

Human Responses to Algorithmic Correction: Evidence from Centre Court *

David Almog[†], Romain Gauriot[‡], Lionel Page[§], and Daniel Martin[¶]

October 15, 2026

Abstract

We study how algorithmic review changes the initial decisions that a system is designed to correct. Using Hawk-Eye data from professional tennis, we compare umpire decisions with and without review against a common accuracy benchmark. Under review, initial errors were 8.6 percentage points higher just outside the line and lower for comparably close balls inside. We find no evidence that average initial accuracy deteriorated. Within matches, reversals were followed by larger adjustments than upheld calls, consistent with adaptation to correction. Overall, correction offsets worse initial decisions for incorrect in calls and reinforces better initial decisions for incorrect out calls. Whether review lowers organizational costs depends jointly on these changes and the costs of remedying corrected errors. A conditional illustration shows how utilization spreads fixed review costs across calls.

*We thank Lucas Lippman, Anastasiia Morozova, and Leshan Xu for their excellent research assistance. We are very grateful to the executives at Hawk-Eye Innovations Ltd and a leading tennis official at the time of Hawk-Eye’s introduction for providing valuable institutional information. We also thank David Argente, Meghan Busse, Colin Camerer, Andrew Caplin, Ben Enke, George Georgiadis, Mitchell Hoffman, Alex Imas, John List, Christopher Mills, Devin Pope, Michael Powell, Ashesh Rambachan, Antonio Rangel, Paul Raschky, Raphaël Raux, Luis Rayo, Yuval Salant, Jörg Spenkuch, Andrew Wooders, and audiences at EC 2024, BDRM, AI and the Future of Work at Wharton, ESIF Economics and AI+ML Meeting, Machine Learning in Economics Summer Conference, Zurich Workshop in AI+Economics, WEBEAS, MIT Sloan, Monash, UT Dallas, UH Manoa, Kellogg Strategy Brown Bag Seminar, and Booth Behavioral Lab for many helpful comments. We thank the Sloan Foundation for supporting this research under the “Cognitive Economics at Work” grant.

[†]Harvard Business School, dalmog@hbs.edu.

[‡]Deakin University, romain.gauriot@deakin.edu.au.

[§]University of Queensland, lionel.page@uq.edu.au.

[¶]University of California, Santa Barbara, danielmartin@ucsb.edu.

1 Introduction

Many organizations now use algorithmic systems to advise workers.¹ A more interventionist use is algorithmic review. A highly accurate system audits human decisions and blocks or reverses likely mistakes. Examples include algorithmic screening of human pricing interventions (Huelsen et al. 2024) and proposed applications in law and medicine (Kleinberg et al. 2018; Rambachan 2024; Raghu et al. 2019). Such review appears to be a straightforward way to improve decisions. Yet its full impact depends on whether workers anticipate the possibility of being corrected and change how they decide. When an algorithm can overturn workers’ decisions, they may respond to the resulting career consequences and operational disruptions, as well as to the personal and social experience of being overruled. These responses can alter both the accuracy and composition of their decisions.

Algorithmic review does more than correct a fixed set of human errors. It changes the decision problem faced by the worker. Review can change the expected consequences of an error because some decisions are more likely to be reviewed or reversed, or because the resulting correction is more visible or disruptive. When these expected consequences differ across errors, workers may change their decision threshold before the algorithm intervenes. The performance of the hybrid system is therefore jointly determined by the algorithm, the review protocol, and the worker’s response to both. Evaluating review while holding human behavior fixed can miss changes in the number and composition of initial decisions that the system is asked to correct. We therefore distinguish initial accuracy, the errors corrected by review, and the costs that remain after correction. A correct reversal need not undo the consequences of the initial decision.

We study how human decision makers respond when a highly accurate algorithm begins to review and potentially reverse their decisions. This question is difficult to study in most workplaces. Researchers may lack both the initial decision before correction and objective ground truth. Outcomes are also often observed only after particular decisions. For example, whether a defendant would flee is observed only if the defendant is released. Moreover, decision makers may possess private information unavailable to the algorithm, so disagreement need not constitute a human error. We address these challenges using the introduction of Hawk-Eye review in professional tennis. The testing period provides initial decisions both with and without review, while Hawk-Eye supplies a highly accurate benchmark without selective labels. Because line calls are repeated binary decisions with little scope for private information relevant to the decision, we can measure changes in the initial decision and in both error types.

