–.89), and Sofascore yielded the most conservative estimates of the three. No reliability evidence comparable to that available for Opta data (Liu et al., 2013) was identified for Sofascore, and all three algorithms are proprietary (García-Fernández & Quian, 2026). Ratings are therefore treated as an observable proxy of task contribution, not as validated measures of leadership or of true performance. Because the statistics relevant to players' contribution differ by position (Gelade & Hvattum, 2020), positional comparisons are exploratory. Plus–minus ratings were not used: over a few matches they are not meaningful, and for players who play almost all minutes they converge on the team's rating (Kharrat et al., 2020).

CaptainPerformance, the principal independent variable, was the Sofascore rating of the starting captain in each match, interpreted throughout as a leadership-relevant performance signal. No leadership behaviours, teammate perceptions, cohesion, collective efficacy, or social identity processes were measured; any leadership reading of this variable is interpretive.

TeammatesPerformance, the collective measure and dependent variable for H1, was the minutes-weighted mean rating of the team's rated players excluding the captain, computed from a purpose-built dataset of individual Sofascore ratings for every player in all 104 matches (3,288 player–match observations). Excluding the captain removes, by construction, any mechanical overlap between the individual and collective measures. Seventy player–match observations (2.1%) had no published rating owing to very short appearances and were excluded.

CaptainTeamGap was a standardised within-match salience measure: the captain's rating minus his teammates' mean rating in that match, divided by the standard deviation of those teammates' ratings, expressing how many standard deviations the captain stood out from his own team. Standardising within the team–match addresses the different scales of individual and aggregate ratings (*SD* = 0.84 vs. 0.34); the raw difference is used descriptively only.

MatchOutcome (loss, draw, win) was the dependent variable for H2. Draws could occur only in the group stage (20 of 72 matches). The 32 knockout matches were coded by their result after extra time, and the four decided by penalty shoot-outs as wins for the advancing team and losses for the eliminated team, because the ratings cover the full match and advancement is the competitively decisive outcome. A sensitivity analysis recoded these four matches as draws.

MinutesPlayed, recorded from entry and exit minutes, was the weighting variable for *TeammatesPerformance* (captains: *M* = 86.4 minutes). Captain-level summaries averaged ratings per captain with equal weight per match, restricted to captains with at least two starts (*n* = 51).

Procedures

Official FIFA squad lists were used to validate player names, team membership, and positions, and official line-ups and match reports to identify the captain at the start of each match. Only starting captains were included; players who received the armband during a match were not treated as the focal captain, because they had not entered the match with the same formally assigned responsibility.

Ratings were collected manually by the two authors from Sofascore match pages between 6 October and 2 November 2026, when published ratings were stable, logging the match URL, capture date, and collector. Each team–match was verified against the official line-up (eleven starters; substitutes with entry and exit minutes).

Analysis

Analyses were conducted in Python (pandas, numpy, scipy.stats, statsmodels). Because the observations are cross-classified—66 captains within 48 teams observed repeatedly, with two paired observations per match—no inferential analysis treats them as independent. H1 was tested with a linear mixed model regressing *CaptainPerformance* on TeammatesPerformance with crossed random intercepts for team and match, as in football studies cross-classified by team and game or opponent (Grund, 2012; Kausel et al., 2019); a sensitivity model grouped by captain instead of team, the two being nearly confounded. H2 was tested with a paired match-level design (*n* = 104 matches): the difference between the two captains' ratings, the difference in TeammatesPerformance, the difference in FIFA ranking (11 June 2026), and a knockout indicator predicted the result (loss < draw < win) in a proportional-odds ordinal regression. Differencing within matches removes within-match dependence and adjusts for the specific opponent. The proportional-odds assumption was checked with separate binary logits at the two cut points.

A further sensitivity model re-estimated the ordinal regression on all 208 observations with cluster bootstrap standard errors by team (1,000 resamples). Outcome-group comparisons used Welch's ANOVA (Levene's *p* = .02), Kruskal–Wallis, and Bonferroni-corrected Dunn tests. Effect sizes are reported as η^2, Cohen's *d*, and odds ratios.

To test whether associations are specific to captains, two counterfactual analyses used the full-squad data: (a) a permutation analysis drawing, in each of 5,000 iterations, one random non-captain starter per team–match (overall and within the captain's positional line) and recomputing the focal statistics; and (b) a matched-player analysis pairing each captain with the starter closest in baseline quality (minutes-weighted mean rating in his other matches), with match-clustered bootstrap intervals (2,000 resamples). Positional analyses are exploratory; goalkeeper captains (*n* = 8) are described only.

 © 2027 ARD Asociación Española

Use of AI tools

In line with the guidance of the Committee on Publication Ethics on authorship and AI tools, the authors disclose that Claude (Anthropic; Claude Opus and Claude Sonnet models, accessed through claude.ai between October and November 2026) was used during the revision of this manuscript. The tool (a) cross-checked the manually collected Sofascore ratings against the captain-level database and flagged discrepancies, which the authors resolved against the dated capture; (b) wrote and executed Python code for the analyses described in this section (mixed models, paired ordinal regressions, bootstrap and permutation procedures) and produced Figures 1–2 from the authors' dataset. The authors collected the data, decided the analytic strategy, selected the literature, the cases reported, and verified the code outputs. The tool was also used for copyediting of language, readability, spelling, and grammar.

RESULTS

Descriptive statistics

Table 1 presents descriptive statistics for the principal variables.

