

The Thin Line: AI Performance Evaluation and Risk-Taking

Andrea Burro

This version: October 26, 2026

[Click here for the most recent version](#)

Abstract

How does replacing human evaluators with AI affect workers' risk-taking? I develop a model in which human leniency reduces the expected cost of risky strategies by allowing some mistakes to go unpenalized. Removing this leniency through AI evaluation lowers risk-taking. To test the model, I exploit the staggered adoption of Electronic Line Calling (ELC), an AI-based camera system that replaces human line judges in professional tennis. The sample includes 12,891 matches, 233 players, and 94 tournaments from 2021 to 2025. I compare the same player across tournaments with and without ELC. I find that ELC reduces the ace rate and the double-fault rate, and raises the first-serve-in rate. Players also change where they aim their serves, which shows that the effects reflect a change in behavior rather than a mechanical consequence of more accurate calls.

University of St. Gallen, SEW-HSG, and LEAP-INSPIRE. andrea.burro@unisg.ch.

I am grateful to Beatrix Eugster for extensive discussions and her mentorship. I thank David Almog, Enzo Brox, Bruno Caprettini, Alex Krumer, Federica Mascolo, Víctor Sancibrián, Jeremia Stalder and seminar participants at the Bocconi PhD Seminar for helpful comments and suggestions. I thank Bocconi University for hosting me while working on this project. All errors are my own.

1 Introduction

AI is increasingly used to evaluate job candidates, workers, and students. Replacing human judgment with an algorithm may also change how the people being evaluated behave. More accurate evaluation catches more mistakes, but it may discourage people from taking productive risks. Yet causal evidence on how people adjust their risk-taking when AI is introduced remains limited.

In this paper, I develop a model of risk-taking under different evaluation regimes: AI or human. A worker chooses how much risk to take, trading off the probability of success against the probability of a mistake. Human evaluators sometimes overlook mistakes, lowering the expected cost of risk-taking. AI evaluation removes this leniency, so workers take less risk.

Testing this model is challenging when an AI evaluator is introduced for all individuals at once, leaving no comparison group, or when AI only assists a human evaluator who retains the final decision. Professional tennis avoids both problems. Electronic Line Calling (ELC), an AI-based camera system that replaces human line judges, was adopted at different times across tournaments. I exploit this staggered adoption to compare players' serving across tournaments with and without ELC, while the task, the stakes, and the player pool remain unchanged.¹

In this setting, the risky strategy is an aggressive serve, and human line judges are asymmetrically lenient on close calls (Almog et al. 2025): they are more likely to call in a ball that is truly out than to call out a ball that is truly in. A serve wrongly called in stays in play, so

¹The comparison regime includes different review procedures: many non-ELC tournaments used Hawk-Eye to review calls challenged by players, while clay tournaments relied on human ball-mark review. Because challenges already corrected part of human leniency, my estimates likely understate, in absolute value, the full effect of moving from human to AI evaluation.

the server keeps a chance of winning the point. If AI removes that margin of human error, players might take less risk.

I test the model using a panel of 12,891 matches on the Association of Tennis Professionals (ATP) tour, covering 233 players and 94 tournaments from 2021 to 2025. I estimate the main specification by ordinary least squares (OLS), including player, year, and tournament fixed effects.

ELC adoption reduces the ace rate by 0.35 percentage points (about 5% of the control mean) and the double-fault rate by 0.16 percentage points (about 5%). The first-serve-in rate increases by 0.44 percentage points (about 1%). If players did not change their serving strategies, removing human leniency would mechanically lower the first-serve-in rate and increase the double-fault rate, since some out serves previously called in would now be recorded as faults. The estimated effects go in the opposite direction, suggesting a behavioral adjustment by players. As a placebo test, I find no effect on the break-point save rate, an outcome that also depends on rally dynamics and pressure, and not only on serve strategies. In addition, using a subsample of matches with detailed information on serve placement, I find that players react to ELC by adjusting the direction of their serves. This reaction, together with the results on serve outcomes that move in the opposite direction of what a mechanical adjustment would predict, points to an actual change in behavior when AI evaluates performance.

The model predicts that scaling both the rewards and the losses by the same positive factor should not affect behavior. What matters instead is pressure: the relative cost of making a fault. Consistent with the model, the effects are similar across tournament prestige, from

ATP 250 events to Grand Slams, while the decline in double faults is larger in the later rounds, when pressure is higher.

The results do not replicate in the sample of professional female players from the Women’s Tennis Association (WTA). I discuss four possible explanations. First, WTA serves are slower than in the ATP sample, which lowers the marginal return to serving aggressively. Second, slower serves leave judges fewer close calls to decide in the server’s favor. Third, male players’ serve choices may be harder for opponents to predict than female players’ (Gauriot, Page, and Wooders 2023), which gives men more scope to adjust their serving when incentives change. Fourth, gender differences in risk-taking may play a role (Croson and Gneezy 2009; Charness and Gneezy 2012).

This paper contributes to three strands of the literature. First, while extensive literature examines algorithms that replace or assist human decision-making, these studies have mostly focused on the human evaluator (Kleinberg et al. 2018; Agarwal et al. 2023; Cho 2024; Angelova, Dobbie, and Yang 2025). A smaller set of papers examines those being evaluated and finds that they change their behavior when an algorithm judges them: workers produce longer output, but of lower quality conditional on its length, under an AI evaluator (Almog, Lippman, and Martin 2026), and they present themselves as more analytical and less intuitive (Goergen, de Bellis, and Klesse 2025). This evidence comes from experiments, while my field setting involves professionals facing real stakes and avoids the biases that can arise in the lab (Palacios-Huerta 2025). Jabarian and Henkel (2025) also study those being evaluated in a field experiment, but in their setting AI only advise, and a human retains the final decision. Under ELC, the AI makes the final call rather than a human, so I provide field evidence on

how those being evaluated react.

Second, the principal-agent literature on subjective performance evaluation shows that an evaluator’s discretion shapes an agent’s effort and risk-taking (Baker, Gibbons, and Murphy 1994; Prendergast and Topel 1996; Prendergast 2002; MacLeod 2003). Discretion allows for systematic leniency, which changes the agent’s optimal behavior. However, empirical evidence on this mechanism remains limited, because leniency is rarely observed directly or cleanly removed. The closest field evidence comes from hiring, where limiting human discretion increases the quality of new hires (Hoffman, Kahn, and Li 2018). In their setting, however, evaluated candidates are unaware of the change in evaluation, making it difficult to study their behavioral response. In contrast, the adoption of ELC in professional tennis is a salient change for the evaluated players, allowing me to identify the moment leniency is removed and estimate both the sign and magnitude of the behavioral response.

Third, economists have long used sports data to study strategic behavior under high stakes. These studies examine whether professionals play equilibrium strategies (Walker and Wooders 2001; Palacios-Huerta 2003; Gauriot, Page, and Wooders 2023; Anderson et al. 2025), their risk-taking behavior (Genakos and Pagliero 2012; Ely, Gauriot, and Page 2017), how performance responds to pressure, momentum, and the order of play (Cohen-Zada et al. 2017; Krumer and Lechner 2017; Cohen-Zada, Krumer, and Shapir 2018; Gauriot and Page 2019; Paserman 2023), and how temperature and home advantage affect performance (Burke et al. 2023; Gucciardi and Ruberti 2025). In all these settings, however, the evaluator stays the same while the underlying incentives shift. In my setting, incentives remain constant and only the evaluator changes; therefore, the reaction I study is not contaminated by a change in

incentives. To the best of my knowledge, no work in this literature has used the staggered rollout of ELC to identify players' behavioral responses.

The closest paper is Almog et al. (2025), which studies the introduction of Hawk-Eye review in tennis but focuses on the umpire, while I study the player. It also studies the earlier transition from human line judges to AI-assisted review through Hawk-Eye challenges. In contrast, I study the later transition, from AI-assisted review to fully automated calls.

The paper proceeds as follows. Section 2 describes the institutional setting. Section 3 develops the theoretical framework. Section 4 describes the data and Section 5 the empirical strategy. Section 6 presents the results. Section 7 discusses the findings and concludes.

2 Institutional Setting

Tennis is an individual sport in which two players hit a ball over a net, each trying to land it inside the opponent's half of a marked rectangular court. A player wins a point when the opponent fails to return the ball into the opponent's court. Every point starts with a serve, and the server gets two attempts to land the ball in a service box.

Tournaments are played on hard, clay, and grass courts, which differ in the speed and height of the ball's bounce. Clay generally slows the ball and produces a higher bounce, giving the receiver more time to return the serve, while hard and grass courts produce a faster bounce, which favors the server. These differences affect the advantage players gain from an aggressive serve and can therefore influence their serving choices.²

²See the International Tennis Federation (ITF) tennis glossary and the ITF court pace classification for the official definitions and surface pace ratings.

Line-call accuracy matters throughout a point, but is most consequential on serves, where the server unilaterally chooses the target. A second serve called out is a double fault and costs the server the point. Historically, nine human line judges positioned around the court made the calls, and the chair umpire adjudicated disputes.

In 2006, the ATP introduced Hawk-Eye, a ball-tracking system originally developed for cricket. Hawk-Eye initially served as an appeal mechanism rather than an automatic line caller: players were granted two to three challenges per set to dispute a line call, and if Hawk-Eye confirmed the call was incorrect, the challenge was not deducted. In a sample of over 2,000 challenges from 35 tournaments, Abramitzky et al. (2012) find that 38% of challenges overturned the original call and that 92% of the challenged calls involved balls within 100 mm of the line. In this phase, human line judges kept primary responsibility for line calls, while Hawk-Eye provided technological oversight. This oversight mechanism introduced an unintended behavioral response: under the challenge system, line judges were more likely to call close balls in even when they were out (Almog et al. 2025).

In 2021, tournaments began replacing the Hawk-Eye challenge system with Electronic Line Calling (ELC), under which Hawk-Eye’s ball-tracking technology automatically makes and announces every line call in real time. This change removes human line judges from the court. It completes the transition from human to AI evaluation, the second step after the introduction of challenges in 2006, with ELC serving as the only evaluator of whether a ball lands in or out.

Rather than providing a video replay for a human to review, ELC uses feeds from 12 to 18 cameras that track the ball at high frame rates, together with neural networks trained to

distinguish it from the background, player shadows, and motion blur in each frame. The system triangulates these feeds into a three-dimensional flight path. It then models the ball’s deformation on impact to reconstruct the footprint it leaves on the court and make an unappealable in-or-out decision.

ELC adoption was staggered across tournaments and over time, as Figure 1 shows. Two of the four Grand Slam tournaments, the Australian Open and the US Open, adopted ELC in 2021, followed by hard-court and grass-court tournaments that adopted the technology gradually over subsequent years (Figure A.1 shows the rollout by surface). The main tournament on grass, Wimbledon, adopted ELC only in 2025, and the other Grand Slam tournament, Roland Garros, never adopted it (Figure A.2 shows the rollout by prestige). The main motivation for introducing ELC was to reduce on-court staff during the COVID-19 pandemic and limit contagion (The New York Times, 2024).

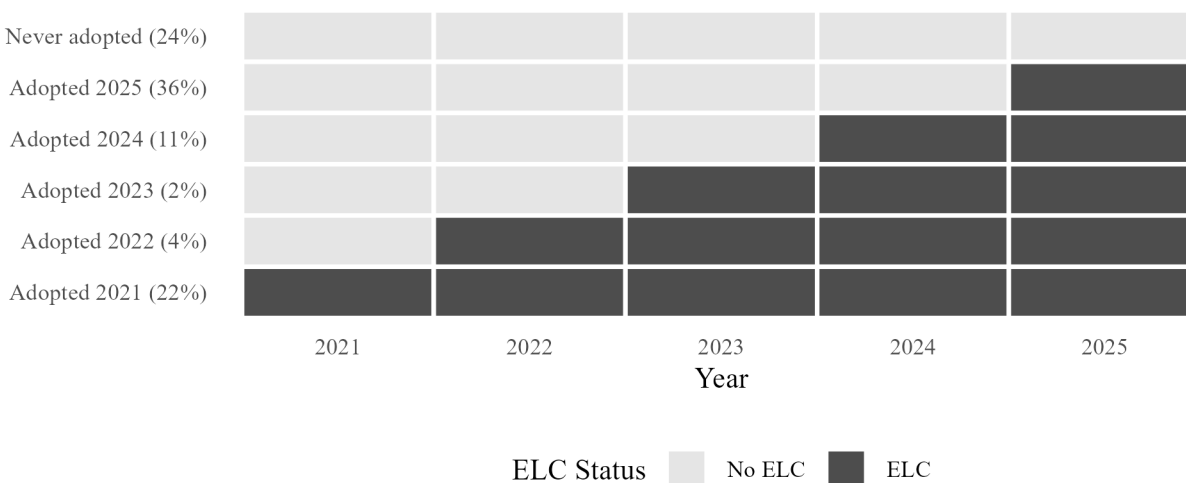


Figure 1: Staggered Adoption of Electronic Line Calling. Rows group ATP tournaments by the year they first adopted ELC; shaded cells mark the years in which a group used ELC. Percentages show each group’s share of all ATP tournaments in the sample. The sample covers official ATP tour events from 2021 to 2025. ELC adoption is manually verified against tournament announcements, broadcasts, and news reports.

3 Theoretical Framework

To motivate the empirical design, I develop a model of risk taking under different evaluation regimes. The framework describes how human leniency changes the incentive to pursue risky strategies. I apply this setup to the tennis setting to derive testable empirical predictions regarding player behavior.

3.1 Setup

A worker chooses the riskiness of a strategy, a , to maximize expected payoff. Risk-taking has two opposing effects. It makes the strategy less likely to meet the required standard, so the probability that the strategy is valid, $p_V(a)$, decreases in a . A strategy that fails the standard is invalid, which I call a mistake. Risk-taking also increases the probability of success conditional on the strategy being accepted by the evaluator, with $q(a)$ increasing in a .

If the evaluator detects a mistake, the strategy is rejected, and the workers receives no reward; if the evaluator overlooks a mistake, strategy is accepted, and the workers has a chance of success. I assume that the probability of success, $q(a)$, is the same whether the strategy is accepted because it is valid or because a mistake was overlooked.

I assume throughout that $p_V(a) q(a)$ and $q(a)$ are strictly concave in a , and that the optimum is interior under both evaluation regimes. Concavity of $q(a)$ means that additional risk adds less to the probability of success, and concavity of $p_V(a) q(a)$ ensures a unique interior optimum.

With payoff W for success and L for failure, where $W > L$, the expected payoff under

evaluation regime θ_j is

$$\begin{aligned} U_j(a) &= \pi^{\theta_j}(a) [q(a)W + (1 - q(a))L] + (1 - \pi^{\theta_j}(a))L \\ &= L + (W - L) \pi^{\theta_j}(a) q(a). \end{aligned}$$

Here, $\pi^{\theta_j}(a)$ is the probability that the evaluator accepts the strategy under regime j , that is, the probability that it is not rejected as a mistake. The constant L and the positive factor $(W - L)$ do not affect the choice, so the worker maximizes the probability of success, $\pi^{\theta_j}(a) q(a)$.

The two evaluation regimes differ only in how mistakes are detected. AI evaluates without error, so it accepts a strategy exactly when it is valid ($\pi^{\theta_{AI}}(a) = p_V(a)$), while humans sometimes overlook a mistake. Let $\beta \equiv \Pr(\text{accepted} \mid \text{invalid}) > 0$ be this conditional favorable-error rate. Under human evaluation, the strategy is accepted if it is valid or if a mistake is overlooked, so the probability that it is accepted is

$$\pi^{\theta_H}(a) = p_V(a) + \beta(1 - p_V(a)).$$

Differentiating, $\pi^{\theta_H'}(a) = (1 - \beta)p'_V(a)$: human leniency flattens the slope of the probability that the strategy is accepted, in addition to shifting its level.