Beginning in 2006, major tennis tournaments allowed players to challenge line calls.

¹For example, see Hoffman, Kahn, and Li (2018), Kawaguchi (2020), Agarwal, Moehring, Rajpurkar, et al. (2023), and Grimon and Mills (2025).

Hawk-Eye combines camera data with the ball’s trajectory and spin to reconstruct where the ball landed. When its assessment contradicted the call, the initial decision was publicly overturned in the stadium.² The protocol also created an important asymmetry. An incorrect in call could usually be corrected by awarding the point to the challenger, or by declaring a fault on a first serve. An incorrect out call had already stopped play, so correction could require the chair umpire to decide whether to award the point or replay it. The second remedy involved more discretion and could be contentious. The introduction of review therefore changed the expected consequences attached to different errors.

Because Hawk-Eye recorded ball locations at some tournaments before players were allowed to challenge calls, we can compare initial umpire decisions before and after the review regime began. But adoption was not randomized and almost all tournament editions without review precede those with review. We therefore interpret the estimates as changes associated with the review regime rather than as the isolated causal effect of review. Observable bounce composition and speed distributions are similar across periods. The central pattern is stable to richer controls. The estimate also remains economically sizable but is less precise when we restrict the sample to events observed under both regimes and include event fixed effects. There were no substantial changes to umpire training and performance assessment, the pool of umpires, or the positioning and instructions given to line judges.³

The main result is a change in the human decision rule. For balls landing within 20 mm of the line, the adjusted rate of initial in calls was 6.2 percentage points higher under review. It rose from 42.8% to 49.0%. In our preferred controlled specification, the mistake rate was 8.6 percentage points higher for balls landing just outside the line and lower for comparably close balls inside. The review regime was therefore associated with a reallocation of errors across the decision boundary rather than a uniform improvement or deterioration in initial decisions.

We also find evidence within matches consistent with adaptation to public correction. After an in-to-out reversal, umpires become less likely to call subsequent close balls in. After an out-to-in reversal, they become more likely to do so. The latter response is stronger in our linear specification. This asymmetry mirrors our main finding because the response is more pronounced following the error that stops play and requires a potentially contentious remedy at the umpire’s discretion. Because challenges and reversals are endogenous, this analysis is not a causal test of a specific mechanism. In particular, it cannot distinguish shame or embarrassment from reputational concerns, learning, strategic responses to the challenge process, or a desire to avoid contentious remedies.

²Before stadium review, Hawk-Eye replays could inform media and television commentators and generate public criticism of incorrect calls. Stadium review nevertheless made correction immediate and visible to umpires, players, and spectators.

³These institutional details were verified by a leading tennis official at the time of Hawk-Eye’s introduction.

A common concern about downstream review is that decision makers may treat correction as a safety net and become less accurate before review. We find no evidence of such deterioration. For calls within 100 mm of the line, the controlled estimate is a 1.1 percentage point reduction in the mistake rate ($SE = 0.008$), although it is not conventionally statistically significant. The estimated decline for non-serves is larger at 2.3 percentage points (17%). The estimate for serves is close to zero. Slower non-serves may leave greater scope for adjustments in attention, although we neither observe effort directly nor establish a difference between these subgroup effects.

Our findings are most relevant to review systems in which algorithmic assessments are accurate, corrections are visible, and different errors carry different expected consequences under review.⁴ In fraud screening, lifting an erroneous block on a payment and recovering a transfer mistakenly allowed to proceed involve different remedies. Similarly, reversing an incorrect denial of a claim may authorize payment without undoing the claimant’s delay. These examples illustrate the distinction between correcting a decision and restoring the situation that would have prevailed without the mistake, but whether the same behavioral response occurs in those settings requires separate evidence.

This paper provides field evidence on how initial human decisions differ under algorithmic review. Recommendations change the information available when a worker decides, whereas review changes the expected consequences of different errors. Section 5 shows why initial accuracy alone is insufficient to evaluate review. In our descriptive comparison of calls within 100 mm of the line, fewer remaining errors arise through different combinations of initial decisions and correction. We derive conditions for organizational savings, allowing the losses from uncorrected errors and the costs of remedying corrected errors to differ across types. The combined savings must cover the cost of providing review, although neither error category needs to pay for itself separately. The illustration using system costs then shows how utilization affects these conditions.