Table 1. Descriptive statistics for the principal study variables.

Variable	*M*	*SD*	Min	Max
CaptainPerformance	7.12	0.84	5.2	10
TeammatesPerformance	6.85	0.34	6.01	7.78
Captain–teammates difference (raw)	0.27	0.78	−1.78	3.41
Standardised captain–teammates gap	0.65	1.65	−2.70	10.57

Note. *N* = 208 captain–team–match observations. All ratings are Sofascore match ratings (0–10 scale). CaptainPerformance is the rating of the starting captain. TeammatesPerformance is the minutes-weighted mean rating of all rated teammates, excluding the captain. The standardised gap is (CaptainPerformance − TeammatesPerformance) / *SD* of teammates' ratings within the same team–match.

Captains received higher ratings than their teammates' average in 58.2% of observations, with a mean raw difference of +0.27 and a mean standardised gap of +0.65. This pattern suggests that captains tend to be among their team's stronger performers, although not necessarily the strongest. Consistent with this, captains ranked, on average, at the 64th percentile among their team's starters in terms of baseline quality. Captain ratings were more dispersed than the teammates' aggregate (*SD* = 0.84 vs. 0.34), as aggregation across players stabilises collective performance.

H1: Captain performance and team performance

H1 was supported; CaptainPerformance correlated with TeammatesPerformance, $r(206) = .39$, 95% CI [.27, .50], $p < .001$ (Spearman $\rho = .45$). In the mixed model with crossed random intercepts for team and match, $b = 0.94$ ($SE = 0.15$, $p < .001$; ICC: team = .18, match = .15); the sensitivity model grouping by captain instead of team gave $b = 0.89$. When the captain was included in the team measure, the correlation rose to $r = .55$, showing that part of the association with team-level ratings reflects the captain's own contribution to them.

Controlling for FIFA ranking reduced but did not eliminate the association (partial $r = .26$, $p < .001$). Among captains with at least two starts ($n = 51$), mean captain and teammates' ratings correlated at $r = .47$ ($p < .001$), and captain rating correlated with the proportion of matches won at $r = .63$ ($p < .001$).

As the counterfactual analysis shows, this individual–collective coupling is not specific to captains; the captain-specific findings concern the association with match outcome and the standardised gap.

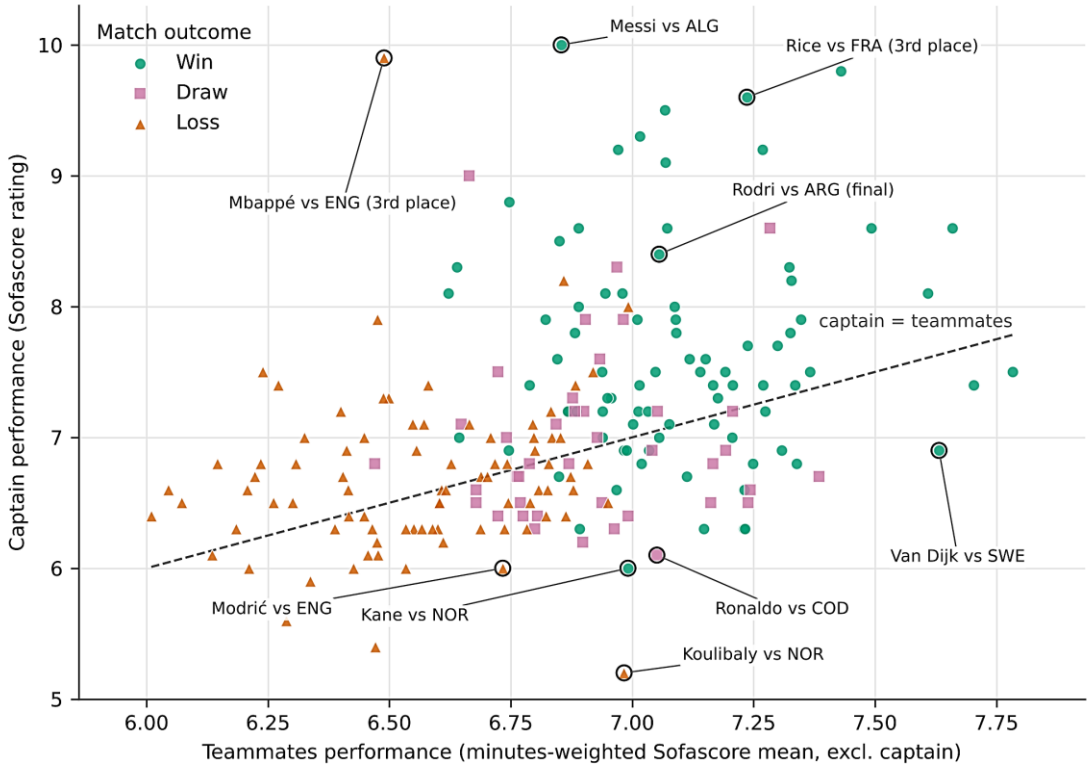


Figure 1. Captain Sofascore rating against teammates' mean rating in each captain–team–match observation, by match outcome (n = 208).