3.2 Main Result: Removing Human Leniency Reduces Risk-Taking

The first-order condition balances the marginal cost and benefit of risk:

$$\underbrace{\pi^{\theta_{j'}}(a) q(a)}_{< 0: \text{ cost of risk}} + \underbrace{\pi^{\theta_j}(a) q'(a)}_{> 0: \text{ benefit of risk}} = 0.$$

Proposition 1 (Less risk under AI). *Let a_{AI}^* and a_H^* denote optimal risk-taking under AI and human evaluation. Whenever $\beta > 0$, $a_H^* > a_{AI}^*$: the worker chooses a riskier strategy under human evaluation.*

Proof. Let a_{AI}^* satisfy the first-order condition under AI, so $p'_V(a_{AI}^*) q(a_{AI}^*) = -p_V(a_{AI}^*) q'(a_{AI}^*)$.

Evaluating the human derivative at that same point and substituting the AI first-order condition,

$$\begin{aligned} \left. \frac{\partial U_H}{\partial a} \right|_{a=a_{AI}^*} &= (W - L) \left\{ (1 - \beta) p'_V(a_{AI}^*) q(a_{AI}^*) \right. \\ &\quad \left. + \left[p_V(a_{AI}^*) + \beta (1 - p_V(a_{AI}^*)) \right] q'(a_{AI}^*) \right\} \\ &= (W - L) \beta q'(a_{AI}^*) > 0. \end{aligned}$$

At the AI optimum the human payoff is still rising in a , so by concavity the human optimum lies to the right: $a_H^* > a_{AI}^*$. □

By occasionally overlooking mistakes, human evaluators reduce the expected cost of risky strategies, making the worker more willing to take risks. AI evaluation removes this subsidy, so the worker chooses a safer strategy.³

³Humans can also make the opposite mistake, rejecting a valid strategy. Let $\alpha \equiv \Pr(\text{rejected} \mid \text{valid})$ be this conditional unfavorable-error rate. Allowing for errors in both states (valid or invalid) and assuming $\alpha + \beta < 1$, $\pi^{\theta_H}(a) q(a) = (1 - \alpha - \beta) p_V(a) q(a) + \beta q(a)$, where $p_V(a) q(a)$ is the AI objective. The AI objective has zero derivative at a_{AI}^* , so scaling it by $(1 - \alpha - \beta) > 0$ leaves that derivative at zero; only the added term contributes a nonzero slope there, $\beta q'(a_{AI}^*) > 0$, and $a_H^* > a_{AI}^*$ still holds.

I assume that β is constant, so a mistake has the same probability of being overlooked regardless of risk-taking. If mistakes associated with riskier strategies are harder to detect, favorable errors may become more likely as a increases. This would give the worker an additional incentive to take risks under human evaluation. The model treats AI evaluation as exact, but the comparison is unchanged if AI retains a smaller constant favorable-error rate, with unfavorable errors absent in both regimes.

3.3 Extension: Failure Costs and Stakes

In the baseline model, a detected mistake and an unsuccessful strategy yield the same payoff. However, the worker may attach an additional cost to a detected mistake because it prevents any chance of success, and this cost may be greater under pressure. I therefore extend the model by allowing a detected mistake to be more costly than failure after the strategy has been accepted.

I distinguish between the payoff from an unsuccessful strategy that was accepted, L_A , and the payoff from a detected mistake, $L_D \leq L_A$, while keeping $W > L_A$ for success. Expected utility under evaluation regime j is

$$\tilde{U}_j(a) = \pi^{\theta_j}(a) \left[q(a) W + (1 - q(a)) L_A \right] + (1 - \pi^{\theta_j}(a)) L_D.$$

The first term captures the payoff when the strategy is accepted, after which the worker can either succeed or fail, while the second captures the payoff from a detected mistake.

To simplify the expression, I normalize the payoff from an unsuccessful strategy to zero and

the payoff from success to one, which leaves the optimal choice unchanged. I define κ as the additional cost of a detected mistake, relative to the payoff difference between success and an unsuccessful strategy:

$$\kappa \equiv \frac{L_A - L_D}{W - L_A} \geq 0.$$

The normalized payoff from a detected mistake is $-\kappa$, so expected utility becomes

$$\tilde{U}_j(a) = \pi^{\theta_j}(a) q(a) - \kappa(1 - \pi^{\theta_j}(a)).$$

The worker trades off the probability of success against the additional cost of a detected mistake. When $\kappa = 0$, a detected mistake costs no more than an unsuccessful strategy, and the expression reduces to the baseline model. I represent greater pressure by increasing κ and assume the objectives remain strictly concave with interior optima.

Human leniency still encourages risk-taking. For any $\kappa \geq 0$, the derivative of the human objective at the AI optimum is $\beta q'(a_{AI}^*) > 0$, so risk-taking remains higher under human evaluation.⁴

Pressure also makes leniency more valuable at a given level of risk-taking. The difference between the human and AI objectives is

$$\tilde{U}_H(a) - \tilde{U}_{AI}(a) = \beta(1 - p_V(a))(q(a) + \kappa).$$

With probability $\beta(1 - p_V(a))$, a mistake is overlooked under human evaluation, allowing the

⁴At a_{AI}^* , the first-order condition is $p'_V(a_{AI}^*)(q(a_{AI}^*) + \kappa) = -p_V(a_{AI}^*)q'(a_{AI}^*)$. Substituting this into the derivative of the human objective leaves $\beta q'(a_{AI}^*) > 0$.

worker to retain a chance of success and avoid the additional cost κ . As this cost increases, the worker has more to lose from removing leniency at the same level of risk-taking. However, a larger difference between the optimal choices requires additional assumptions and does not follow from this result alone.

A proportional increase in payoffs has a different role. Multiplying W , L_A , and L_D by a common positive factor leaves κ and the optimal choice unchanged. In this formulation, higher stakes do not change behavior unless they also change the relative cost of a detected mistake.

3.4 Application to Tennis and Empirical Predictions

In tennis, the worker is the player and the risky strategy is an aggressive serve. I focus on serves for two reasons: under ELC, AI evaluates them on every point, and the server controls the serving decision. I leave the form of aggressiveness unspecified: a player could change serve speed, choose whether to aim at the lines, exploit the opponent’s weak side, or a combination of the three.

A mistake corresponds to a serve landing out, so $p_V(a)$ is the probability that the serve truly lands in. The evaluator’s verdict is the line call: $\pi^{\theta_j}(a)$ is the probability that the serve is called in. A serve called in allows the point to continue, and $q(a)$ is the probability of winning conditional on that call. Human leniency means a truly out serve is sometimes called in, giving the player a chance to win the point. This corresponds to the asymmetric leniency documented by Almog et al. (2025). Figure 2 illustrates the resulting difference between the two evaluation regimes.

The model predicts less aggressive serving under ELC. This response should reduce the

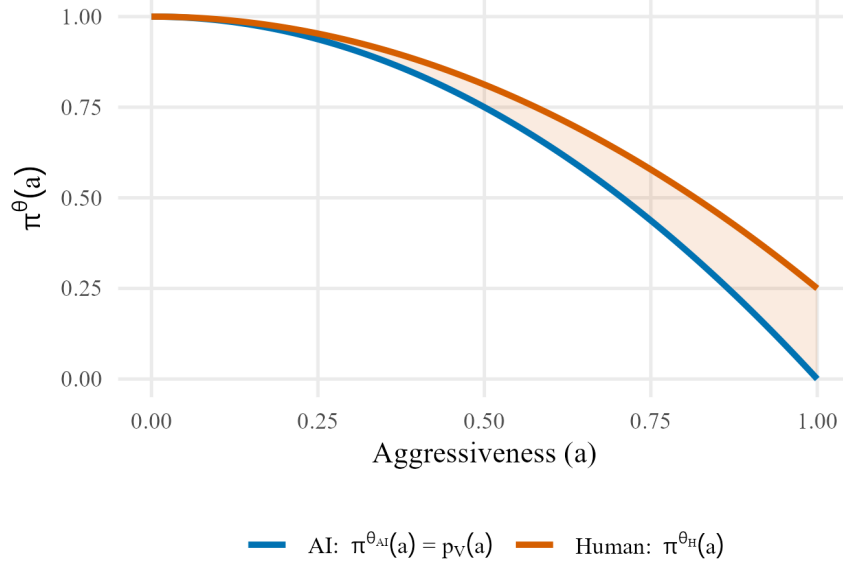


Figure 2: Human Leniency and Serve Aggressiveness. $\pi^{\theta_j}(a)$ is the probability that the serve is called “in” with aggressiveness level a under evaluation regime j (Human or AI); under AI it equals the probability $p_V(a)$ that the serve truly lands in. In this example, $\beta = 0.25$ is the rate at which balls that are truly out are called “in” under human evaluation; the shaded gap is the leniency subsidy $\beta(1 - p_V(a))$, largest where aggressiveness is highest.

chance of a serve landing out and, if aces increase with aggressiveness, lower the ace rate. However, observed fault rates also depend on how serves are called. Holding strategy fixed, ELC mechanically lowers the first-serve-in rate, because more first serves are called out, and raises the double-fault rate, because more second serves are called out. Therefore, a higher first-serve-in rate and a lower double-fault rate require the behavioral response to more than offset these mechanical effects. Beyond these predictions, the framework generates heterogeneous comparative statics along four dimensions.

Pressure versus prestige. In tennis, L_A is the payoff from losing a rally and L_D the payoff from a double fault. Players may attach an additional cost to a double fault because it ends the rally without giving them a chance to compete. I represent greater pressure as an increase in this relative cost, κ . I assume that tournament prestige, reflected in prize money, ranking

points, and public visibility, raises payoffs proportionally, which leaves the optimal choice unchanged. I therefore compare ELC effects across tournament prestige and ask whether the behavioral response is larger in later rounds, when players face greater pressure.

Leniency. β is not directly observed, but pre-ELC double-fault rates may contain information about it: because β is defined conditional on the ball being out, the observed fault probability is $1 - \pi^{\theta_H}(a) = (1 - p_V(a))(1 - \beta)$, which falls as β rises. Lenient calling mechanically suppresses observed faults relative to the true out-rate. Low pre-ELC double-fault rates may therefore indicate greater leniency, although they may also reflect safer serving. I test whether the response is concentrated in these tournaments.

Surface. The response to ELC may differ across surfaces because court conditions affect the return to aggressive serving. I allow $q(a)$ to differ across surfaces, since the same serve may give the player a different advantage depending on how the ball bounces. Where this return is greater, favorable calling errors can provide a stronger incentive to take risk.

Gender. Where the serve gives little advantage, $q'(a)$ is small and the incentive provided by favorable calling errors may be weaker. Since serves are less decisive on the women’s (WTA) tour, the framework motivates testing whether the response is smaller in that sample.

4 Data

I use three datasets that previous studies have also used to answer economic questions (Cohen-Zada, Krumer, and Shapir 2018; Burke et al. 2023; Anderson et al. 2025; Gucciardi and Ruberti 2025). For match-level statistics, I use Jeff Sackmann’s ATP match database,

which covers all official ATP tour events. To replicate the analysis for the WTA tour, I use the same source’s WTA data. To estimate the effect of ELC on serve placement, I use the Match Charting Project.⁵

The ATP panel spans 2021 to 2025 and includes 12,891 matches involving 233 players across 94 tournaments. I restrict the sample to official tour events and exclude exhibition matches, national team competitions (Davis Cup), the Olympics, and the Laver Cup, which have different competitive incentives and player effort levels. I also keep only active players, defined as players with at least 10 matches in a season, to limit the influence of players who appear rarely, for example because of injuries or retirements.

The Match Charting Project contains point-by-point data hand-coded by volunteers, including the direction of each serve, which the official statistics do not record. I match it to the main panel by tournament, date, and player name. The matched subsample includes 2,308 ATP matches played by 199 players across 87 tournaments. The two datasets differ in coverage and in how the data are recorded: the main panel covers every official match but reports only aggregate serve statistics, while the Match Charting Project reports serve direction for a smaller subsample of matches, selected by volunteers rather than at random, and coded by hand rather than tracked automatically. Serve-direction outcomes may therefore contain measurement error and may not represent all matches, but prior work has already used these data to test players’ serve strategies (Anderson et al. 2025).

The treatment variable $ELC_{gt} \in \{0, 1\}$ indicates whether tournament g operated under

⁵All three sources were publicly available at <https://github.com/JeffSackmann> at the time of data collection (March–May 2026). The original ATP and WTA repositories are no longer available, but an archived copy of these databases is available at <https://github.com/Aneeshers/tennis-sackmann-archive>. The datasets used in this paper are available upon request.

Electronic Line Calling in year t . I construct it through a manual verification protocol, cross-checking each tournament-year against official communications, tournament announcements, and match highlights. Where official records disagree or are missing, I verify ELC status through tournament broadcasts and news reports. Once a tournament adopted ELC, it used the technology in every match, so all players within a tournament-year faced the same conditions. Of the 12,891 matches in the ATP sample, 52% were played under ELC.

I use match-level statistics to construct four outcome variables that capture different aspects of serving performance. The ace rate, aces divided by total serve points, is my main proxy for aggressive serving success. In contrast, the double-fault rate, double faults divided by total serve points, captures the downside of excessive aggressiveness on second serves. The first-serve-in rate is the share of total serve points where the first serve lands in, and it is expected to increase as players serve less aggressively. The fourth outcome, the break-point save rate, is the share of break points faced that the server wins. A break point lets the receiver win the game on the opponent's serve; the server "saves" it by winning that point. Because saving a break point depends on both players' performance, including rally dynamics and their responses to the game situation, I use this outcome as a placebo for which I predict no effect of ELC.

Table 1 reports descriptive statistics by ELC status for the ATP sample. Under human evaluation, the mean ace rate is 7.02%, the mean double-fault rate is 3.37%, and the mean first-serve-in rate is 62.88%. Under ELC, these are 8.89%, 3.65%, and 62.16%, respectively. These raw differences partly reflect tournament selection, as ELC was adopted first on faster hard-court surfaces, which mechanically produce more aces and faults. The fixed effects in

the main specification absorb these stable tournament differences.

Table 1: Descriptive Statistics by ELC Status

	No ELC (Human)	ELC (AI)
Ace rate	0.070	0.089
First-serve-in rate	0.629	0.622
Double-fault rate	0.034	0.037
Break-point save rate	0.588	0.603
Player rank (mean)	58.556	56.600
Observations	11,361	12,261

The unit of observation is a player-match. ELC is a dummy variable equal to one if the tournament operated under Electronic Line Calling in that year. Each row reports the mean by ELC status. Ace rate, first-serve-in rate, double-fault rate, and break-point save rate are expressed as proportions. Player rank is the ATP singles ranking, where lower values indicate better-ranked players. The sample covers official ATP tour events from 2021–2025. Match data are from Jeff Sackmann’s tennis repositories (<https://github.com/JeffSackmann>).

5 Empirical Strategy

To estimate how ELC affects serve outcomes, I use the following fixed-effects regression:

$$y_{imgt} = \lambda_i + \lambda_g + \lambda_t + \gamma \text{ELC}_{gt} + \delta \mathbf{X}_{imgt} + \varepsilon_{imgt},$$

where y_{imgt} is the outcome for player i in match m at tournament g in year t . Player fixed effects λ_i absorb stable differences in playing style. Tournament fixed effects λ_g absorb stable tournament characteristics: surface, prestige, and court conditions. Year fixed effects λ_t absorb common trends in serve behavior, equipment, and rule changes. The coefficient of interest, γ , is identified from variation in ELC adoption across tournaments and over time.

This source of variation also explains why I do not use an estimator developed for staggered adoption (e.g., Callaway and Sant’Anna 2021; Sun and Abraham 2021; Borusyak, Jaravel, and Spiess 2024). Those estimators assume that a unit stays treated once treated, and assign

each unit to a cohort by the date it is first treated. Most players, however, compete under both regimes within a few weeks, so treatment is not absorbing at the player level, and I study their response by comparing the same player across the two regimes.

The control vector $\mathbf{X}_{imgt}$ includes player-rank and opponent-rank at the time of the match, which absorb variation in serve behavior driven by ability and opponent quality. I cluster standard errors at the player level to account for serial correlation in serve behavior across matches and tournaments.