1.1 Related Literature

Our evaluation framework is closely related to work on the design of human–AI collaboration. Agarwal, Moehring, and Wolitzky (2025) characterize collaboration policies using human accuracy responses under assumptions that allow these responses to predict performance across policies. Agarwal, Moehring, Rajpurkar, et al. (2023) evaluate collaboration between radiologists and AI using diagnostic accuracy and the cost of human time. Both papers relate human behavior to the objective of a designer choosing how to use AI. We study review after the human decision, where a correct classification can leave costs of the initial

⁴In Appendix B, we assess the external validity of our setting using the framework proposed by List (2020).

mistake. Our framework relates the observed human response to the quantities needed for evaluation, without identifying an optimal policy.

This paper contributes to the literature on human–algorithm interaction by studying ex post algorithmic review. Most empirical work examines systems that provide recommendations before a human decision. This literature shows that system performance depends on how decision makers use algorithmic advice. Discretion may reflect biases rather than private information (Hoffman, Kahn, and Li 2018). Adherence varies with context and experience (Kawaguchi 2020), and risk scores may reallocate attention rather than induce mechanical compliance (Grimon and Mills 2025). Belief updating from qualitative AI recommendations depends systematically on prior beliefs and on whether those recommendations confirm or contradict them (Kovach, Martin, and Tserenjigmid 2026), while miscalibrated beliefs can produce systematic overreliance or underreliance on algorithmic advice (Caplin et al. 2025; Balakrishnan, Ferreira, and Tong 2026). Accordingly, providing AI predictions need not improve average performance (Agarwal, Moehring, Rajpurkar, et al. 2023), while appropriately coarsening the information conveyed by AI risk scores can improve decisions (Hoong and Dreyfuss 2026). These studies examine how the information supplied by an algorithm shapes assisted decisions. Our setting instead removes the contemporaneous recommendation. Line judges make their initial calls without observing the system’s assessment. The relevant behavioral margin is therefore how the prospect of subsequent review changes their decision rule before the algorithm intervenes.

Algorithmic review lies between recommendations and full automation. Work on human–algorithm decision rights considers how authority should be allocated (Athey, Bryan, and Gans 2020; Ide and Talamàs 2025) and how predictions can support automation, triage, or oversight (Kleinberg et al. 2018; Raghu et al. 2019; Rambachan 2024; Huelden et al. 2024). Under review, the human acts first but an algorithm can subsequently alter the outcome. The review protocol thus affects both which mistakes are corrected and the consequences attached to different initial decisions. Algorithmic systems can therefore shape choices not only by conveying information but also by changing action costs. McLaughlin and Spiess (2024) model recommendations as defaults that make deviations costly, while Albright (2024) finds that lenient bail recommendations affect judges’ decisions even when the underlying risk information is unchanged, consistent with reputational cover. Our findings are consistent with analogous incentive effects under ex post review, where correction changes the consequences attached to different errors. Accurate review can consequently shift the initial decision threshold and reallocate errors across types, so system performance cannot be inferred from a fixed distribution of human decisions. On a different dynamic margin, Cho (2024) examines skill depreciation following automated assistance among baseball umpires. More recent work on MLB’s 2026 Automated Ball-Strike challenge system shows that asymmetric challenge pressure can shift umpires’ decision thresholds within games (Michopoulos

et al. 2026) and documents brief adjustments near the decision boundary following overturns and players’ decisions to invoke review (Lee et al. 2026). These findings provide evidence from a subsequent setting consistent with our central result that selective algorithmic review is associated with differences in initial human decisions.

Related evidence shows that people respond differently to human and algorithmic assessment. Avery, Leibbrandt, and Vecchi (2023) show that anticipated assessment by AI rather than humans changes job seekers’ application behavior, while Almog, Lippman, and Martin (2026) find that workers alter the quantity and quality of their output when evaluated by AI. Our setting does not allow us to distinguish a response specific to algorithmic review from a response to the introduction of review more generally. At the same time, it illustrates why algorithms can expand the scope of review. Hawk-Eye makes it possible to assess and correct decisions in real time with a combination of speed and accuracy that human review could not feasibly match.