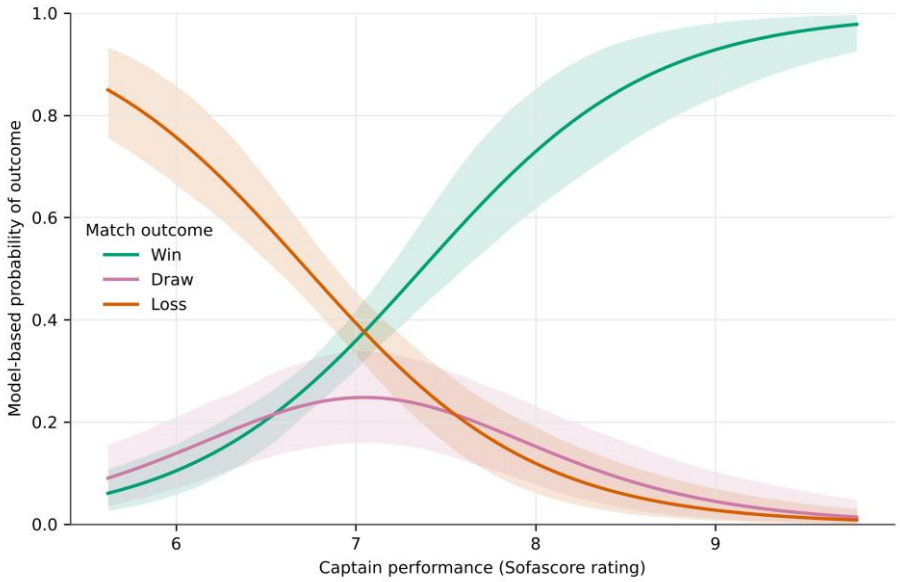
Figure 1 plots each captain's rating against his teammates' mean rating. Because both axes use the individual-rating scale, the vertical distance from the equality line is the raw captain–teammates difference. Wins concentrate above the line and at high captain ratings, whereas in losses captain and teammates decline together, leaving the average distance near zero. Nine labelled observations illustrate the range of configurations. The two largest captain advantages occurred in opposite outcomes: Messi rated 10.0 against his teammates' 6.85 in a win over Algeria, and Mbappé 9.9 against 6.49 in the third-place defeat by England, a match in which England's captain also stood out (Rice, 9.6 vs. 7.24). The result therefore followed the difference between the two captains rather than either captain's level alone. The champion captain also stood out in the final (Rodri, 8.4 vs. 7.06). Conversely, England and the Netherlands won with their captain rated below his teammates (Kane, 6.0 vs. 6.99; Van Dijk, 6.9 vs. 7.63), Portugal drew with Ronaldo rated below his (6.1 vs. 7.05), and Croatia lost with captain and teammates both rated low (Modrić, 6.0 vs. 6.73). The lowest captain rating of the tournament, which was also the largest negative gap, accompanied a defeat (Koulibaly, 5.2 vs. 6.98).

H2: Captain performance and match outcome

H2 was supported; in the paired match-level model (n = 104), the difference between the two captains' ratings predicted the result after controlling for the difference in TeammatesPerformance, the FIFA ranking difference, and stage, b = 1.75 (SE = 0.44, p < .001), OR = 5.78 per rating point, 95% CI [2.43, 13.75]. The teammates' difference was the dominant predictor, and the ranking difference was not significant once performance terms entered; without the teammates' control, the captain OR was 4.57. In the observation-level sensitivity model (n = 208) with cluster bootstrap by team, each SD of captain rating multiplied the odds of a better outcome by 3.76 (b = 1.32, bootstrap 95% CI [0.62, 1.78]). Recoding penalty-decided matches as

© 2027 ARD Asociación Española

draws left the estimate unchanged (b = 1.83), and binary logits at the two cut points yielded similar coefficients (1.85 and 1.93), supporting the proportional-odds assumption. Captain ratings differed across outcomes (wins 7.62, draws 6.96, losses 6.69; Welch $F(2, 107.5) = 32.30$, $p < .001$, $\eta^2 = .26$; $d = 1.24$ for wins vs. losses; Kruskal–Wallis $H = 61.9$, $p < .001$), with Dunn tests locating the difference in wins (draws vs. losses, $p = .17$). Because ratings incorporate outcome-defining events, these estimates quantify coupling, not an independent contribution (Figure 2).



Note. Model-based predicted probabilities of each match outcome by captain Sofascore rating (observation-level ordinal regression, n = 208), with 95% confidence bands. The curves describe in-sample associations between post-match ratings and results, not predictions for future matches; the crossing point of the win and loss curves is model-dependent and should not be interpreted as a performance threshold.

Figure 2. Predicted probability of loss, draw, and win by captain Sofascore rating.

Table 2. Captain performance and standardised captain–teammates gap by match outcome.

Outcome	*n*	Captain *M* (*SD*)	Std. gap *M* (*SD*)	Captain above teammates (%)
Win	84	7.62 (0.85)	1.18 (2.02)	72.6
Draw	40	6.96 (0.65)	0.22 (1.22)	42.5
Loss	84	6.69 (0.63)	0.31 (1.24)	51.2

Note. N = 208 starting-captain observations. M = mean; SD = standard deviation. Outcomes refer to the result after regular and extra time; draws occur only in the group stage. Matches decided on penalties are coded as won or lost (a sensitivity analysis recoding them as draws is reported in the text). "Captain above teammates" is the percentage of observations in which the captain's rating exceeded the minutes-weighted teammates mean. Descriptive statistics; inferential tests account for the paired match structure (see Results).

The standardised captain–teammates gap was clearly highest in wins (Table 2), but it did not follow the win > draw > loss ordering of absolute ratings: +1.18 teammate standard deviations in wins, +0.22 in draws, and +0.31 in losses, $F(2, 205) = 8.02$, $p < .001$; captains exceeded their teammates' mean rating in 73% of wins, 42% of draws, and 51% of losses.