The main identifying assumption is that, conditional on the fixed effects and controls, ELC status is uncorrelated with other determinants of a player’s serve behavior in that tournament. The design does not rely only on before-and-after comparisons within tournaments: it also compares the same player across tournaments with and without ELC.

6 Empirical Results

6.1 Main Results

Table 2 reports the main fixed-effects results.⁶ ELC adoption reduces the ace rate by 0.35 percentage points ($p < 0.05$), a 5% decrease relative to the control mean. The double-fault rate falls by 0.16 percentage points ($p < 0.10$), or 5% of the control mean. The first-serve-in rate rises by 0.44 percentage points ($p < 0.10$), or 1% of the control mean, consistent with the idea that AI evaluation leads to less risk-taking.

⁶Some matches contribute only one player observation because the requirement to play at least 10 matches in a season applies separately to each player. The main regressions exclude 28 observations with missing rankings; the break-point-save regression also excludes 1,220 observations with no break points faced.

As a placebo test, I use the break-point save rate, which the model does not cover. Column 4 of Table 2 re-runs the main specification with this outcome and finds no significant effect. If the effects were solely due to increased accuracy, the first-serve-in rate would fall, and the double-fault rate would rise, the opposite of what I find. Together with the decline in aces and the lack of a statistically significant effect on break-point saving, this pattern is consistent with a behavioral reaction to ELC.

Table 2: Effect of Electronic Line Calling on Player Risk-Taking

	Ace rate	First-serve-in rate	Double-fault rate	Break-point save rate
ELC	-0.0035** (0.0015)	0.0044* (0.0023)	-0.0016* (0.0008)	0.0047 (0.0078)
Control mean	0.0702	0.6288	0.0337	0.5878
Observations	23,594	23,594	23,594	22,374
Player FE	✓	✓	✓	✓
Tournament FE	✓	✓	✓	✓
Year FE	✓	✓	✓	✓

The unit of observation is a player-match. ELC is a dummy variable equal to one if the tournament operated under Electronic Line Calling in that year. The outcome variables are ace rate (Column 1), first-serve-in rate (Column 2), and double-fault rate (Column 3). Column 4 reports the ELC effect on the placebo outcome, break-point save rate, which is dominated by rally dynamics the model does not cover. All specifications include player, tournament, and year fixed effects, and control for time-varying player rank and opponent rank. Standard errors are clustered at the player level. ***, **, and * denote statistical significance at the 1%, 5%, and 10% levels, respectively.

6.2 Robustness

The changes in serve outcomes are consistent with players taking less risk under ELC, but could also reflect changes in who plays or in conditions correlated with ELC adoption, such as surface. I assess these alternative explanations and the robustness of the estimates using the specifications reported in Table 3.

I begin by assessing sensitivity to the controls. The main estimates in column 1 are robust to excluding the controls (column 2) and to replacing the rank controls with the absolute

Table 3: Robustness: Effect of ELC on Player Risk-Taking

Panel A: Ace Rate											
	Baseline	Without Controls	Rank Distance Control	Winsorized Outcomes	Tournament, Player-Year FE	Full Sample	No Clay Tournaments	No Clay, Tournament-Year Clust.	Player, Tournament, Surface-Year FE	Tournament, Year FE	First Round Only
ELC	-0.0035** (0.0015)	-0.0036** (0.0015)	-0.0036** (0.0016)	-0.0031** (0.0014)	-0.0043*** (0.0016)	-0.0040*** (0.0015)	-0.0040** (0.0019)	-0.0040 (0.0031)	-0.0034** (0.0017)	-0.0009 (0.0022)	-0.0042* (0.0023)
Control mean	0.0702	0.0702	0.0702	0.0698	0.0702	0.0689	0.0901	0.0901	0.0702	0.0702	0.0704
Controls	✓	✓	Rank Distance	✓	✓	✓	✓	✓	✓	✓	✓
Clustering	Player	Player	Player	Player	Player	Player	Player	Tournament-Year	Player	Player	Player
Observations	23,594	23,622	23,594	23,594	23,594	25,990	16,364	16,364	23,594	23,594	9,892
Player FE	✓	✓	✓	✓		✓	✓	✓	✓		✓
Tournament FE	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Year FE	✓	✓	✓	✓		✓	✓	✓		✓	✓
Player-Year FE					✓						
Surface-Year FE									✓		

Panel B: First-Serve-In Rate											
	Baseline	Without Controls	Rank Distance Control	Winsorized Outcomes	Tournament, Player-Year FE	Full Sample	No Clay Tournaments	No Clay, Tournament-Year Clust.	Player, Tournament, Surface-Year FE	Tournament, Year FE	First Round Only
ELC	0.0044* (0.0023)	-0.0045* (0.0023)	0.0044* (0.0023)	-0.0042* (0.0023)	0.0043* (0.0023)	0.0042* (0.0022)	0.0083*** (0.0025)	0.0083** (0.0035)	0.0066*** (0.0024)	0.0025 (0.0026)	0.0066** (0.0032)
Control mean	0.6288	0.6288	0.6288	0.6286	0.6288	0.6274	0.6270	0.6270	0.6288	0.6288	0.6230
Controls	✓	✓	Rank Distance	✓	✓	✓	✓	✓	✓	✓	✓
Clustering	Player	Player	Player	Player	Player	Player	Player	Tournament-Year	Player	Player	Player
Observations	23,594	23,622	23,594	23,594	23,594	25,990	16,364	16,364	23,594	23,594	9,892
Player FE	✓	✓	✓	✓		✓	✓	✓	✓		✓
Tournament FE	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Year FE	✓	✓	✓	✓		✓	✓	✓		✓	✓
Player-Year FE					✓						
Surface-Year FE									✓		

Panel C: Double-Fault Rate											
	Baseline	Without Controls	Rank Distance Control	Winsorized Outcomes	Tournament, Player-Year FE	Full Sample	No Clay Tournaments	No Clay, Tournament-Year Clust.	Player, Tournament, Surface-Year FE	Tournament, Year FE	First Round Only
ELC	-0.0016* (0.0008)	-0.0016* (0.0008)	-0.0016* (0.0008)	-0.0015* (0.0008)	-0.0015* (0.0009)	-0.0012 (0.0008)	-0.0021** (0.0010)	-0.0021* (0.0011)	-0.0011 (0.0010)	-0.0018* (0.0010)	-0.0027** (0.0012)
Control mean	0.0337	0.0337	0.0337	0.0335	0.0337	0.0342	0.0344	0.0344	0.0337	0.0337	0.0353
Controls	✓	✓	Rank Distance	✓	✓	✓	✓	✓	✓	✓	✓
Clustering	Player	Player	Player	Player	Player	Player	Player	Tournament-Year	Player	Player	Player
Observations	23,594	23,622	23,594	23,594	23,594	25,990	16,364	16,364	23,594	23,594	9,892
Player FE	✓	✓	✓	✓		✓	✓	✓	✓		✓
Tournament FE	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Year FE	✓	✓	✓	✓		✓	✓	✓		✓	✓
Player-Year FE					✓						
Surface-Year FE									✓		

The unit of observation is a player-match. ELC is a dummy variable equal to one if the tournament operated under Electronic Line Calling in that year. The table reports robustness checks for the ELC effect on the outcome, separately for each panel. Column 1 reports the main specification, with player rank and opponent rank controls, three-way fixed effects (player, tournament, and year), and standard errors clustered at the player level. Column 2 drops the player rank and opponent rank controls. Column 3 replaces the two rank controls with the absolute distance between the player and opponent rankings. Column 4 winsorizes outcomes at the 1st and 99th percentiles. Column 5 replaces player and year fixed effects with player-by-year fixed effects. Column 6 uses the full sample, including players with fewer than 10 matches in a season (the other columns use active players only). Column 7 excludes clay-court tournaments. Column 8 excludes clay-court tournaments and clusters standard errors at the tournament-year level. Column 9 replaces the common year fixed effects with surface-by-year fixed effects. Column 10 drops player fixed effects, keeping only tournament and year fixed effects and the rank controls. Column 11 restricts the sample to first-round matches. ***, **, and * denote statistical significance at the 1%, 5%, and 10% levels, respectively.

distance between the player’s and the opponent’s rankings (column 3). This distance captures the quality mismatch between the two players in a single measure. To limit the influence of outliers, I winsorize the outcomes at the 1st and 99th percentiles in column 4; the coefficients are similar.

ELC adoption may change which players enter treated tournaments, biasing the comparison of player-match outcomes. The fixed effects absorb stable differences across players and tournaments, but they do not absorb changes in who chooses to compete. If more aggressive players selectively avoid ELC tournaments, or if ELC events attract differently ranked players, the estimated effect might partly reflect changes in who plays rather than how players respond to ELC. For example, in professional golf, the presence of a superstar reduces the performance of other players (Brown 2011). I test for selection in two ways.

First, I estimate a linear probability model of tournament entry where the outcome equals one if a given player competed in a given tournament-year. I regress this outcome on ELC status with player-by-year and tournament fixed effects. I cluster standard errors at the player and tournament levels. I restrict the sample to players who entered at least five tournaments in a given year, since active players are more likely to skip a tournament by choice than because of injury or retirement. The coefficient in Table 4 is small and statistically insignificant, indicating that ELC has no effect on the probability that a given player enters a given tournament.

Second, I test whether ELC shifts the composition of entrants by regressing tournament-level entrant statistics (mean and median rank, the top-10 and top-50 shares, mean age, and the number of entrants) on ELC status with tournament and year fixed effects. None of the

coefficients in Table 5 are significant. Together, these two tests help mitigate concerns about selection bias.

Table 4: Selection: Probability of Tournament Entry

	=1 if participated in the tournament
ELC	-0.0190 (0.0135)
Observations	47,471
Player-Year FE	✓
Tournament FE	✓

The unit of observation is a player-tournament-year. The dependent variable is a dummy equal to one if the player participated in the tournament in that season. ELC is a dummy variable equal to one if the tournament operated under Electronic Line Calling in that year. The specification includes player-by-year and tournament fixed effects. Standard errors are clustered at the player and tournament levels. The sample includes all players, including those with fewer than 10 matches in a season. ***, **, and * denote statistical significance at the 1%, 5%, and 10% levels, respectively.

Table 5: Selection: Entrant Composition

	Mean rank	Median rank	Top-10 share	Top-50 share	Mean age	Number of entrants
ELC	0.1953 (5.4396)	4.2204 (2.6166)	-0.0014 (0.0085)	0.0075 (0.0254)	0.1402 (0.1923)	-2.8697 (2.0835)
Mean	91.439	67.984	0.069	0.377	27.125	42.169
SD	34.462	24.450	0.123	0.195	0.775	28.606
Observations	296	296	296	296	296	296
Tournament FE	✓	✓	✓	✓	✓	✓
Year FE	✓	✓	✓	✓	✓	✓

The unit of observation is a tournament-year. ELC is a dummy variable equal to one if the tournament operated under Electronic Line Calling in that year. Each column reports a regression of an entrant composition summary on the ELC dummy. All specifications include tournament and year fixed effects. Standard errors are clustered at the tournament level. The sample includes all players, including those with fewer than 10 matches in a season. ***, **, and * denote statistical significance at the 1%, 5%, and 10% levels, respectively.

Because ELC status varies across tournaments within a year, a player typically competes at both ELC and non-ELC events in the same year. In column 5 of Table 3, I replace the player and year fixed effects with player-by-year fixed effects. This identifies the effect from the same player serving under the two evaluation regimes in the same year, and the estimates are

similar to the baseline. In column 6, I use the full sample, including players with fewer than 10 matches in a season. The ace-rate and first-serve-in estimates are similar to the baseline, while the double-fault estimate is slightly smaller and no longer statistically significant.

Column 10 partially addresses player selection. If more aggressive players disproportionately played at ELC tournaments, dropping player fixed effects should affect the estimate: the comparison would then also draw on across-player differences in aggressiveness, not just within-player changes. The double-fault coefficient is stable when player fixed effects are dropped.

ELC adoption is also correlated with surface. Clay tournaments adopted ELC only in 2025 or not at all. This generates a potential confound: players may adjust serving strategy differently by surface for reasons unrelated to ELC. The tournament fixed effects absorb the time-invariant surface effect. Serve aggressiveness might still evolve differently by surface. For example, if players become progressively more aggressive on clay relative to hard courts over the sample period, the tournament fixed effects would not absorb these differential trends.

I assess this concern in two ways. In column 7 of Table 3, I exclude clay-court tournaments, removing the never-treated clay Slam and the late-adopting clay events. The effect is then identified from hard- and grass-court tournaments alone. In column 8, I exclude clay-court tournaments and cluster standard errors at the tournament-year level. The first-serve-in and double-fault estimates remain similar to the main specification. In column 9, I add surface-by-year fixed effects and identify the effect only from tournaments that differ in ELC status within the same surface and year. The effects are stable in columns 7 and 9.

Column 11 addresses a different threat: a player who serves better because of ELC may win

more, advance further, and thus face stronger opponents later in the event, purely because of their own success. Opponent rank may not fully capture this selection, so I restrict the sample to first-round matches, where opponents are determined by the draw before the event begins and therefore cannot depend on a player's performance during the tournament. The estimates are, if anything, larger in this subsample.

The pooled ELC effect could also be driven by a small number of tournaments. To check, I re-estimate the baseline specification separately within each tournament with at least one pre-ELC and one post-ELC year. I replace the year fixed effects with a linear time trend, keep player fixed effects and player and opponent rank controls, and cluster at the player level. Figure A.3 displays the tournament-specific ELC coefficients for the ace rate, with the pooled estimate shown as a dotted vertical line. Although individual estimates are noisy because each regression uses only matches from one tournament, the pattern suggests that the response is not limited to a few events. The analogous plots for the double-fault rate and first-serve-in rate (Figure A.4 and Figure A.5) show a similar pattern.

Similarly, a handful of players with many match observations could drive the result. I therefore re-estimate the baseline separately for each of the 30 players with the most observations, including tournament and year fixed effects. The player-specific estimates for ace rates (Figure A.6), double-fault rates (Figure A.7), and first-serve-in rates (Figure A.8) suggest that the response is spread across players rather than driven by a few aggressive servers.

ELC may also change the match itself, not just how players serve. For example, it removes challenges and disputes over line calls, which could shorten matches. I estimate the effect of ELC on match duration and the total number of serve points. The unit of observation is the

match, so I include fixed effects for each of the two players and cluster standard errors at the tournament level. Table A.1 shows no significant change in the match length.

As an additional placebo check, I assign each tournament that appears in both 2014–2018 and 2021–2025 the ELC status it had in 2021–2025, shifted back seven years, and re-estimate the main specification for active players in 2014–2018, when ELC did not exist. None of the coefficients are statistically significant (Table A.2).

I also vary the definition of active players, using cutoffs from at least 5 to at least 20 matches per season, and show that the estimates are similar across definitions (Table A.3).

Two-way fixed-effects estimates can be biased under staggered adoption if the effect differs across adoption cohorts. I therefore estimate a stacked difference-in-differences on 2018–2025 that compares each adoption cohort only with tournaments not yet treated. The estimates have the same sign as the baseline and the first-serve-in and double-fault estimates are significant (Figure A.9).

6.3 Mechanism: Serve Direction

The baseline framework concerns how much risk a worker takes. Changes in evaluation can also affect which strategy a worker chooses when evaluation errors differ across strategies. In tennis, risk corresponds to how aggressively players serve, and strategy to where they direct their serves. I ask whether ELC changes where players aim. More cautious serving might translate into fewer serves aimed at the lines (down the T or wide) and more at the body.

Calling errors may also affect this choice: human judges can call an in serve out or an out serve in. Because these errors may vary by target, removing them can shift where players

choose to serve.

To capture this mechanism, I extend the baseline model by letting the server choose between a line target (either down the T or wide) and a body target directed toward the receiver, while holding aggressiveness fixed. Figure 3 shows these directions and the boundaries of the service box. Because changing direction can affect how close a serve lands to both the sideline and the service line, I let the probability of landing in differ across targets.