Anticipated review also connects our paper to the literature on monitoring and accountability. Monitoring can change opportunistic behavior, compliance, and worker productivity (Nagin et al. 2002; Olken 2007; Gosnell, List, and Metcalfe 2020; Zou 2021). A distinctive feature of algorithmic review is that monitoring can both expose an error and remedy it in real time. These functions create potentially offsetting incentives. The availability of correction may provide a safety net, while disruptive or consequential reversals may induce workers to avoid the decisions most likely to be overturned. Moreover, when the likelihood or consequences of correction differ across errors, workers may shift their decision thresholds rather than simply exert more or less effort. Our evidence documents differences in error composition under review that aggregate accuracy alone does not capture.

Finally, the visibility of reversals connects to work on social image and observability (Bursztyn and Jensen 2017). Experimental evidence shows that image concerns can lead decision makers to disregard AI recommendations even when others’ perceptions carry no monetary consequences (Almog 2025a), and that making workers’ reliance on AI visible to an evaluator reduces adoption and performance (Almog 2025b). In our setting, an algorithmic reversal can similarly make an incorrect decision publicly observable, but visibility, reversal, and remediation are bundled together.

2 Setting and Data

Hawk-Eye review was first used at an ATP Tour event in March 2006 at the Nasdaq-100 Open in Key Biscayne (currently known as the Miami Open).⁵ Later that year, the US

⁵The ATP Tour is the top tennis tour organized by the Association of Tennis Professionals.

Open and many other tournaments on surfaces other than clay adopted this new technology by allowing players to challenge calls.⁶ Hawk-Eye uses six to ten computer-linked television cameras positioned around the court to collectively create a three-dimensional representation of the ball’s trajectory. Hawk-Eye reports having an average error of just 3.6 mm, so just like the ATP Tour we will treat Hawk-Eye readings as if they are the ground truth.⁷

2.1 Professional Tennis, Umpires, and Hawk-Eye

Professional tennis is a racket sport that is played either one-on-one (singles) or two-on-two (doubles). In singles tennis, two players compete against each other on opposite sides of the tennis court. The objective is to score points by hitting the ball “in” (within the bounds of the opponent’s side of the court). We will concentrate solely on men’s singles matches (singles matches where both players are men) because the analysis of other formats is underpowered in our data.⁸ The scoring system for tennis is based on a series of points, games, and sets. Players accumulate points to win a game and accumulate games to win a set. Matches are typically played as the best of three or five sets, with each set requiring a player to win at least six games.

A crew of up to ten umpires is involved in officiating a match. The roles typically include one chair umpire who oversees the entire match, making decisions on points, penalties, and overall match control. Additionally, there are usually nine line umpires positioned around the court, each responsible for specific lines, with the sole duty to determine whether the ball bounced in or out when it is relatively close to their respective line.⁹ The chair umpire has the authority to overrule the decisions of the line umpires if necessary, so drawing inferences on the performance of an individual umpire is complicated, especially since we do not have data on whether the chair umpire overruled an individual line umpire. Thus, in this paper, we evaluate the performance of the umpire crew as a whole, but the absence of chair umpire and crew identities prevents us from tracking the same officials across matches to assess their response to Hawk-Eye review.

International chair umpires are certified with Gold, Silver, or Bronze badges, while line umpires are graded according to national or other systems. Certification and grading methods remained unchanged during the study period, and Hawk-Eye metrics were not incorporated into these evaluations for the first couple of years. A leading tennis official who was

⁶Players have a fixed number of challenges that they can use along the match, but successful challenges do not count against this limit. Players were not allowed to challenge chair umpire decisions on non-clay courts before this system was put in place.

⁷For perspective, the standard diameter of a tennis ball is 67 mm.

⁸If more data were available on the other formats, it might be of interest to examine whether responses to algorithmic review vary by gender or team composition.

⁹Figure C.1 provides a visual representation of the positioning of each umpire.

involved in testing and introducing Hawk-Eye provided valuable institutional insights that will be used throughout the paper.

The Hawk-Eye review protocol works by endowing players with 2-3 challenges per set.¹⁰ If a challenge is successful, players do not lose a challenge opportunity. When a player challenges a call, a computerized path and the final landing location of the ball are displayed on a large screen in the stadium for the umpire, players, and the crowd to observe the outcome of the challenge.