This asymmetry is informative: winning was characterised by captain overperformance, but losing was not characterised by captain underperformance. In defeats, captain and teammates declined together, leaving

the gap near zero, so losses appear predominantly collective. Rather than the captain dragging the team down, captain and teammates' ratings fell jointly. This pattern is consistent with the reciprocal, system-level conceptualisation of the captain–team relationship advanced in the introduction. In the third-place match, for example, Kylian Mbappé recorded a 9.9 rating, the second-highest captain rating in the dataset, in a losing effort, underscoring that elite individual performance does not guarantee a positive team outcome within an interdependent competitive system.

Table 3. Match outcomes by quartile of (a) the standardised captain–teammates gap and (b) absolute captain performance.

Quartile	Range	*n*	Win (%)	Draw (%)	Loss (%)
Panel (a): Standardised captain–teammates gap					
Q1 (lowest)	−2.70 to −0.33	52	21.2	28.8	50
Q2	−0.32 to 0.42	52	36.5	21.2	42.3
Q3	0.42 to 1.27	52	46.2	13.5	40.4
Q4 (highest)	1.28 to 10.57	52	57.7	13.5	28.8
Panel (b): Absolute captain performance					
Q1 (lowest)	5.20 to 6.50	55	9.1	23.6	67.3
Q2	6.60 to 7.00	60	28.3	21.7	50
Q3	7.10 to 7.50	47	55.3	17	27.7
Q4 (highest)	7.60 to 10.00	46	78.3	13	8.7

Note. *N* = 208 starting-captain observations; percentages are computed within each quartile. Quartile sizes in panel (b) are unequal because Sofascore ratings are reported to one decimal and tied values were kept in the same quartile; ranges in panel (a) overlap at 0.42 because of rounding. The steeper gradient in panel (b) is expected, because ratings incorporate the decisive actions that accompany winning; quartiles are descriptive and are not used for inference.

Win rates increased monotonically across quartiles of both the standardised gap (21.2% to 57.7%) and absolute captain performance (9.1% to 78.3%; Table 3). However, the steeper gradient for absolute performance does not imply a stronger causal effect, as post-match ratings incorporate events that determine match outcomes. Rather, it reflects how strongly these ratings distinguish between results.

Positional analysis

All positional analyses are exploratory (Table 4). The captain's rating was associated with outcome within each field position (OR per *SD*: defenders 5.87, midfielders 3.48, forwards 3.43), and the rating × position interaction was not significant (LR = 1.72, *df* = 2, *p* = .423). With 64–70 observations per position, however, this test has limited power, and a non-significant interaction is not evidence of positional invariance. Predicted win probabilities at a rating of 7.5 overlapped fully (defenders .65, midfielders .53, forwards .53).

The standardised gap related to outcome most clearly among forward captains (ρ = .34, *p* = .006), marginally among defenders (ρ = .24, *p* = .050), and not among midfielders (ρ = .12, *p* = .347). A plausible reading is measurement-related: event-based ratings register decisive offensive actions sharply, so a forward captain's overperformance is captured in the events that most directly produce victories, whereas midfield contributions are absorbed more diffusely into collective performance.

Counterfactual benchmark: captains versus comparable non-captains

A central question is whether these associations are specific to the captain. In the permutation analysis, the individual–collective coupling was not: the captain–teammates correlation (*r* = .39) sat near the centre of the permutation distribution (mean .37, 95% range [.26, .47], *p* = .36). The outcome associations, by contrast,

 © 2027 ARD Asociación Española

were specific: the captain's rating–outcome correlation (ρ = .54) exceeded virtually all permuted values (means .38 overall and .41 within position; $p < .001$), as did the gap–outcome correlation (ρ = .23 vs. permutation means of .02 and .01; $p < .001$).

Table 4. Association between captain rating and match outcome by position.

Position	*n*	OR per 1 SD [95% CI]	Interaction vs. DF, β [95% CI]	P(win) at rating 7.5 [95% CI]	Std. gap → outcome, ρ (*p*)
Defenders	70	5.87 [2.56, 13.45]	ref.	.65 [.47, .85]	.24 (.050)
Midfielders	66	3.48 [1.73, 6.99]	−0.45 [−1.52, 0.61]	.53 [.40, .71]	.12 (.347)
Forwards	64	3.43 [1.82, 6.46]	−0.66 [−1.66, 0.34]	.53 [.39, .82]	.34 (.006)
Rating × position (LR test)	200	—	LR = 1.72, *df* = 2, *p* = .423	—	—

Note. Exploratory proportional-odds models, field positions only. OR = odds ratio for a more favourable outcome (loss < draw < win) per 1 *SD* of the captain's Sofascore rating, from separate models by position; CI = confidence interval; DF = defenders; LR = likelihood-ratio test. Interaction coefficients are from a pooled model with defenders as reference. Predicted probabilities are from the pooled model, with 95% intervals from 1,000 match-clustered bootstrap replicates. ρ = Spearman correlation between the standardised captain–teammates gap and match outcome. Goalkeeper captains (*n* = 8) are described in the text only. All analyses are exploratory and uncorrected for multiple comparisons.

In the matched-player analysis (captains' mean baseline 7.10 vs. matched teammates' 7.01), the captain's advantage persisted: the difference in rating–outcome correlations was .20 (match-clustered bootstrap 95% CI [.04, .35]) and in gap–outcome correlations .26 (95% CI [.07, .46]).