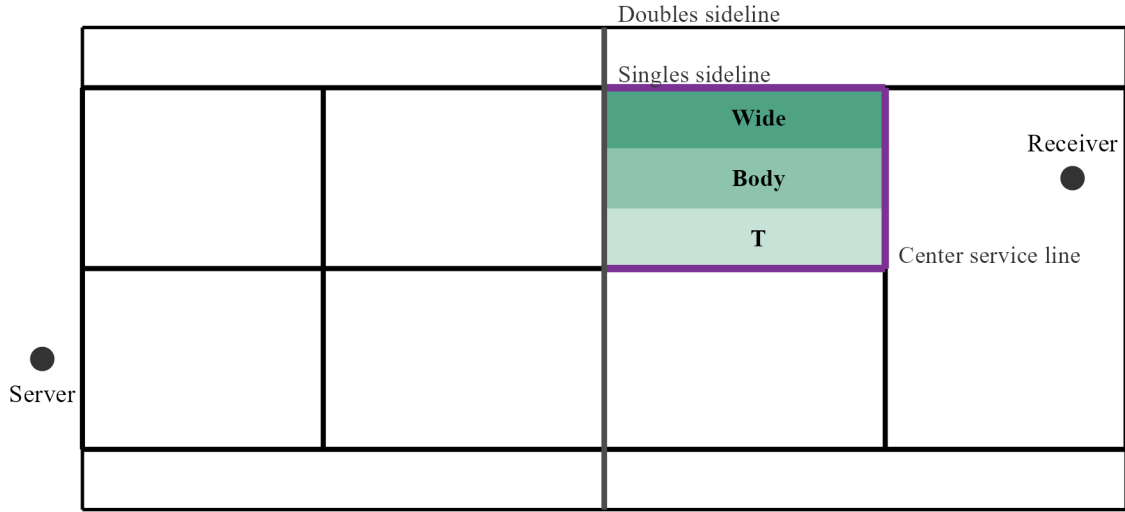


Figure 3: Service Box and Serve Direction. The service box the serve is aimed at is shaded and split into three corridors marking the serve-direction targets: T, Body, and Wide. The purple lines mark the box’s outer sideline, center line, and service line.

For each target $\tau \in \{\text{line}, \text{body}\}$, let p_τ be the probability that the serve truly lands in and q_τ the probability of winning conditional on the serve being called in. As in the baseline model, the player chooses the target that gives the higher probability of winning the point.

I allow both types of human error to differ across targets. Let α_τ be the probability that a truly in serve is called out and β_τ the probability that a truly out serve is called in.

Under AI evaluation, the value of target τ is

$$V_{AI}(\tau) = p_{\tau}q_{\tau}.$$

Under human evaluation, some in serves are called out and some out serves are called in, giving

$$V_H(\tau) = \left[(1 - \alpha_{\tau})p_{\tau} + \beta_{\tau}(1 - p_{\tau}) \right] q_{\tau}.$$

The change in the value of the target when ELC replaces human judges is therefore

$$\Delta_{\tau} \equiv V_{AI}(\tau) - V_H(\tau) = \left[\alpha_{\tau}p_{\tau} - \beta_{\tau}(1 - p_{\tau}) \right] q_{\tau}.$$

The first term captures the benefit from no longer having in serves called out, while the second captures the loss of favorable calls on out serves. The balance depends on both error rates and on how often serves aimed at that target actually land in.

ELC makes line targets more attractive relative to body targets when $\Delta_{\text{line}} > \Delta_{\text{body}}$, meaning that removing calling errors benefits line targets more. If this difference is large enough, players who previously preferred body serves switch to serving toward the lines. Players can therefore serve less aggressively while directing more serves toward the lines, although whether they do so depends on the calling errors associated with each target.

Whether players shift toward line or body targets is an empirical question, which I examine using serve-direction data from the Match Charting Project merged with the ATP panel. I estimate the effect of ELC on the share of line-targeted and body-targeted serves, separately

for first and second serves.

Figure 4 shows that ELC increases the share of line-targeted first serves (down the T or wide) by 0.75 percentage points ($p < 0.10$), and decreases the share of body first serves by the same amount (since each serve is targeted either at the lines or at the body). The extension offers one possible explanation for this shift: removing calling errors may benefit line targets more than body targets. Together with the main results, this shows that players adjust both how aggressively they serve and where they direct their serves.

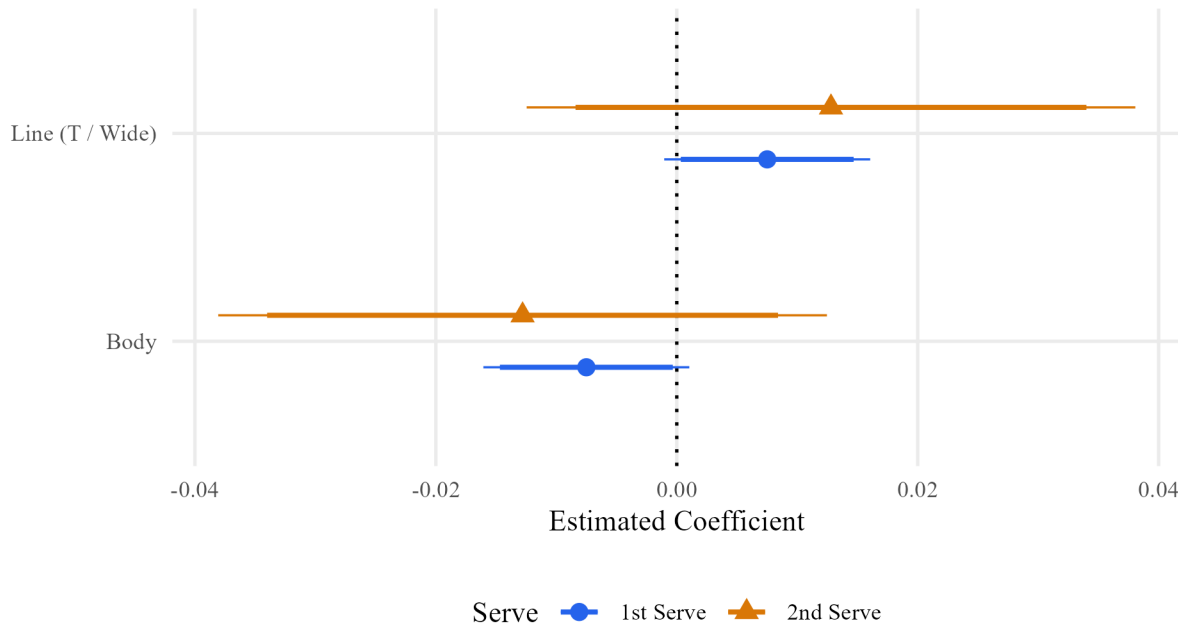


Figure 4: Effect of ELC on Serve Direction. The sample is the Match Charting Project subsample matched to the ATP panel (2,308 matches, 199 players, 87 tournaments); the unit of observation is a player-match. Outcomes are the shares of line (down the T or wide) and body serves, separately for first and second serves. Without ELC, 92.1% of first serves and 68.0% of second serves are aimed at the lines. Each estimate comes from a regression with player, tournament, and year fixed effects and player and opponent rank controls. Thick bars show 90% and thin bars show 95% confidence intervals based on standard errors clustered at the player level.

6.4 Heterogeneity Analysis

The pooled result could be driven by lower-prestige events, where stakes are lower, and players are, on average, less experienced, so their responses might differ from those of highly ranked players. To test this, Figure A.10 (estimates in Table A.4) plots treatment effects separately for high-prestige tournaments (Grand Slams, the ATP Finals, and Masters 1000 events) and lower-prestige tournaments (ATP 500 and 250 events).

The effects are similar across the two groups, with a negative and significant response in ace rate in both high-prestige and lower-prestige tournaments. The double-fault rate and first-serve-in rate show similar patterns across groups, as the model predicts when prestige scales all payoffs by a common factor.

To see whether the response to ELC varies with pressure within a tournament, I compare early rounds with the quarterfinals, semifinals, and final, when players are closer to the title and face greater attention. The model motivates this comparison: an increase in the relative cost of a double fault raises the value of human leniency. Because the rank gap also narrows across rounds, and skill disparities might affect performance on their own (Brox and Goller 2026), I control for player and opponent rank. Figure A.11 presents estimates by tournament round, with the corresponding coefficients in Table A.5.

In both early and later rounds, the ace rate falls, suggesting that the response is present throughout the tournament. For the double-fault rate, the coefficient is small and insignificant in early rounds but larger and significant in quarterfinals, semifinals, and finals. Similarly, the first-serve-in rate also responds more positively in later rounds. These patterns are consistent with a stronger behavioral response as players advance and pressure rises, while effects vary

little across tournament prestige.⁷

How lenient line judges were before ELC may also shape the response. I use the pre-ELC double-fault rate as a proxy for leniency, classifying tournaments by players' average double-fault rate in 2018–2019, before any tournament in the sample adopted the technology. I split tournaments at the median into higher-risk and lower-risk groups. Figure A.12 shows the estimates, reported in Table A.6.

The ace-rate and first-serve-in responses are concentrated in tournaments with a low pre-ELC double-fault rate, while the double-fault responses are similar in the two groups, although this pattern has more than one possible explanation. In the model, the observed double-fault rate equals $(1 - p_V)(1 - \beta)$, so a low rate can reflect either lenient judges (high β) or players who were already serving safely (high p_V). The median split cannot distinguish between these explanations.⁸ Under the leniency interpretation, tournaments recorded fewer double faults because line judges called marginal serves in, giving players a larger subsidy under human evaluation and therefore more to lose when ELC was introduced.

Figure A.13 reports ELC effects by court surface, with the corresponding estimates in Table A.7. Since ELC adoption is correlated with court surface, I interpret this comparison as descriptive evidence. The ace-rate response is similar on hard and clay courts, while there is no effect on grass. The first-serve-in and double-fault responses are statistically significant only on hard courts, which, like grass, favor the server. This is consistent with the model's prediction that the subsidy provided by human leniency is worth more where the serve is

⁷I interpret these results with caution. The early- and late-round coefficients are estimated in a single interacted specification. The two coefficients differ in statistical significance, but a test of their equality does not reject the null.

⁸The two channels offset each other: leniency lowers observed faults directly but also induces riskier serving (higher a^*), which raises the number of truly out serves.

more decisive and $q'(a)$ is larger.

I also estimate the main specification for female professionals. The results do not replicate in this sample: the ELC coefficients are small and statistically insignificant for all outcomes (Table A.8). I discuss four possible explanations. First, the serve is a less decisive shot on the WTA tour: first serves are slower, so the marginal gain from an aggressive serve $q'(a)$ is small. Second, and closely related, slower serves are easier for human judges to call correctly, so the favorable-error rate β is plausibly smaller than in the ATP sample. Because the ELC effect works entirely through removing this leniency, a small β predicts a negligible response. Third, Gauriot, Page, and Wooders (2023) find that male professionals are harder for opponents to predict on serve than female professionals. Their result points to differences in strategic responsiveness, so the null results may instead reflect weaker adjustment of serving strategy to the changed incentives. Fourth, gender differences in risk-taking may play a role (Croson and Gneezy 2009; Charness and Gneezy 2012). The data cannot separate the technological channel (a smaller leniency subsidy) from the expertise channel (weaker adjustment) or the risk channel.

Line judges may feel crowd pressure to favor home players. I test this by splitting the ELC effect between home players, whose nationality matches the host country, and all other players (Table A.9). The ace-rate and first-serve-in effects do not differ significantly between the two groups. The only difference is in double faults: they fall under ELC for away players but not for home players, although the gap is only marginally significant ($p = 0.08$). This could reflect crowd pressure leading human judges to call some in away serves out, which ELC no longer does.

7 Conclusion

Individuals adapt their risk-taking to how leniently they are judged. I develop a model in which human evaluators encourage risk-taking by allowing some mistakes to go unpenalized. Removing this leniency through accurate automated evaluation leads individuals to choose safer strategies. I test this framework using the staggered adoption of Electronic Line Calling in professional tennis, and find that replacing human line judges with AI reduces player aggressiveness: the ace rate and double-fault rate decrease, while the first-serve-in rate increases. The effects are similar across tournament prestige, while the estimates suggest a stronger response in high-pressure rounds.

Players also adjust serve direction, aiming more first serves toward the lines. These changes in serve outcomes align with the framework, suggesting that the behavioral response more than offset the mechanical rise in recorded faults when previously forgiven serves are called out. The extension to serve direction offers one explanation for how less aggressive serving can coexist with a shift toward the lines.

Many non-ELC tournaments in my sample operated under the Hawk-Eye challenge system, which already corrected some human errors; clay tournaments instead relied on ball-mark review. In the model, the comparison therefore captures the removal of the remaining leniency. Hence, the estimates likely understate the full effect of moving from human to AI evaluation in absolute terms.

The analysis has two main limitations. First, match-level averages may hide stronger responses at specific high-pressure points, such as break points. Point-level data, including serve speed

and exact placement, would help identify the mechanism more precisely. Second, professional tennis players are highly skilled and face clear well-understood incentives. The response may therefore differ for lower-ability workers, in settings with ambiguous evaluation criteria, or where evaluation is less salient.

The setting also allows me to observe professionals making decisions in their own competitive environment. Field behavior may differ from choices elicited in the lab (Levitt, List, and Reiley 2010). Outcomes are measured objectively at high frequency across many observations, and the treatment is large and externally imposed, reducing concerns about selection into treatment.

The results extend beyond tennis. In this paper, I show that improving the evaluation accuracy changes workers' incentives to take risks. Field evidence on monitoring technology shows an increase in effort and a reduction in risk-taking among bus drivers (Kelley, Lane, and Schönholzer 2024), a reduction in opportunistic behavior among call-center employees (Nagin et al. 2002), and lower risk-taking when risk is costly, such as banks holding safer portfolios under closer supervision (Hirtle, Kovner, and Plosser 2020). At the same time, monitoring can also signal distrust and crowd out the effort it was intended to raise (Falk and Kosfeld 2006).

The subjective performance evaluation literature shows how discretion shapes these incentives (Baker, Gibbons, and Murphy 1994; Prendergast 2002; MacLeod 2003). In tennis, players adjust their serving when AI replaces human line judges, showing that their response depends directly on the leniency they previously received.

My results suggest another potential cost of monitoring: removing an evaluator's leniency may

lower workers' willingness to take productive risks. Fewer failures can come at the expense of fewer high-reward outcomes, highlighting a key trade-off when evaluating the gains from more accurate monitoring. More broadly, MacLeod (2026) argues that better AI predictions do not necessarily lead to better decisions, since decisions also depend on how people weigh the costs of different errors. For example, two patients with the same condition might have different risk preferences, and the AI assigns them the same treatment. A human evaluator can instead adjust the decision to each patient's preferences. Just as the AI does not weigh risk preferences for patients, it does not accommodate productive risk-taking for workers. Organizations adopting AI evaluation must therefore consider the evaluated side: in domains where high-risk strategies are socially valuable, such as innovation and entrepreneurship, more precise AI evaluation could induce excessive caution.

The results suggest two directions for future work. First, study how workers and evaluators adjust together when AI is introduced, particularly when AI assists human evaluators who retain discretion over the final decision. A related question is whether organizations can design evaluation systems with a small, calibrated degree of leniency that preserves valuable risk-taking without reintroducing human bias. Second, examine how responses to AI evaluation vary across workplaces where workers' ability, the stakes, and the returns to risk-taking differ. The response may differ in domains where risk-taking generates long-term value that is difficult to assess in the short term, such as innovation, entrepreneurship, or scientific research.