The remedies differ according to the initial call. Reversing an incorrect in call establishes that the ball was out and ordinarily awards the point to the challenger. A first serve is an exception: a fault ordinarily gives the server a second serve, rather than ending the point. When an incorrect out call is reversed, the chair umpire must judge whether the call hindered play before awarding or replaying the point.¹¹ This is discretion exercised under the rules. The reversal corrects the line classification, but an interrupted rally cannot resume from the position it would have reached without the call. A replay therefore need not restore the outcome of uninterrupted play.

2.2 Data

We combine Hawk-Eye records of ball locations from over 100 tournaments, challenge records, and video audits to reconstruct initial umpire calls.¹² The location records provide the accuracy benchmark, challenge records identify reversals, and video audits guide and assess the rules for matching records and classifying calls. We were able to merge 2,038 out of the 2,108 challenges registered for the 28 tournaments common to the two data sets (a 97% merge rate).

The consolidated data set produced by merging the Hawk-Eye Base data set and the Challenge data set comprises a total of 698 matches across 35 tournament editions, and summary statistics for this data set can be found in Table 1.¹³ This data set includes 109 matches from seven tournament editions that were played before Hawk-Eye review and 589 matches from 28 tournament editions after Hawk-Eye review was active. The samples before and after Hawk-Eye review are very similar in the characteristics of points most relevant

¹⁰In practice, players very rarely exhaust their challenges. This is advantageous for our research question, as umpires are almost always under the threat of being challenged.

¹¹See the [ITF Rules of Tennis](#), Rules 20 and 24 and Appendix VII, Case 7 and Electronic Review Procedures.

¹²We excluded from the analysis the 15 clay court tournaments. We elaborate on this decision in Appendix A.1.

¹³Table C.1 provides information on the month, category, court type, and the number of matches for each tournament.

for identification, especially the shares of bounces within 20 mm and 100 mm of the line. The differences in tournament and round composition motivate the inclusion of match-level controls in the analyses below. Figure C.3 shows that the distribution of players' strokes was highly similar before and after the introduction of Hawk-Eye review, both in proximity to the line and in speed.

	Before Hawk-Eye review (PostHK=0)	After Hawk-Eye review (PostHK=1)	Total	Post - Pre [Std. diff.]
<i>A. Sample counts</i>				
Players	71	161	174	
Tournament editions	7	28	35	
Matches	109	589	698	
Points	15,435	83,902	99,337	
Bounces	71,630	402,467	474,097	
<i>B. Tournament tier (share of tournaments)</i>				
ATP 250	28.6%	42.9%	40.0%	+14.3 p.p. [0.30]
ATP 500	0.0%	10.7%	8.6%	+10.7 p.p. [0.49]
ATP 1000	71.4%	46.4%	51.4%	-25.0 p.p. [0.53]
<i>C. Match round (share of matches)</i>				
Final	6.4%	4.6%	4.9%	-1.8 p.p. [0.08]
Semifinal	12.8%	9.0%	9.6%	-3.8 p.p. [0.12]
Quarterfinal	17.4%	14.9%	15.3%	-2.5 p.p. [0.07]
Round of 16	12.8%	19.2%	18.2%	+6.3 p.p. [0.17]
Other	50.5%	52.3%	52.0%	+1.8 p.p. [0.04]
<i>D. Bounce composition (share of all bounces)</i>				
< 20 mm from the line	0.78%	0.69%	0.70%	-0.09 p.p. [0.01]
< 100 mm from the line	3.66%	3.53%	3.55%	-0.13 p.p. [0.01]
Serves	29.2%	27.6%	27.8%	-1.64 p.p. [0.04]
Non-serves	70.8%	72.4%	72.2%	+1.64 p.p. [0.04]
<i>E. Average speed in km/h (s.d.)</i>				
Serves	147.7 (24.2)	150.0 (23.1)	149.6 (23.3)	+2.3 [0.10]
Non-serves	81.7 (21.4)	83.3 (20.4)	83.1 (20.5)	+1.6 [0.08]
All	101.0 (37.4)	101.7 (36.5)	101.6 (36.6)	+0.7 [0.02]

Table 1: Summary statistics for the consolidated data set. The last column reports the difference in the sample characteristics before and after Hawk-Eye review. Bracketed values are absolute standardized differences.