Thus, any starter's rating tracks his team's collective performance, but the captain's rating, and especially his salience relative to his own team, is unusually coupled with match outcome compared with random, positionally matched, and quality-matched teammates. Because all measures are post-match ratings, this specificity concerns association, not established influence.

DISCUSSION

Formal captaincy does not necessarily imply effective leadership, but the two may be linked when captains demonstrate strong individual performance. The association between captain and team performance remained robust after accounting for the nested and paired structure of the data and excluding captains from the team-level measure. Moreover, captain performance was more strongly associated with match outcomes than that of similarly skilled non-captain starters.

These findings do not support the notion of the heroic captain. Consistent with evidence that formal captains are rarely perceived as their team's best leader (Fransen et al., 2014), the results do not establish that captains direct team performance or that their individual performance causes team success. Instead, they suggest that formal designation alone does not ensure effective leadership, while demonstrated task competence through effort, decision-making, and consistent execution may reinforce leadership credibility, which is consistent with the concept of leading by example.

The counterfactual analyses clarify what is specific to the captain. That an individual's rating tracks collective performance is a general property of interdependent play, shared by randomly selected starters. What distinguished captains was the association of their performance, and especially of their relevance within their own team, with match outcome. The captain-specific signal therefore lies not in playing well in absolute terms, which strong players of any designation do, but in the distinctive coupling between the designated leader's

performance and the team's result. This converges with evidence from seven international tournaments in which a generic standout performance had no clear association with outcomes independent of the team's mean rating (Ball et al., 2026): here the same held for random starters, but not for captains. Because that study used another provider and a threshold rather than a continuous gap, the comparison is just indicative.

Three alternative explanations deserve consideration. First, squad quality could drive both variables, as stronger teams perform better and appoint stronger captains; the partial correlation controlling for FIFA ranking (r = .26) and the paired design, in which the ranking difference lost significance once performance entered, indicate that team quality accounts for only part of the association. Second, the coupling could be compositional; this is excluded by construction, since the collective measure omits the captain, and individual–collective coupling is in any case a property of any starter, as expected when fielded players' inputs are complements (Franck & Nüesch, 2010). Third, causation may run through the measurement instrument: ratings reward decisive actions, winning teams generate more of them, and favourable outcomes may raise captain ratings rather than the reverse. Sofascore ratings averaged 7.89, 7.58, and 7.14 for players on winning, drawing, and losing sides in a 60-player sample (García-Fernández & Quian, 2026), a gradient similar to that observed here, and reporters' ratings shifted by about a quarter of a standard deviation with the result of penalty shoot-outs (Kausel et al., 2019). Neither study separates legitimate performance differences from rating bias in algorithmic scores, but both indicate that a rating–outcome association cannot by itself show that performance caused the result; more generally, when individual and team outcomes are jointly determined, their association cannot reveal which influences which (Manski, 1993). This explanation cannot be adjudicated with the present data. These findings show that captain performance is more strongly associated with match outcomes than that of comparable teammates, although the direction of influence remains unclear. Importantly, the captain's rating reflects individual task contribution rather than leadership effectiveness (Judge et al., 2002).

Reciprocal performance and outcome asymmetry

The asymmetry between wins and losses supports a reciprocal interpretation. If captain performance were a direct and symmetrical cause of collective performance, poor captain performances would accompany defeats as clearly as strong ones accompany victories. Instead, collective success had a sharper individual signature than collective failure. This interpretation is consistent with the view of football as a complex adaptive system (Passos et al., 2013). Poor coordination, tactical mismatches, opposition quality, and individual errors may simultaneously undermine captain and team performance. Conversely, team success may provide opportunities for individual contributions to become more visible and receive greater recognition.

One plausible explanation is that individual performance becomes more visible in successful teams. Although captains were generally strong players, they rarely ranked among their squad's best (64th percentile in baseline quality). Yet, in victories, their performance exceeded the team average by more than one standard deviation. If the role concentrates functions that become visible when the team succeeds—assuming responsibility in decisive moments and taking high-leverage actions such as set pieces and penalties—event-based ratings will register captain salience most strongly in wins. Further, the captain's rating will couple with outcome more tightly than that of comparable teammates, as observed. Distinguishing this account from an independent captain effect requires designs that measure leadership behaviour directly.

Together, the results support a reciprocal, loosely coupled captain–team model (Orton & Weick, 1990): captain performance may strengthen collective confidence and execution, while effective team play creates the conditions for the captain to perform.

 © 2027 ARD Asociación Española

Position and the legibility of leadership performance

The positional analysis provides a qualification to consider: the association between captain performance and match outcomes did not differ significantly across positions, although the relationship between captains outperforming their teammates and match outcomes was most evident among forwards. This pattern suggests that individual contributions are more readily captured by performance measures in attacking positions, rather than indicating that leadership is more relevant in these roles.

Rating algorithms do not reward all actions equally. In all three platforms used here, offensive actions such as shots on target, key passes, and successful take-ons raise a player's rating the most, whereas defensive actions such as clearances raise it less (Ball et al., 2025). In addition, the statistics that reflect a player's contribution to the team differ by position (Gelade & Hvattum, 2020). The distributed work of midfield leadership, such as controlling tempo, supporting the press, and maintaining compactness, is central to coordination but hard to register as discrete events. The approach is therefore most sensitive to leadership expressed through visible, decisive actions and least sensitive to leadership enacted through communication, spatial organisation, or emotional regulation; pairing ratings with tracking data and observational coding would show whether midfield captains lead through channels that current models cannot see.