References

- Abramitzky, Ran, Liran Einav, Shimon Kolkowitz, and Roy Mill, “On the Optimality of Line Call Challenges in Professional Tennis,” *International Economic Review*, 53 (2012), 939–964.
- Agarwal, Nikhil, Alex Moehring, Pranav Rajpurkar, and Tobias Salz, “Combining Human Expertise with Artificial Intelligence: Experimental Evidence from Radiology,” Working Paper, 31422, 2023 (National Bureau of Economic Research).
- Almog, David, Romain Gauriot, Lionel Page, and Daniel Martin, “AI Oversight and Human Mistakes: Evidence from Centre Court,” arXiv:2401.16754, 2025 (arXiv).
- Almog, David, Lucas Lippman, and Daniel Martin, “When an AI Judges Your Work: The Hidden Costs of Algorithmic Assessment,” arXiv:2603.02076, 2026 (arXiv).
- Anderson, Axel, Jeremy Rosen, John Rust, and Kin-Ping Wong, “Disequilibrium Play in Tennis,” *Journal of Political Economy*, 133 (2025), 190–251.
- Angelova, Victoria, Will Dobbie, and Crystal S Yang, “Algorithmic Recommendations and Human Discretion,” *The Review of Economic Studies*, (2025).
- Baker, George, Robert Gibbons, and Kevin J. Murphy, “Subjective Performance Measures in Optimal Incentive Contracts,” *The Quarterly Journal of Economics*, 109 (1994), 1125–1156.
- Borusyak, Kirill, Xavier Jaravel, and Jann Spiess, “Revisiting Event-Study Designs: Robust and Efficient Estimation,” *Review of Economic Studies*, 91 (2024), 3253–3285.
- Brown, Jennifer, “Quitters Never Win: The (Adverse) Incentive Effects of Competing with Superstars,” *Journal of Political Economy*, 119 (2011), 982–1013.
- Brox, Enzo, and Daniel Goller, “Tournaments, Contestant Heterogeneity, and Performance,” *Journal of Political Economy Microeconomics*, 4 (2026), 583–619.

Burke, Marshall, Vincent Tanutama, Sam Heft-Neal, Miyuki Hino, and David Lobell, “Game, Sweat, Match: Temperature and Elite Worker Productivity,” Working Paper, 31650, 2023 (National Bureau of Economic Research).

Callaway, Brantly, and Pedro H. C. Sant’Anna, “Difference-in-Differences with multiple time periods,” *Journal of Econometrics*, 225 (2021), 200–230.

Charness, Gary, and Uri Gneezy, “Strong Evidence for Gender Differences in Risk Taking,” *Journal of Economic Behavior & Organization*, 83 (2012), 50–58.

Cho, Sungwoo, “The Effect of Robot Assistance on Skills,” SSRN Scholarly Paper, 4902149, 2024 (Social Science Research Network).

Cohen-Zada, Danny, Alex Krumer, Mosi Rosenboim, and Offer Moshe Shapir, “Choking under pressure and gender: Evidence from professional tennis,” *Journal of Economic Psychology*, 61 (2017), 176–190.

Cohen-Zada, Danny, Alex Krumer, and Offer Moshe Shapir, “Testing the effect of serve order in tennis tiebreak,” *Journal of Economic Behavior & Organization*, 146 (2018), 106–115.

Croson, Rachel, and Uri Gneezy, “Gender Differences in Preferences,” *Journal of Economic Literature*, 47 (2009), 448–474.

Ely, Jeffrey, Romain Gauriot, and Lionel Page, “Do agents maximise? Risk taking on first and second serves in tennis,” *Journal of Economic Psychology*, 63 (2017), 135–142.

Falk, Armin, and Michael Kosfeld, “The Hidden Costs of Control,” *American Economic Review*, 96 (2006), 1611–1630.

Gauriot, Romain, and Lionel Page, “Does Success Breed Success? A Quasi-Experiment on Strategic Momentum in Dynamic Contests,” *The Economic Journal*, 129 (2019), 3107–3136.

Gauriot, Romain, Lionel Page, and John Wooders, “Expertise, gender, and equilibrium play,”

- Quantitative Economics*, 14 (2023), 981–1020.
- Genakos, Christos, and Mario Pagliero, “Interim Rank, Risk Taking, and Performance in Dynamic Tournaments,” *Journal of Political Economy*, 120 (2012), 782–813.
- Goergen, Jonas, Emanuel de Bellis, and Anne-Kathrin Klesse, “AI assessment changes human behavior,” *Proceedings of the National Academy of Sciences*, 122 (2025), e2425439122.
- Gucciardi, Gianluca, and Massimo Ruberti, “Individual performance and environment: Home advantage in ATP tennis,” *Journal of Economic Behavior & Organization*, 238 (2025), 107236.
- Hirtle, Beverly, Anna Kovner, and Matthew Plosser, “The Impact of Supervision on Bank Performance,” *The Journal of Finance*, 75 (2020), 2765–2808.
- Hoffman, Mitchell, Lisa B Kahn, and Danielle Li, “Discretion in Hiring,” *The Quarterly Journal of Economics*, 133 (2018), 765–800.
- Jabarian, Brian, and Luca Henkel, “Voice AI in Firms: A Natural Field Experiment on Automated Job Interviews,” SSRN Scholarly Paper, 5395709, 2025 (Social Science Research Network).
- Kelley, Erin M., Gregory Lane, and David Schönholzer, “Monitoring in Small Firms: Experimental Evidence from Kenyan Public Transit,” *American Economic Review*, 114 (2024), 3119–3160.
- Kleinberg, Jon, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan, “Human Decisions and Machine Predictions,” *The Quarterly Journal of Economics*, 133 (2018), 237–293.
- Krumer, Alex, and Michael Lechner, “First in first win: Evidence on schedule effects in round-robin tournaments in mega-events,” *European Economic Review*, 100 (2017), 412–427.
- Levitt, Steven D., John A. List, and David H. Reiley, “What Happens in the Field Stays

- in the Field: Exploring Whether Professionals Play Minimax in Laboratory Experiments,” *Econometrica*, 78 (2010), 1413–1434.
- MacLeod, W. Bentley, “Optimal Contracting with Subjective Evaluation,” *American Economic Review*, 93 (2003), 216–240.
- , “The Economics of Professional Decision-Making: Can Artificial Intelligence Reduce Decision Uncertainty?” *Annual Review of Economics*, (2026).
- Nagin, Daniel S., James B. Rebitzer, Seth Sanders, and Lowell J. Taylor, “Monitoring, Motivation, and Management: The Determinants of Opportunistic Behavior in a Field Experiment,” *American Economic Review*, 92 (2002), 850–873.
- Palacios-Huerta, Ignacio, “Professionals Play Minimax,” *The Review of Economic Studies*, 70 (2003), 395–415.
- , “The Beautiful Dataset,” *Journal of Economic Literature*, 63 (2025), 1363–1423.
- Paserman, M. Daniele, “Gender Differences in Performance in Competitive Environments? Evidence from Professional Tennis Players,” *Journal of Economic Behavior & Organization*, 212 (2023), 590–609.
- Prendergast, Canice, “The Tenuous Trade-off between Risk and Incentives,” *Journal of Political Economy*, 110 (2002), 1071–1102.
- Prendergast, Canice, and Robert H. Topel, “Favoritism in Organizations,” *Journal of Political Economy*, 104 (1996), 958–978.
- Sun, Liyang, and Sarah Abraham, “Estimating dynamic treatment effects in event studies with heterogeneous treatment effects,” *Journal of Econometrics*, 225 (2021), 175–199.
- Walker, Mark, and John Wooders, “Minimax Play at Wimbledon,” *American Economic Review*, 91 (2001), 1521–1538.

Appendix

Tables

Table A.1: Effect of Electronic Line Calling on Match Duration and Length

	Duration (minutes) (1)	Total serve points (2)
ELC	-1.9947 (1.3798)	-2.9874 (1.8819)
Control mean	115.78	161.36
Observations	12,511	12,884
Player 1 FE	✓	✓
Player 2 FE	✓	✓
Tournament FE	✓	✓
Year FE	✓	✓

The unit of observation is a match. ELC is a dummy variable equal to one if the tournament operated under Electronic Line Calling in that year. The outcome variables are the match duration in minutes (Column 1) and the total number of serve points played by both players (Column 2). All specifications include player fixed effects for both players, tournament fixed effects, and year fixed effects, and control for the rank of both players. Standard errors are clustered at the tournament level. ***, **, and * denote statistical significance at the 1%, 5%, and 10% levels, respectively.

Table A.2: Placebo: Pseudo-ELC Assigned to 2014–2018

	Ace rate	First-serve-in rate	Double-fault rate	Break-point save rate
ELC	0.0003 (0.0016)	-0.0018 (0.0023)	0.0004 (0.0008)	-0.0014 (0.0079)
Control mean	0.0767	0.6160	0.0343	0.5887
Observations	19,574	19,574	19,574	18,469
Player FE	✓	✓	✓	✓
Tournament FE	✓	✓	✓	✓
Year FE	✓	✓	✓	✓

The unit of observation is a player-match in 2014–2018. ELC is a placebo dummy variable: each tournament that also appears in 2021–2025 is assigned the ELC status it had in 2021–2025, shifted back by seven years (2021 to 2014, ..., 2025 to 2018). No tournament used Electronic Line Calling in 2014–2018. The sample is restricted to active players with at least 10 matches in a season. The outcome variables are ace rate (Column 1), first-serve-in rate (Column 2), double-fault rate (Column 3), and break-point save rate (Column 4). All specifications include player, tournament, and year fixed effects, and control for time-varying player rank and opponent rank. Standard errors are clustered at the player level. ***, **, and * denote statistical significance at the 1%, 5%, and 10% levels, respectively.

Table A.3: Sensitivity to the Definition of Active Players

Minimum matches	Ace rate	First-serve-in rate	Double-fault rate	Observations	Players
≥ 20	-0.0051*** (0.0017)	0.0045* (0.0024)	-0.0015* (0.0009)	21,346	185
≥ 19	-0.0049*** (0.0017)	0.0048** (0.0024)	-0.0016* (0.0009)	21,516	187
≥ 18	-0.0047*** (0.0017)	0.0047* (0.0024)	-0.0017* (0.0009)	21,696	189
≥ 17	-0.0044*** (0.0016)	0.0044* (0.0024)	-0.0015* (0.0009)	22,086	197
≥ 16	-0.0043*** (0.0016)	0.0045* (0.0024)	-0.0015* (0.0009)	22,198	201
≥ 15	-0.0043*** (0.0016)	0.0042* (0.0023)	-0.0014* (0.0009)	22,438	208
≥ 14	-0.0041** (0.0016)	0.0045* (0.0024)	-0.0015* (0.0009)	22,661	210
≥ 13	-0.0039** (0.0016)	0.0045* (0.0023)	-0.0014* (0.0008)	22,947	219
≥ 12	-0.0037** (0.0016)	0.0039* (0.0023)	-0.0013 (0.0008)	23,151	225
≥ 11	-0.0036** (0.0015)	0.0042* (0.0023)	-0.0014* (0.0008)	23,414	229
≥ 10 (main specification)	-0.0035** (0.0015)	0.0044* (0.0023)	-0.0016* (0.0008)	23,594	233
≥ 9	-0.0036** (0.0015)	0.0043* (0.0023)	-0.0015* (0.0008)	23,854	246
≥ 8	-0.0035** (0.0015)	0.0043* (0.0023)	-0.0015* (0.0008)	24,092	254
≥ 7	-0.0034** (0.0015)	0.0042* (0.0023)	-0.0014 (0.0008)	24,371	266
≥ 6	-0.0037** (0.0015)	0.0042* (0.0022)	-0.0013 (0.0008)	24,611	277
≥ 5	-0.0041*** (0.0015)	0.0043* (0.0022)	-0.0013 (0.0008)	24,901	295

The unit of observation is a player-match. ELC is a dummy variable equal to one if the tournament operated under Electronic Line Calling in that year. Each row re-estimates the main specification on the sample of players with at least the indicated number of matches in a season; the main sample uses at least 10 matches. All specifications include player, tournament, and year fixed effects, and control for time-varying player rank and opponent rank. Each cell reports the coefficient on ELC, with its standard error in parentheses. Standard errors are clustered at the player level. ***, **, and * denote statistical significance at the 1%, 5%, and 10% levels, respectively.

Table A.4: Heterogeneity by Tournament Prestige

	Ace rate	First-serve-in rate	Double-fault rate
ELC $\times$ Prestige = High prestige	-0.0040** (0.0020)	0.0034 (0.0027)	-0.0010 (0.0010)
ELC $\times$ Prestige = Low prestige	-0.0033* (0.0017)	0.0051* (0.0029)	-0.0019* (0.0010)
Observations	23,594	23,594	23,594
Player FE	✓	✓	✓
Year FE	✓	✓	✓
Tournament FE	✓	✓	✓

The unit of observation is a player-match. ELC is a dummy variable equal to one if the tournament operated under Electronic Line Calling in that year. The table reports the ELC effect interacted with tournament prestige tier. High prestige includes Grand Slam, ATP Finals, and Masters 1000 events. Low prestige includes ATP 500 and ATP 250 events. All specifications include player, tournament, and year fixed effects, and control for time-varying player rank and opponent rank. Standard errors are clustered at the player level. ***, **, and * denote statistical significance at the 1%, 5%, and 10% levels, respectively.

Table A.5: Heterogeneity by Tournament Round

	Ace rate	First-serve-in rate	Double-fault rate
ELC $\times$ Round = Other	-0.0035** (0.0016)	0.0038 (0.0024)	-0.0013 (0.0008)
ELC $\times$ Round = QF/SF/F	-0.0034* (0.0019)	0.0065** (0.0029)	-0.0023** (0.0011)
Observations	23,594	23,594	23,594
Player FE	✓	✓	✓
Year FE	✓	✓	✓
Tournament FE	✓	✓	✓

The unit of observation is a player-match. ELC is a dummy variable equal to one if the tournament operated under Electronic Line Calling in that year. The table reports the ELC effect interacted with round group. QF/SF/F denotes late-stage matches (quarterfinals, semifinals, and final), and Other denotes earlier rounds. All specifications include player, tournament, and year fixed effects, a dummy variable for late-round matches, and control for time-varying player rank and opponent rank. Standard errors are clustered at the player level. ***, **, and * denote statistical significance at the 1%, 5%, and 10% levels, respectively.

Table A.6: Heterogeneity by Pre-ELC Double-Fault Risk

	Ace rate	First-serve-in rate	Double-fault rate
ELC $\times$ Pre-ELC Risk = High Risk	-0.0008 (0.0022)	0.0031 (0.0035)	-0.0018 (0.0013)
ELC $\times$ Pre-ELC Risk = Low Risk	-0.0052*** (0.0017)	0.0055** (0.0027)	-0.0017* (0.0010)
Observations	21,072	21,072	21,072
Player FE	✓	✓	✓
Year FE	✓	✓	✓
Tournament FE	✓	✓	✓

The unit of observation is a player-match. ELC is a dummy variable equal to one if the tournament operated under Electronic Line Calling in that year. The table reports the ELC effect interacted with pre-ELC double-fault risk. High Risk tournaments had an above-median average double-fault rate among players in 2018–2019, before any tournament in the sample had adopted ELC; Low Risk tournaments were below the median. All specifications include player, tournament, and year fixed effects, and control for time-varying player rank and opponent rank. Standard errors are clustered at the player level. ***, **, and * denote statistical significance at the 1%, 5%, and 10% levels, respectively.

Table A.7: Heterogeneity by Court Surface

	Ace rate	First-serve-in rate	Double-fault rate
ELC $\times$ Surface = Clay	-0.0044** (0.0020)	-0.0006 (0.0038)	-0.0016 (0.0011)
ELC $\times$ Surface = Grass	-0.0017 (0.0029)	0.0057 (0.0038)	0.0000 (0.0016)
ELC $\times$ Surface = Hard	-0.0036 (0.0022)	0.0086*** (0.0030)	-0.0022* (0.0011)
Observations	23,594	23,594	23,594
Player FE	✓	✓	✓
Year FE	✓	✓	✓
Tournament FE	✓	✓	✓

The unit of observation is a player-match. ELC is a dummy variable equal to one if the tournament operated under Electronic Line Calling in that year. The table reports the ELC effect interacted with the court surface (Grass, Clay, Hard). All specifications include player, tournament, and year fixed effects, and control for time-varying player rank and opponent rank. Standard errors are clustered at the player level. ***, **, and * denote statistical significance at the 1%, 5%, and 10% levels, respectively.

Table A.8: Effect of Electronic Line Calling on Player Risk-Taking (WTA Sample)

	Ace rate	First-serve-in rate	Double-fault rate
ELC	0.0003 (0.0010)	0.0011 (0.0025)	0.0000 (0.0011)
Control mean	0.0392	0.6245	0.0510
Observations	21,050	21,050	21,046
Player FE	✓	✓	✓
Year FE	✓	✓	✓
Tournament FE	✓	✓	✓

The unit of observation is a player-match. ELC is a dummy variable equal to one if the tournament operated under Electronic Line Calling in that year. The outcome variables are ace rate (Column 1), first-serve-in rate (Column 2), and double-fault rate (Column 3) for the WTA sample. All specifications include player, tournament, and year fixed effects, and control for time-varying player rank and opponent rank. Standard errors are clustered at the player level. ***, **, and * denote statistical significance at the 1%, 5%, and 10% levels, respectively.

Table A.9: Heterogeneity by Home Status

	Ace rate	First-serve-in rate	Double-fault rate
ELC $\times$ Home status = Away	-0.0034** (0.0015)	0.0042* (0.0024)	-0.0018** (0.0009)
ELC $\times$ Home status = Home	-0.0045* (0.0026)	0.0063 (0.0040)	0.0002 (0.0013)
Home = Away (p-value)	0.592	0.565	0.075
Observations	23,594	23,594	23,594
Player FE	✓	✓	✓
Tournament FE	✓	✓	✓
Year FE	✓	✓	✓

The unit of observation is a player-match. ELC is a dummy variable equal to one if the tournament operated under Electronic Line Calling in that year. The table reports the ELC effect separately for home players, whose nationality matches the country hosting the tournament, and away players. Host countries are taken from the official ATP Tour results archive. All specifications include player, tournament, and year fixed effects, a home dummy variable, and control for time-varying player rank and opponent rank. Standard errors are clustered at the player level. ***, **, and * denote statistical significance at the 1%, 5%, and 10% levels, respectively.

Figures

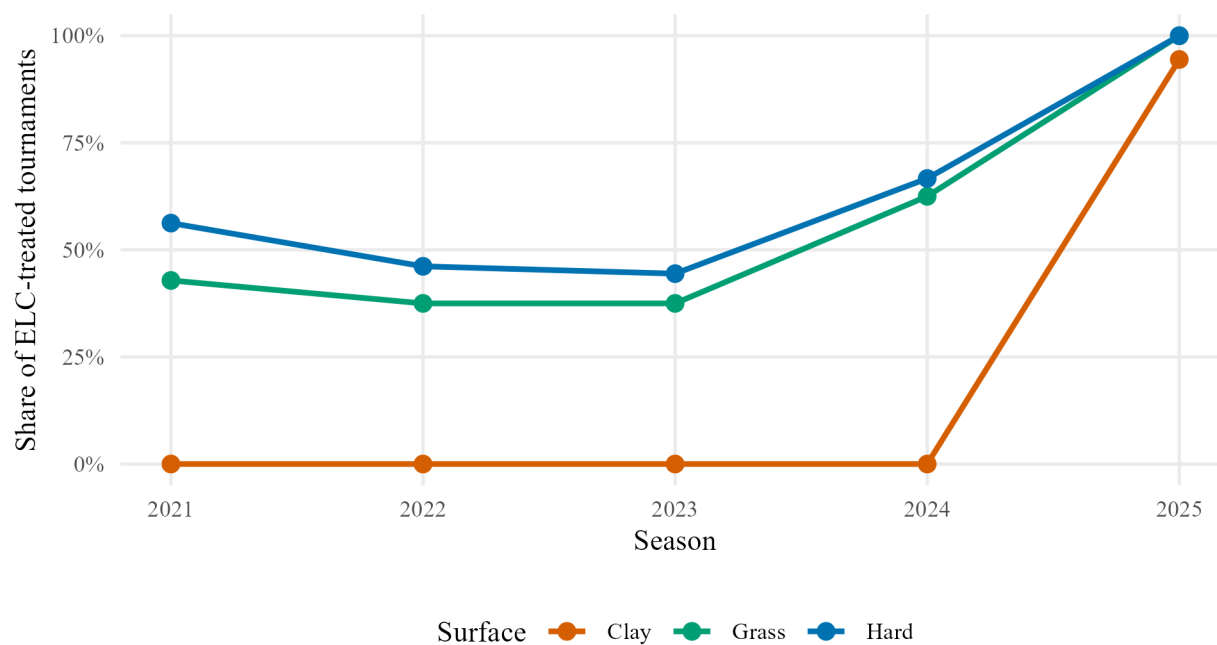


Figure A.1: Staggered Adoption of Electronic Line Calling by Court Surface. Colored lines indicate court surface. ELC adoption is manually verified against tournament announcements, broadcasts, and news reports.

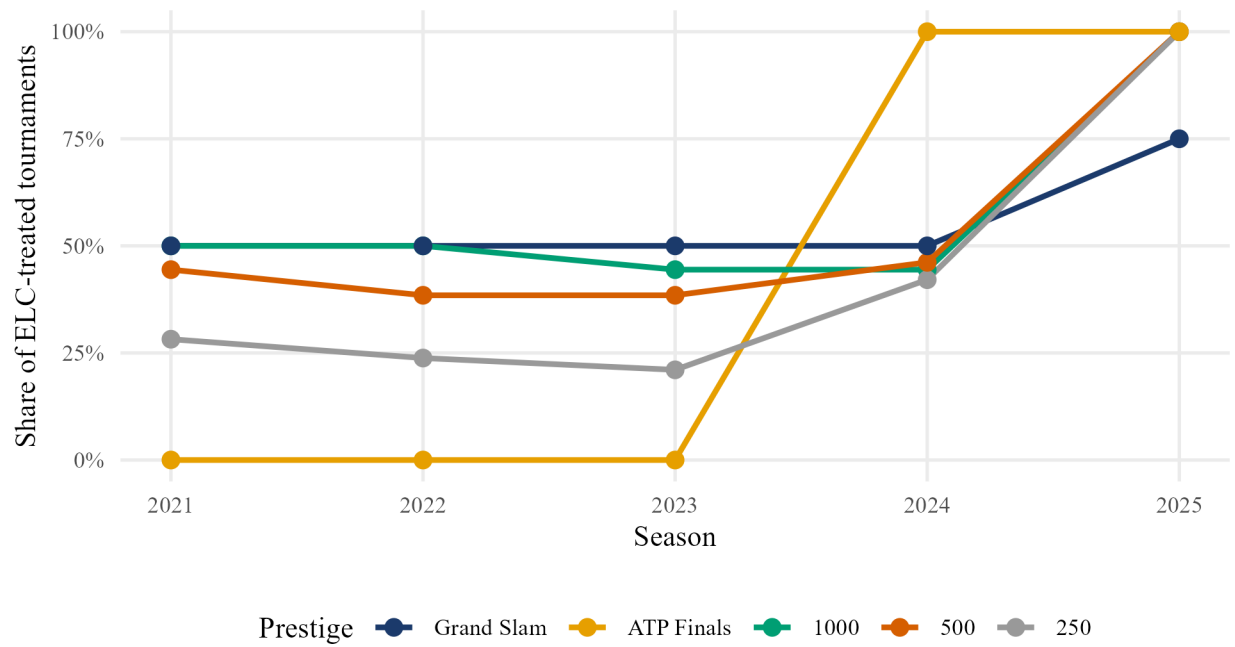


Figure A.2: Staggered Adoption of Electronic Line Calling by Tournament Prestige. Colored lines indicate tournament prestige. ELC adoption is manually verified against tournament announcements, broadcasts, and news reports.

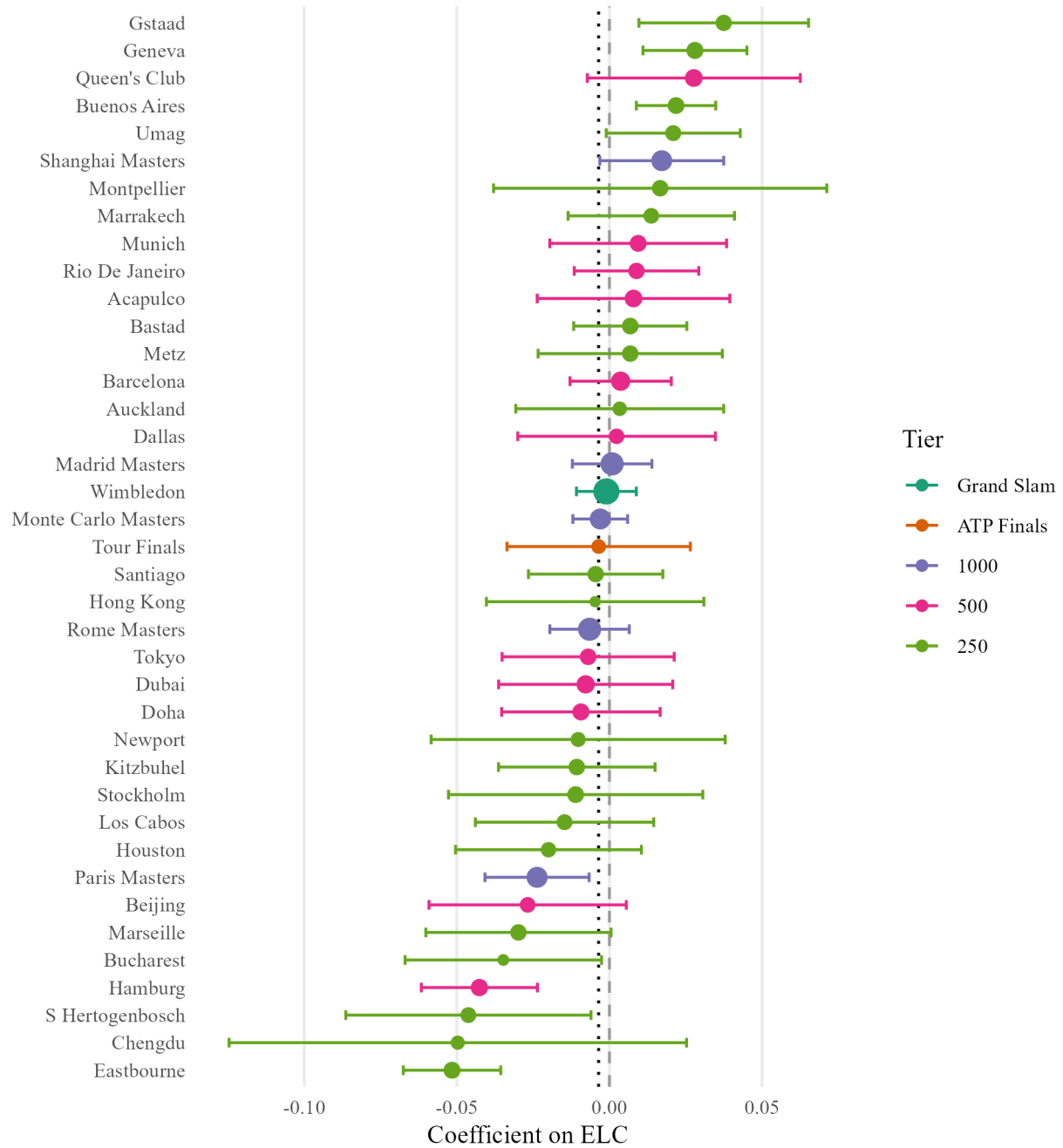


Figure A.3: Tournament-Specific ELC Effects on the Ace Rate. Each estimate comes from a within-tournament fixed-effects regression with player fixed effects, a linear time trend, and player and opponent rank controls. Bars show 95% confidence intervals based on standard errors clustered at the player level. The dotted vertical line marks the pooled estimate from the main specification. Point size is proportional to the number of observations.

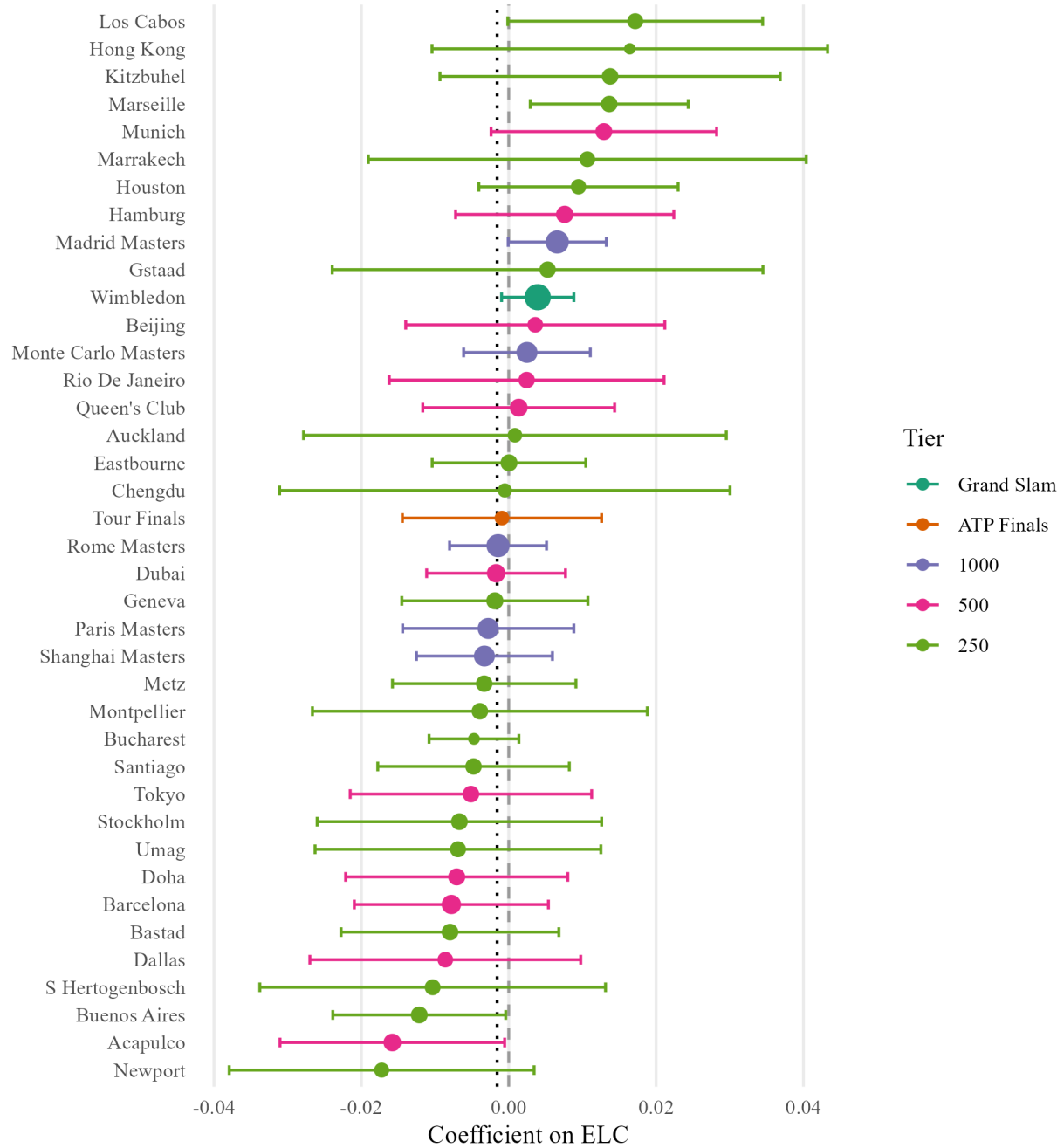


Figure A.4: Tournament-Specific ELC Effects on the Double-Fault Rate. Each estimate comes from a within-tournament fixed-effects regression with player fixed effects, a linear time trend, and player and opponent rank controls. Bars show 95% confidence intervals based on standard errors clustered at the player level. The dotted vertical line marks the pooled estimate from the main specification. Point size is proportional to the number of observations.