Exceptional cases and distributed leadership

Individual cases illustrate how the captain's contribution and the team's result combine. Messi is the clearest instance of a captain who stood out from his team throughout a run: he rated above 9.0 in five of his seven starts (mean 8.90, the highest among captains with at least two starts) while his teammates averaged 6.91, and Argentina won all six matches he started before the final. In the final, his advantage narrowed to 0.31 points (7.1 vs. 6.79) and Argentina lost. The pattern fits a team whose results were closely tied to its captain's salience, although the ratings cannot show whether he raised his teammates' level or compensated for it. Mbappé shows the limit of this route: his 9.9 in the third-place defeat, the largest captain advantage of the tournament, could not overcome weaknesses in the wider system, and the opposing captain also stood out in that match.

Rodri represents a different route. Among captains with at least two starts, he ranked fourth in mean rating (7.76) and stood out only modestly from teammates whose ratings were the second highest in the tournament (mean gap 0.62 points). Yet he never rated below 6.8, exceeded his teammates in six of eight matches, and Spain lost none of them, winning the final with one of his highest ratings (8.4). His case reflects the distributed work of midfield leadership: controlling tempo, supporting the press, maintaining compactness.

Other teams progressed although their captain stood out little. England reached the semi-finals, winning five of Kane's seven starts, while his mean advantage over his teammates was only 0.34 points, and the quarter-final against Norway was won with his rating a point below theirs (6.0 vs. 6.99). Morocco reached the quarter-finals with Hakimi, a defender, rated essentially at team level (mean gap 0.09 points), including a round-of-32 win over the Netherlands on penalties with a 6.3 rating. Such cases are consistent with shared-leadership and social-identity approaches, in which informal leaders assume task, motivational, or social roles (Mertens et al., 2021) and shape performance through identification and teamwork (Fransen et al., 2020), and with evidence that less centralised teams score more (Grund, 2012).

These exceptions refine rather than weaken the argument: formal captaincy acquires competitive relevance when the captain visibly contributes at a high level, while outcomes depend on reciprocal, distributed, and positionally differentiated processes. The captain is neither an automatic hero nor reducible to a rating, but one consequential node within a wider network.

Practical implications

Practical implications must be drawn with restraint. The study did not compare captain-selection criteria, and the findings do not show that appointing a stronger performer as captain improves outcomes. More modestly, the captain's match contribution and significance relative to his team can be profiled across a competition, and staff may monitor it alongside, not instead of, the motivational, social, and organisational functions of the role. Mid-match armband transfers and wins achieved despite below-team captain performance further show that the role is enacted situationally and that distributed leadership structures matter.

Scope: national-team tournaments and club football

Several findings are specific to the World Cup setting. National teams assemble intermittently, so captains lead groups with shorter shared histories than club captains; the format compresses competition into at most eight matches with single-elimination rounds, which may heighten the prominence of decisive actions; and national-team captaincy carries a symbolic, representational weight with no exact club equivalent. By contrast, the coupling of individual and collective ratings, the measurement logic of event-based ratings, and the analytical requirements identified here—separating the captain from the collective measure, modelling repeated and paired observations, and benchmarking against non-captains—should transfer to club research. Whether the captain-specific coupling with outcome holds over a league season remains an open question.

Limitations

First, the design is observational; causal direction cannot be established, and reverse causality through rating algorithms remains embedded in every estimate. Only four matches were decided on penalties, too few to exploit the quasi-exogenous variation used by Kausel et al. (2019). Second, all performance measures are post-match ratings from proprietary systems (García-Fernández & Quian, 2026); they capture observable technical–tactical output rather than motivational, social, or communicational leadership, and event-level statistics explain only 22–38% of the variance in players' contribution to team goal difference (Gelade & Hvattum, 2020). Human assessments offer no validated benchmark either, given their poor inter-rater reliability (ICC = .04–.14; Bergkamp et al., 2022). Third, the primary analyses rest on Sofascore; the focal associations replicated across WhoScored, FotMob, and the composite, but provider-specific biases cannot be excluded. Fourth, with 66 captains in 48 teams, captain-level and team-level variance cannot be fully separated. Finally, the sample is limited to the 2026 men's World Cup; generalisation to women's, youth, or club football requires dedicated data.

These limitations can outline future research in three layers. The first is analytical: replicating the design in other tournaments, including the 2027 Women's World Cup, and in longitudinal club data, where repeated seasons would separate captain- from team-level variance and allow captaincy changes to serve as natural experiments. The second is behavioural: coding on-field leadership acts, such as communication, tactical correction, and emotional regulation, together with tracking data, to establish whether the captain's outcome coupling reflects leadership behaviour or merely visible, decisive actions. The third is perceptual and relational: combining match data with teammates' perceptions of leadership quality and network analyses of leadership distribution, to test whether the captain's credibility mediates any association with outcomes.

CONCLUSIONS

Captaincy matters conditionally, not absolutely. Any starter's rating rose and fell with that of his teammates, and the captain's did so no more than theirs. What distinguished the captain was the coupling of his performance, and especially of his salience within his own team, with the match result: this coupling exceeded

 © 2027 ARD Asociación Española

that of random, positionally matched, and quality-matched teammates. These findings, obtained from match-level ratings rather than perceptual instruments, show that observational designs can complement the psychometric tradition of athlete-leadership research in elite settings where questionnaires are rarely feasible.