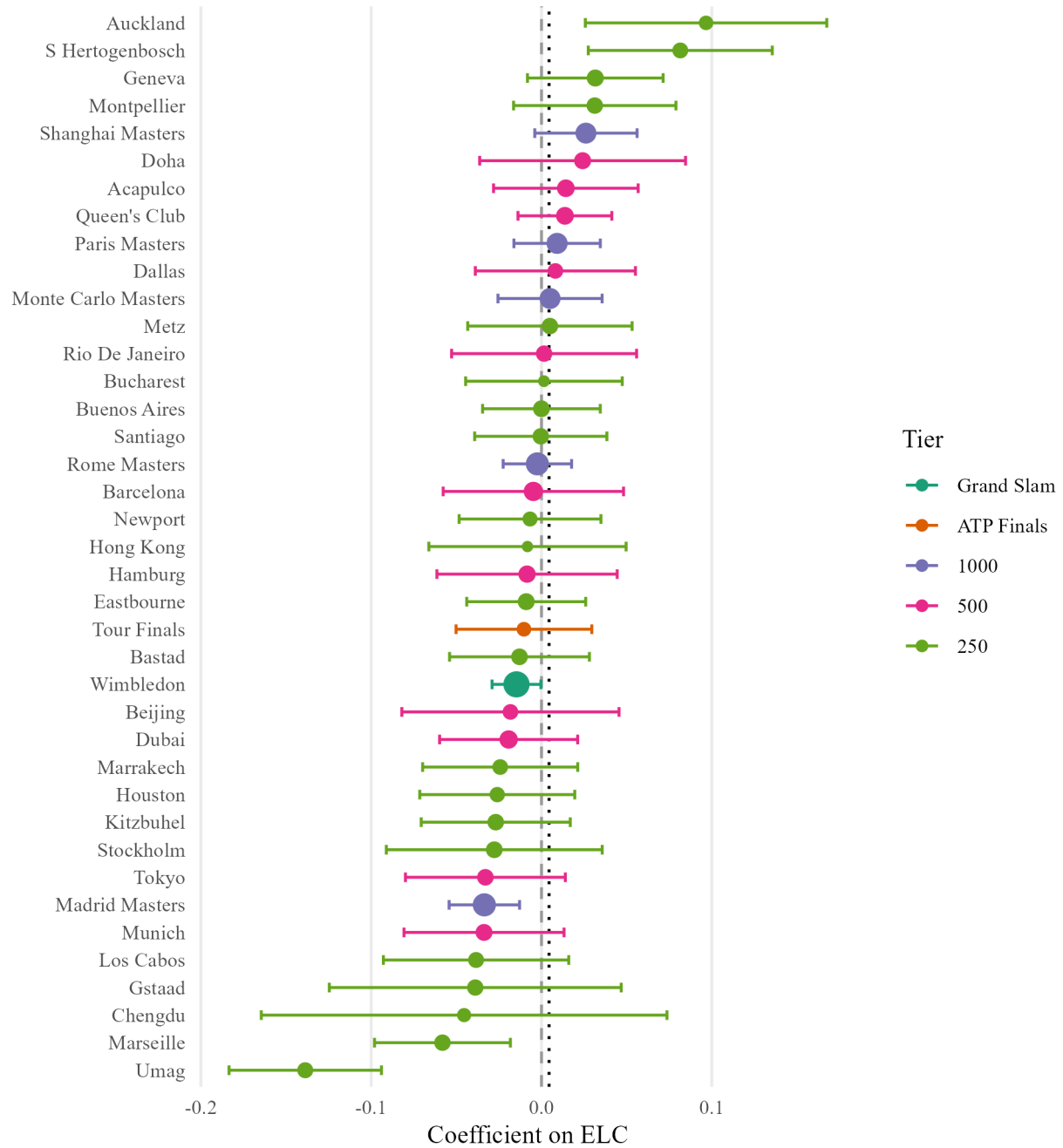


Figure A.5: Tournament-Specific ELC Effects on the First-Serve-In Rate. Each estimate comes from a within-tournament fixed-effects regression with player fixed effects, a linear time trend, and player and opponent rank controls. Bars show 95% confidence intervals based on standard errors clustered at the player level. The dotted vertical line marks the pooled estimate from the main specification. Point size is proportional to the number of observations.

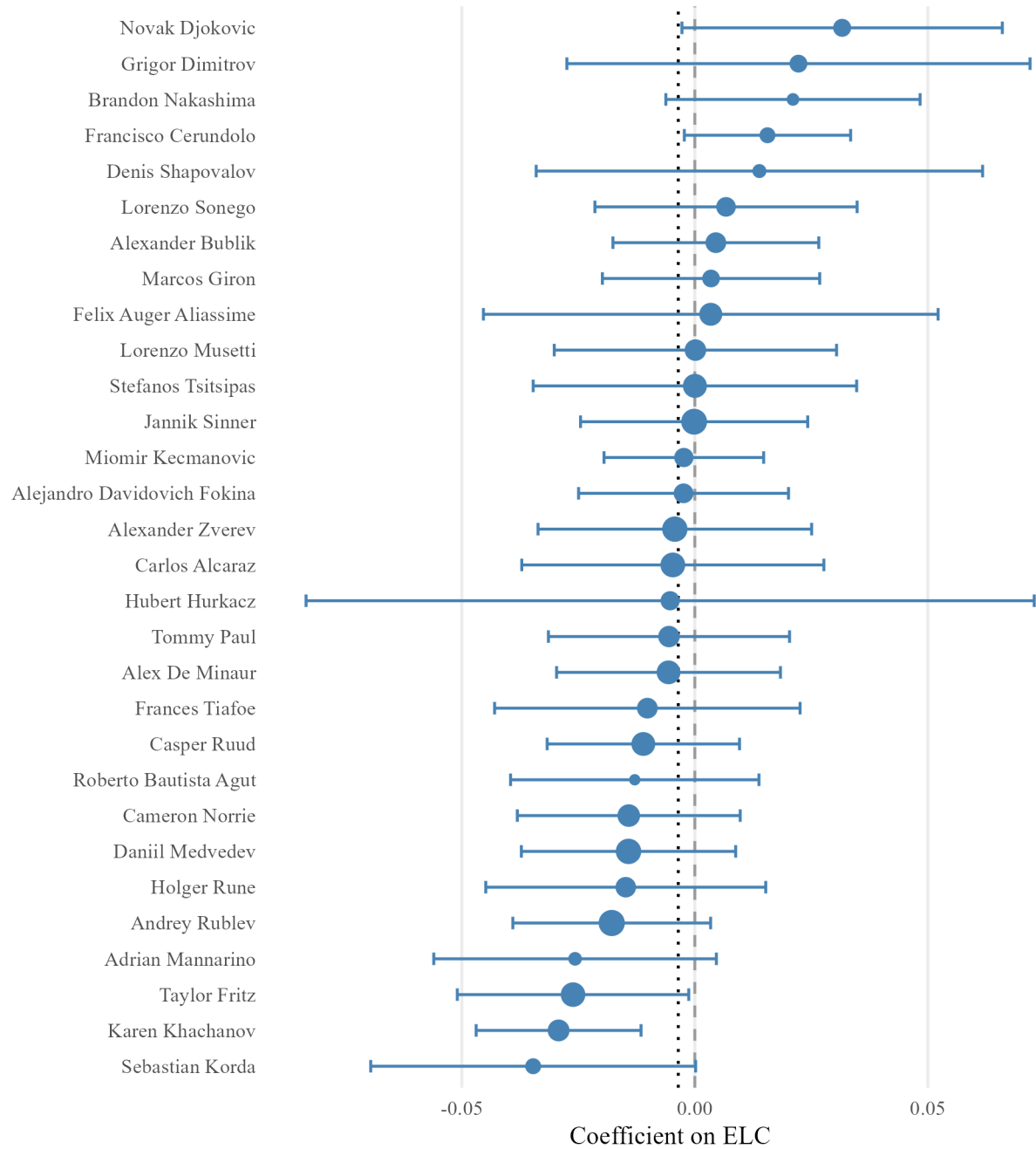


Figure A.6: Player-Specific ELC Effects on the Ace Rate. Each estimate comes from a within-player fixed-effects regression with tournament and year fixed effects and player and opponent rank controls. Bars show 95% confidence intervals based on standard errors clustered at the tournament level. The sample includes the 30 players with the most match observations. The dotted vertical line marks the pooled estimate from the main specification. Point size is proportional to the player's number of match observations.

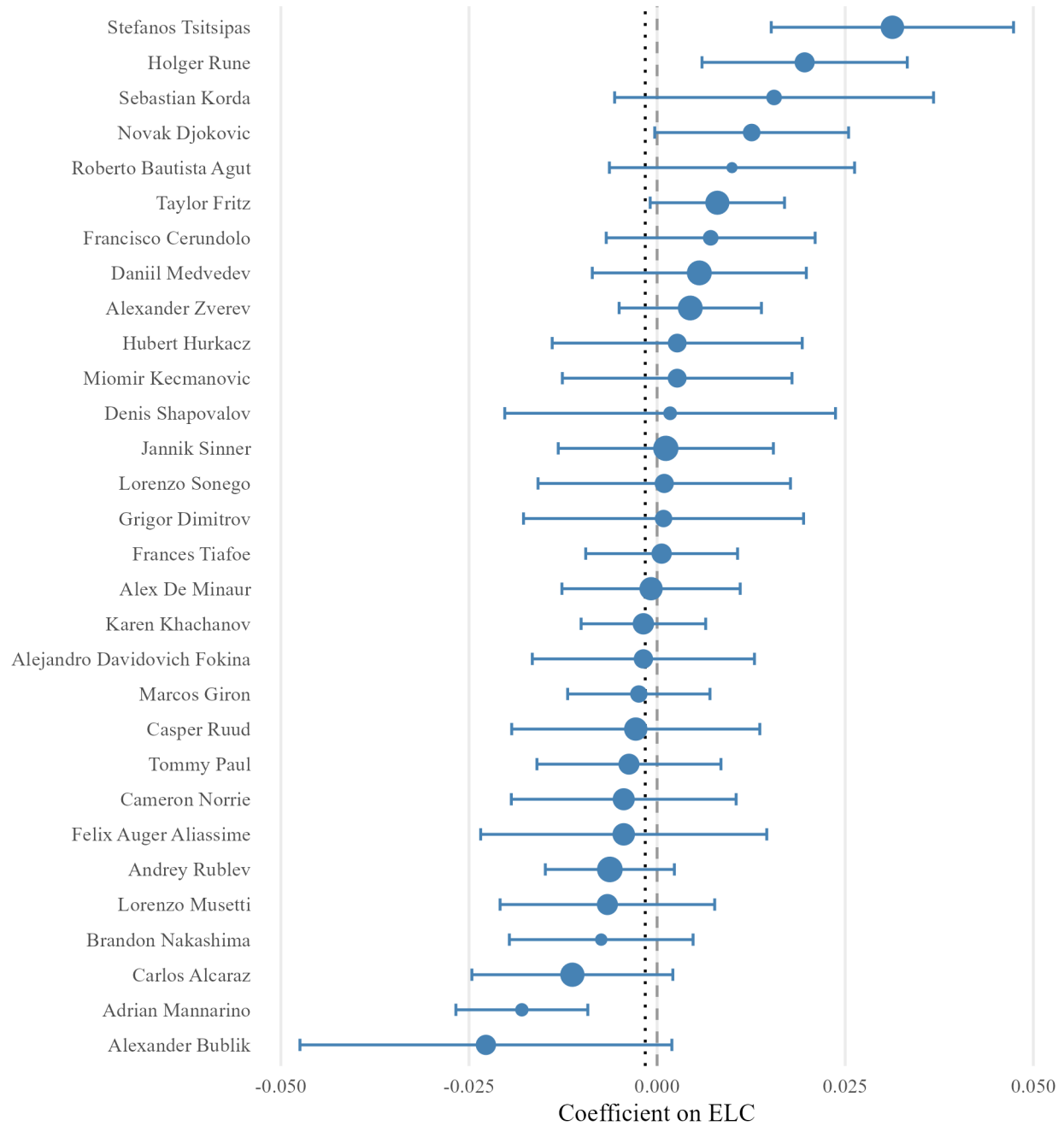


Figure A.7: Player-Specific ELC Effects on the Double-Fault Rate. Each estimate comes from a within-player fixed-effects regression with tournament and year fixed effects and player and opponent rank controls. Bars show 95% confidence intervals based on standard errors clustered at the tournament level. The sample includes the 30 players with the most match observations. The dotted vertical line marks the pooled estimate from the main specification. Point size is proportional to the player's number of match observations.

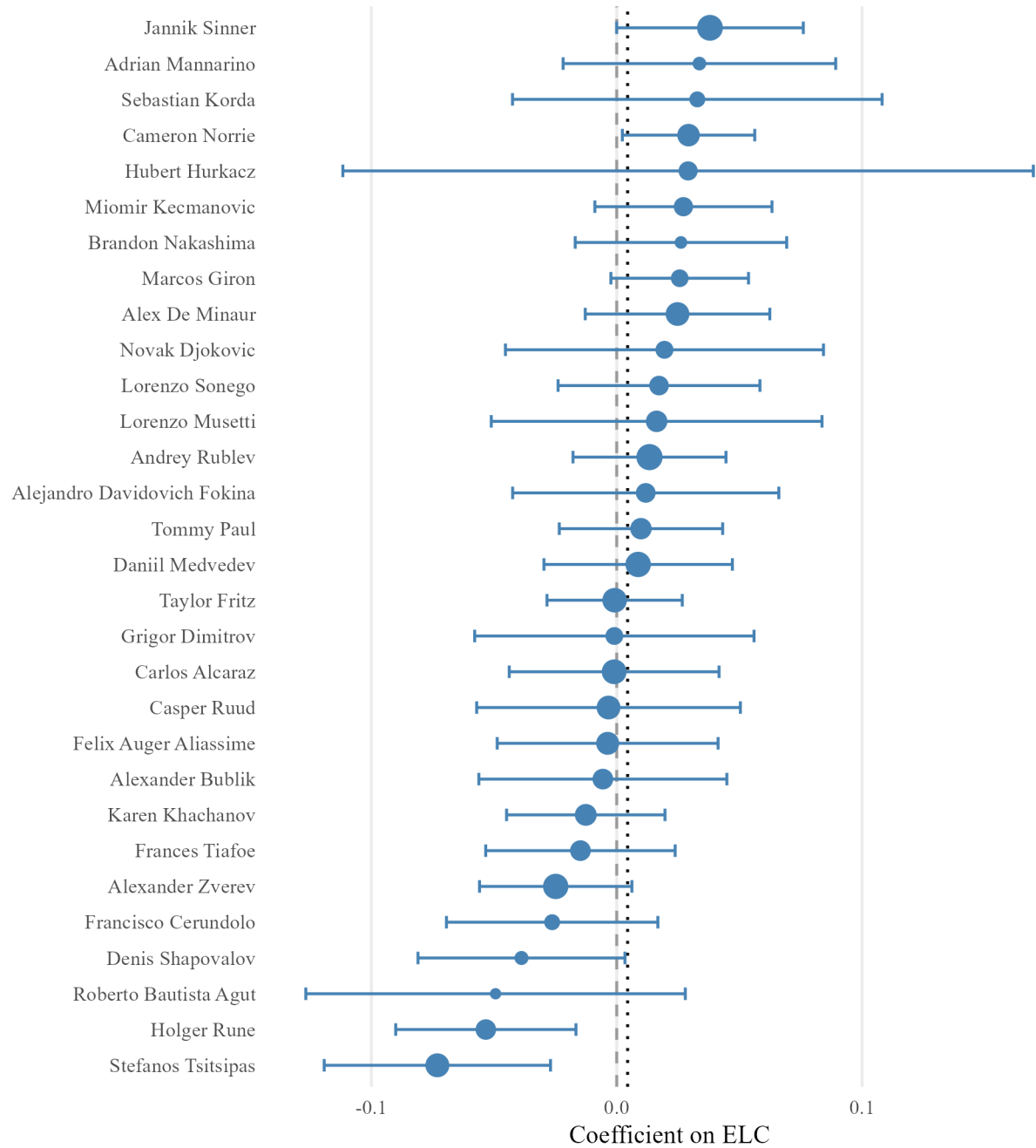


Figure A.8: Player-Specific ELC Effects on the First-Serve-In Rate. Each estimate comes from a within-player fixed-effects regression with tournament and year fixed effects and player and opponent rank controls. Bars show 95% confidence intervals based on standard errors clustered at the tournament level. The sample includes the 30 players with the most match observations. The dotted vertical line marks the pooled estimate from the main specification. Point size is proportional to the player's number of match observations.