This coupling was asymmetric. Victories carried a clear individual signature, with captains standing out from their teammates by more than one teammate standard deviation, whereas in defeats captain and teammates declined together. The asymmetry favours a reciprocal, loosely coupled model over a captain who directly causes results: collective success creates the conditions for the captain's contribution to become visible, while collective failure can arise from many sources that no individual performance offsets. Because captains were good but rarely the best players of their squads, their relevance in victories may reflect the role's concentration of decisive actions; whether this constitutes leadership or visible task contribution cannot be established with post-match ratings, and the direction of influence remains open.

Moreover, the routes to success were plural. Some teams advanced with a captain who stood out throughout (Messi), others with a midfield captain whose contribution was consistent rather than prominent (Rodri), and others with a captain who barely exceeded his teammates (Kane, Hakimi). Because event-based ratings reward offensive actions most, the approach sees leadership expressed through decisive actions more clearly than leadership exercised through coordination, communication, or emotional regulation. The captain thus emerges as one consequential node within a distributed leadership network rather than its sole driver.

A broader lesson follows. One element of formal leadership is the leader's own performance within the team: the armband acquired competitive relevance when its holder visibly contributed at a high level to the shared task. The credibility of designated leadership therefore rests partly on leading by example, on performing well and not only on holding the role. This principle may extend to other collaborative settings, such as work teams, organisations, and classrooms, where formal leaders are expected both to coordinate others and to contribute to the collective work, although its transfer requires evidence from those contexts.

AUTHOR CONTRIBUTIONS

All authors meet the established criteria for authorship. Contributions are reported according to the CRediT taxonomy as follows:

Conceptualisation: Fernando A. Mora-Davila. Methodology: Natalia Rocha. Formal analysis: Fernando A. Mora-Davila and Natalia Rocha. Investigation: Fernando A. Mora-Davila. Data curation: Fernando A. Mora-Davila. Writing – original draft: Fernando A. Mora-Davila. Writing – review & editing: Natalia Rocha. Supervision: Fernando A. Mora-Davila.

All authors have critically reviewed and approved the final version of the manuscript and agree to be accountable for all aspects of the work.

FUNDING

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

CONFLICT OF INTEREST

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this manuscript.

ETHICAL CONSIDERATIONS

This study analysed publicly available match-level information from the 2026 FIFA World Cup (official squad lists, line-ups, and match reports) and performance ratings published by commercial football-data platforms. It involved no recruitment of participants, no intervention, and no information beyond what is publicly reported, so formal ethics approval was not required. Because the dataset nevertheless processes named, player-level information, three safeguards were applied: processing was limited to publicly reported professional match performance, with no private or off-field information; individual players are named only where editorially necessary, with exceptional cases documented in supplementary material; and results are interpreted at the aggregate level, with ratings framed as post-match associations rather than judgments of individual ability or leadership. All procedures complied with the ethical standards and laws of the country in which the research was conducted, the Declaration of Helsinki, and applicable data-protection principles.

AVAILABILITY OF DATA AND MATERIALS

The dataset was compiled from public sources: official FIFA 2026 World Cup squad lists, line-ups, and match reports, and player ratings published by Sofascore, FotMob, and WhoScored, collected manually from the platforms' public match pages without automated extraction. Because the ratings are proprietary and their redistribution is governed by the platforms' terms of service, the compiled dataset is not publicly deposited; it is available from the corresponding author on reasonable request for non-commercial verification, and the per-match documentation (match identifiers, page references, capture dates) allows every rating to be re-collected. The analysis code is available on the same basis.

AI USE DISCLOSURE

In accordance with current publishing ethics and transparency recommendations, artificial intelligence (AI) tools were used as described in the Material and Methods (Use of AI tools): Claude (Anthropic) supported data verification, the coding and execution of the statistical analyses, the production of figures, copyediting for language, readability, spelling, and grammar. No AI tools were used to generate the scientific content: the authors designed the study, collected the data, chose the analytic strategy, interpreted the results, and formulated the conclusions, and they verified the AI-assisted code outputs. The authors retain full responsibility for the content of the manuscript and confirm its originality, integrity, and accuracy.

REFERENCES

Ball, J., Huynh, M., & Varley, M. C. (2025). Comparing player rating systems as a metric for assessing individual performance in soccer. *Journal of Sports Sciences, 43*(7), 676–686. https://doi.org/10.1080/02640414.2025.2471208

Ball, J., Huynh, M., & Varley, M. C. (2026). Understanding match outcomes in international soccer tournaments through team performance, team strength, and individual contributions. *Journal of Sports Analytics*. Advance online publication. https://doi.org/10.1177/22150218261495195

 © 2027 ARD Asociación Española

Bergkamp, T. L. G., Meijer, R. R., den Hartigh, R. J. R., Frencken, W. G. P., & Niessen, A. S. M. (2022). Examining the reliability and predictive validity of performance assessments by soccer coaches and scouts: The influence of structured collection and mechanical combination of information. *Psychology of Sport and Exercise, 63*, Article 102257. https://doi.org/10.1016/j.psychsport.2022.102257

Chelladurai, P., & Saleh, S. D. (1980). Dimensions of leader behavior in sports: Development of a leadership scale. *Journal of Sport Psychology, 2*(1), 34–45. https://doi.org/10.1123/jsp.2.1.34