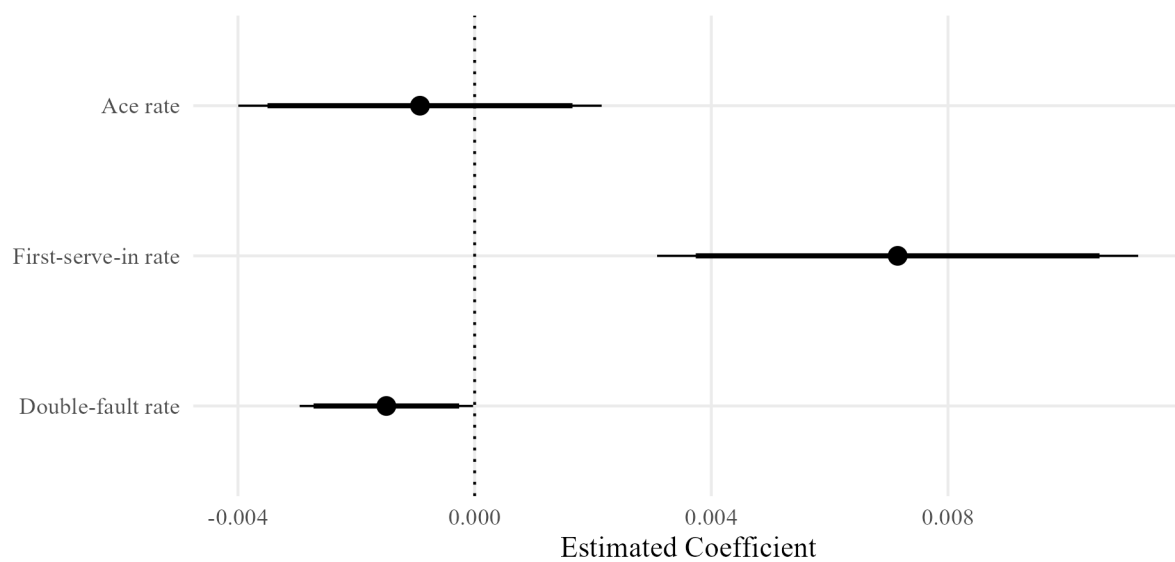


Figure A.9: Stacked Difference-in-Differences Estimates of the ELC Effect. Each estimate comes from a stacked difference-in-differences regression on 2018–2025 that compares each adoption cohort with tournaments not yet treated. Thick bars show 90% and thin bars show 95% confidence intervals based on standard errors clustered at the player level.

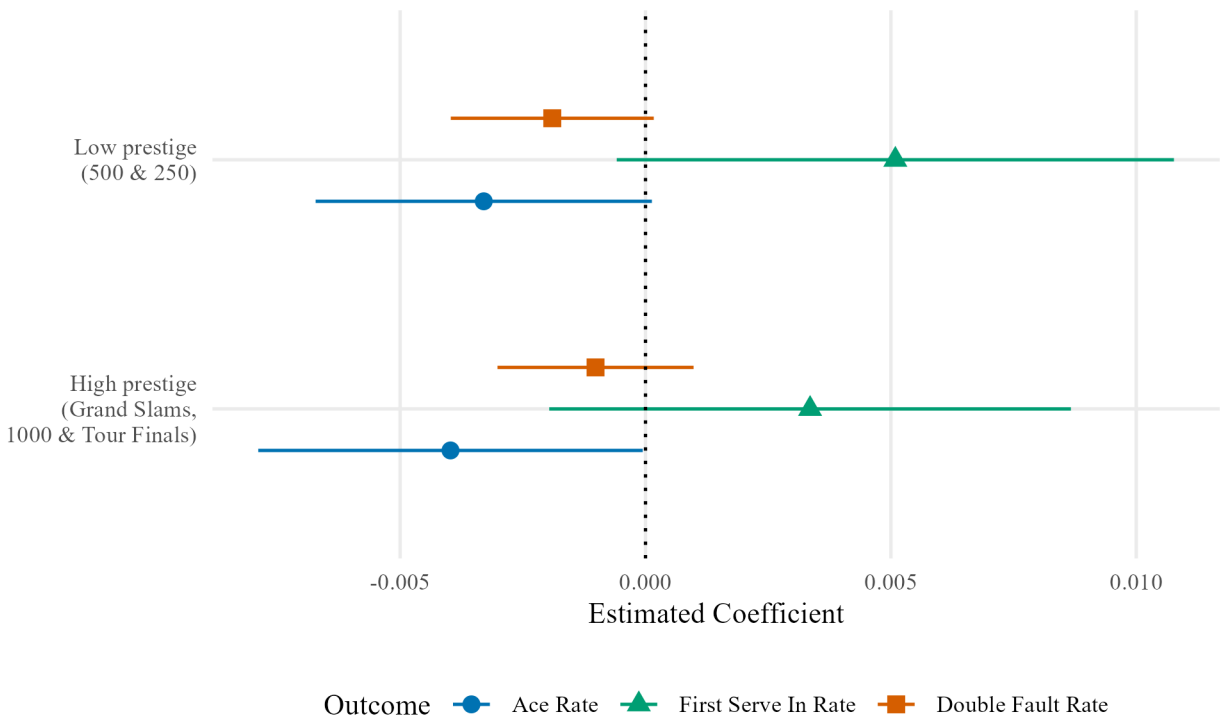


Figure A.10: ELC Effects by Tournament Prestige. Each panel plots fixed-effects point estimates and 95% confidence intervals for the ELC effect on each outcome, separately by tournament prestige. Grand Slams, ATP Finals, and Masters 1000 events represent the highest-prestige, highest-stakes tournaments. Estimates come from a single specification interacting ELC with the group dummy, with player, tournament, and year fixed effects and player and opponent rank controls. Confidence intervals are based on standard errors clustered at the player level.

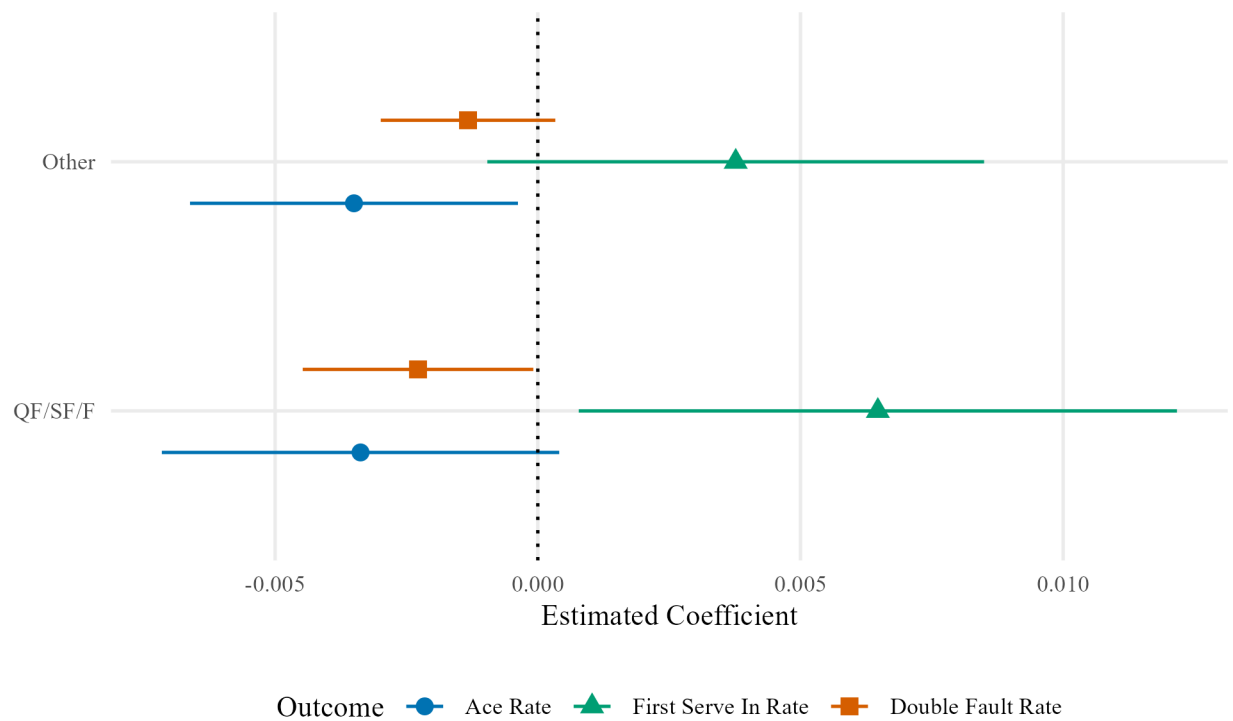


Figure A.11: ELC Effects by Tournament Round. Each panel plots fixed-effects point estimates and 95% confidence intervals for the ELC effect on each outcome, separately for late rounds (quarterfinals, semifinals, and final) and all earlier rounds. Estimates come from a single specification interacting ELC with the group dummy, with player, tournament, and year fixed effects and player and opponent rank controls. Confidence intervals are based on standard errors clustered at the player level.

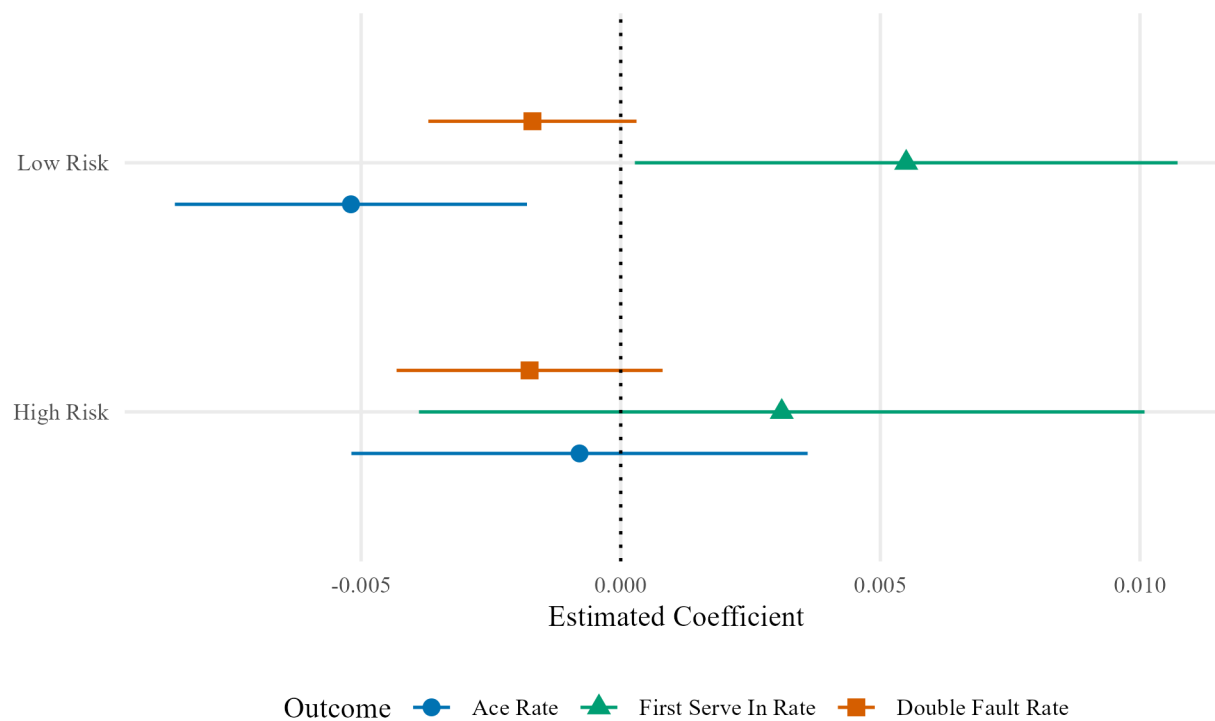


Figure A.12: ELC Effects by Pre-ELC Double-Fault Risk. Tournaments are split at the median of players' average double-fault rates in the pre-ELC years (2018–2019); the high-risk group had above-median double-fault rates and the low-risk group below-median. Each panel plots fixed-effects point estimates and 95% confidence intervals for the ELC effect on each outcome, separately by group. Estimates come from a single specification interacting ELC with the group dummy, with player, tournament, and year fixed effects and player and opponent rank controls. Confidence intervals are based on standard errors clustered at the player level.

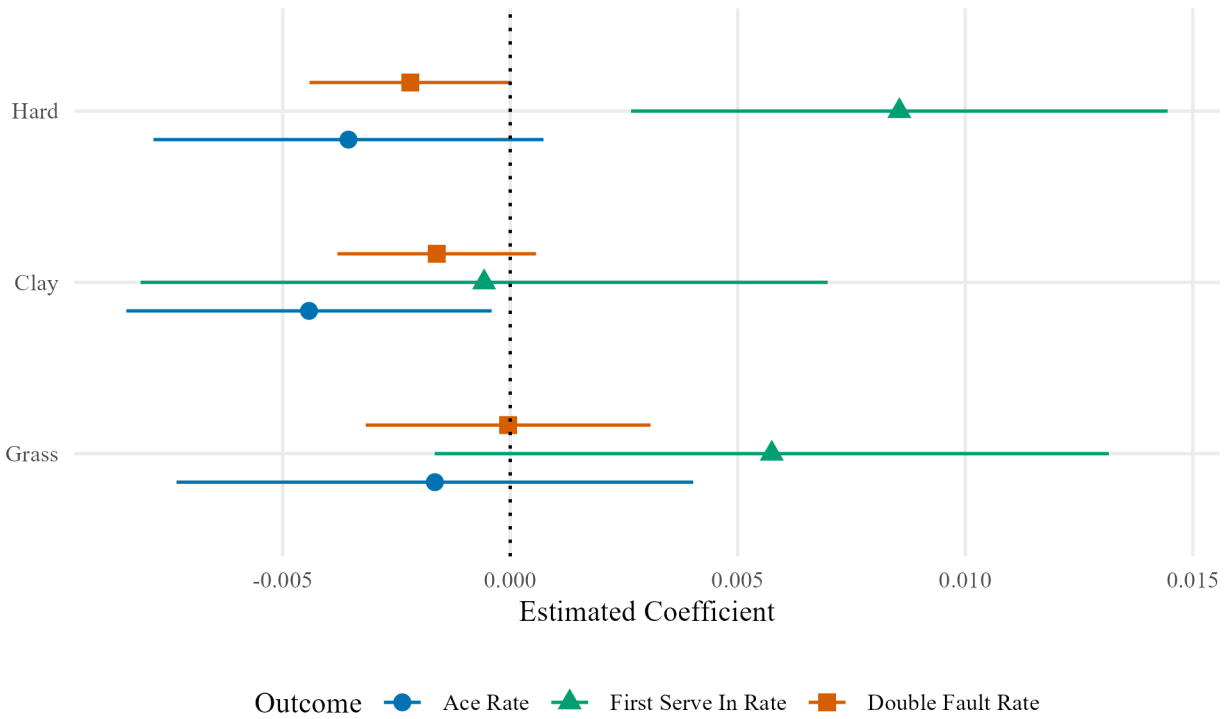


Figure A.13: ELC Effects by Court Surface. Each panel plots fixed-effects point estimates and 95% confidence intervals for the ELC effect on each outcome, separately by court surface. Estimates come from a single specification interacting ELC with the surface dummies, with player, tournament, and year fixed effects and player and opponent rank controls. Confidence intervals are based on standard errors clustered at the player level.