Cotterill, S. T., & Fransen, K. (2016). Athlete leadership in sport teams: Current understanding and future directions. *International Review of Sport and Exercise Psychology, 9*(1), 116–133. https://doi.org/10.1080/1750984X.2015.1124443

Cotterill, S. T., Loughead, T. M., & Fransen, K. (2022). Athlete leadership development within teams: Current understanding and future directions. *Frontiers in Psychology, 13*, Article 820745. https://doi.org/10.3389/fpsyg.2022.820745

Decroos, T., Bransen, L., Van Haaren, J., & Davis, J. (2019). Actions speak louder than goals: Valuing player actions in soccer. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (pp. 1851–1861). ACM. https://doi.org/10.1145/3292500.3330758

Franck, E., & Nüesch, S. (2010). The effect of talent disparity on team productivity in soccer. *Journal of Economic Psychology, 31*(2), 218–229. https://doi.org/10.1016/j.joep.2009.12.003

Fransen, K., Haslam, S. A., Mallett, C. J., Steffens, N. K., Peters, K., & Boen, F. (2017). Is perceived athlete leadership quality related to team effectiveness? A comparison of three professional sports teams. *Journal of Science and Medicine in Sport, 20*(8), 800–806. https://doi.org/10.1016/j.jsams.2016.11.024

Fransen, K., McEwan, D., & Sarkar, M. (2020). The impact of identity leadership on team functioning and well-being in team sport: Is psychological safety the missing link? *Psychology of Sport and Exercise, 51*, Article 101763. https://doi.org/10.1016/j.psychsport.2020.101763

Fransen, K., Vanbeselaere, N., De Cuyper, B., Vande Broek, G., & Boen, F. (2014). The myth of the team captain as principal leader: Extending the athlete leadership classification within sport teams. *Journal of Sports Sciences, 32*(14), 1389–1397. https://doi.org/10.1080/02640414.2014.891291

García-Fernández, R., & Quian, A. (2026). El impacto de las aplicaciones estadísticas en el periodismo deportivo: Estudio comparativo entre valoraciones algorítmicas y humanas en el fútbol de élite [The impact of statistical applications on sports journalism: A comparative study of algorithmic and human ratings in elite football]. *Estudios sobre el Mensaje Periodístico, 32*(2), 453–475. https://doi.org/10.5209/esmp.103702

Gelade, G. A., & Hvattum, L. M. (2020). On the relationship between +/− ratings and event-level performance statistics. *Journal of Sports Analytics, 6*, 85–97. https://doi.org/10.3233/JSA-200432

Grund, T. U. (2012). Network structure and team performance: The case of English Premier League soccer teams. *Social Networks, 34*, 682–690. https://doi.org/10.1016/j.socnet.2012.08.004

Judge, T. A., Bono, J. E., Ilies, R., & Gerhardt, M. W. (2002). Personality and leadership: A qualitative and quantitative review. *Journal of Applied Psychology, 87*(4), 765–780. https://doi.org/10.1037/0021-9010.87.4.765

Kausel, E. E., Ventura, S., & Rodríguez, A. (2019). Outcome bias in subjective ratings of performance: Evidence from the (football) field. *Journal of Economic Psychology, 75*, Article 102132. https://doi.org/10.1016/j.joep.2018.12.006

Kharrat, T., McHale, I. G., & López Peña, J. (2020). Plus-minus player ratings for soccer. *European Journal of Operational Research, 283*, 726–736. https://doi.org/10.1016/j.ejor.2019.11.026

Liu, H., Hopkins, W., Gómez, M. A., & Molinuevo, J. S. (2013). Inter-operator reliability of live football match statistics from OPTA Sportsdata. *International Journal of Performance Analysis in Sport, 13*(3), 803–821. https://doi.org/10.1080/24748668.2013.11868690

Loughead, T. M., Hardy, J., & Eys, M. A. (2006). The nature of athlete leadership. *Journal of Sport Behavior, 29*(2), 142–158.

Manski, C. F. (1993). Identification of endogenous social effects: The reflection problem. *The Review of Economic Studies, 60*(3), 531–542. https://doi.org/10.2307/2298123

Mertens, N., Boen, F., Steffens, N. K., Haslam, S. A., & Fransen, K. (2021). Will the real leaders please stand up? The emergence of shared leadership in semi-professional soccer teams. *Journal of Science and Medicine in Sport, 24*(3), 281–290. https://doi.org/10.1016/j.jsams.2020.09.007

Orton, J. D., & Weick, K. E. (1990). Loosely coupled systems: A reconceptualization. *Academy of Management Review, 15*(2), 203–223. https://doi.org/10.2307/258154

Passos, P., Araújo, D., & Davids, K. (2013). Self-organization processes in field-invasion team sports: Implications for leadership. *Sports Medicine, 43*(1), 1–7. https://doi.org/10.1007/s40279-012-0001-1

Piasecki, P. A., & Filho, E. (2025). The influence of social and task peer-leaders on team cohesion and collective efficacy in sports: A meta-analytical review. *Sport Sciences for Health, 21*(3), 1349–1362. https://doi.org/10.1007/s11332-025-01375-0

Price, M. S., & Weiss, M. R. (2011). Peer leadership in sport: Relationships among personal characteristics, leader behaviors, and team outcomes. *Journal of Applied Sport Psychology, 23*(1), 49–64. https://doi.org/10.1080/10413200.2010.520300



This work is licensed under a Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0).

 © 2027 ARD Asociación Española