2.2.1 Hawk-Eye Base Data Set

The main source of data for this paper is the Hawk-Eye Base data set, which consists of the official Hawk-Eye data for matches played at the international professional level between March 2005 and March 2009. This data set provides a number of pieces of information for

each point, including the position of every bounce captured by the Hawk-Eye system during that point, the serving player, the ongoing score, and the point winner. Altogether, this data set includes information on 1.8 million bounces from more than 1,800 men’s singles matches, spanning various prestigious tournaments such as Grand Slam tournaments, Masters 1000 tournaments, and International series tournaments. This data set has been used in the past to study important economics questions such as risk management (Ely, Gauriot, and Page 2017), tournament incentives (Gauriot and Page 2019), and mixed strategy play (Gauriot, Page, and Wooders 2023). These papers have focused on the players’ perspective, and the Hawk-Eye Base data set is well-suited for this purpose. However, this data set alone is insufficient to analyze umpires’ performance, as it does not permit us to identify whether the final call comes from the umpire or via an overturned challenge.

2.2.2 Challenge Data Set

The Challenge data set was originally obtained from the ATP Tour and encompasses all challenges recorded during 35 tournaments across the initial three years following the introduction of Hawk-Eye (2006-2008). It captures comprehensive information on 2,784 challenges, including details about the players involved, the match itself, the specific point in the match when the challenge was made, and the outcome of each challenge. Of the 35 tournaments, we use the 28 that also appear in our Hawk-Eye Base data set. In order to compile this data set, Abramitzky et al. (2012) undertook the arduous effort of collecting and compiling umpire match sheets. Abramitzky et al. (2012) acknowledge the selection problem involved in only observing the challenged points (around 2.6% of total points). Nevertheless, they are able to draw inferences on the optimality of decision-making from the players’ standpoint by relying on the idea that a higher propensity to challenge implies challenging less conservatively. However, without the point-by-point data, it is hard to assess the umpire’s performance once Hawk-Eye review was active. Combining the first two data sets allows us to analyze initial calls beyond the selected set of challenged points.

2.2.3 Video Auditing

In order to assess the quality of our merging algorithm, we audited full-match video replays using TennisTV, the official ATP streaming service. We identified 43 matches that appeared in both the Hawk-Eye Base and Challenge data sets, providing us with an assessment of how well a given merging algorithm would work on the 144 challenges witnessed during these matches.

Using variables such as match, point, distance, and who hit the ball, we developed a merging algorithm with a 99.3% merge rate, for which only 4.9% of the video-audited

challenges were merged to an incorrect point. The algorithm consists of eight iterations, starting with the most stringent merging rule and gradually relaxing criteria for challenges that persist without merging.¹⁴

As an additional benefit, auditing match replays enabled us to test the effectiveness of the four criteria we implemented to identify incorrect calls (two criteria for each type of mistake). We deem an in call to be incorrect when a player’s stroke is recorded as bouncing out in the Hawk-Eye Base data set, yet they still win the point, or if the data set records at least three more strokes after an out bounce. We deem an out call to be incorrect if all the strokes of a point are recorded as in, and the player delivering the final hit loses, or if there is a second serve after the first one was recorded as in the Hawk-Eye Base data set.¹⁵

3 Empirical Results

We examine how umpiring decisions changed after Hawk-Eye review was introduced. We begin with raw evidence of a shift in the decision rule and the composition of errors. We then estimate the pattern across distances with point and match controls and discuss identification and supporting diagnostics. Next, we present suggestive evidence within matches about possible mechanisms and consider the average mistake rate as a secondary outcome. Section 4 examines heterogeneity, and Section 5 connects these findings to the evaluation of review systems.

3.1 Error Reallocation Near the Line

Figure 1 shows how umpire decisions differ with the introduction of Hawk-Eye review. It plots the probability that a ball is initially called in against its signed distance from the line, separately before and after review. As expected, the probability of an in call rises sharply as the true landing location moves from outside to inside the line. Near the boundary, the curve under review lies above the curve without review. Balls landing at similar locations were more likely to be initially called in under review.

To determine whether the distance-specific pattern persists after accounting for observable differences across periods, we estimate the following specification:

$$\mathbb{1}(Incorrect\ Call)_{ipm} = \alpha_0 + \sum_{b \in \mathcal{B}} \alpha_b (PostHK_m \times D_{ib}) + \beta X_p + \gamma Y_m + \epsilon_{ipm} \quad (1)$$

¹⁴A detailed explanation of the merging algorithm is provided in Appendix A.2.

¹⁵Appendix A.3 documents these criteria further.

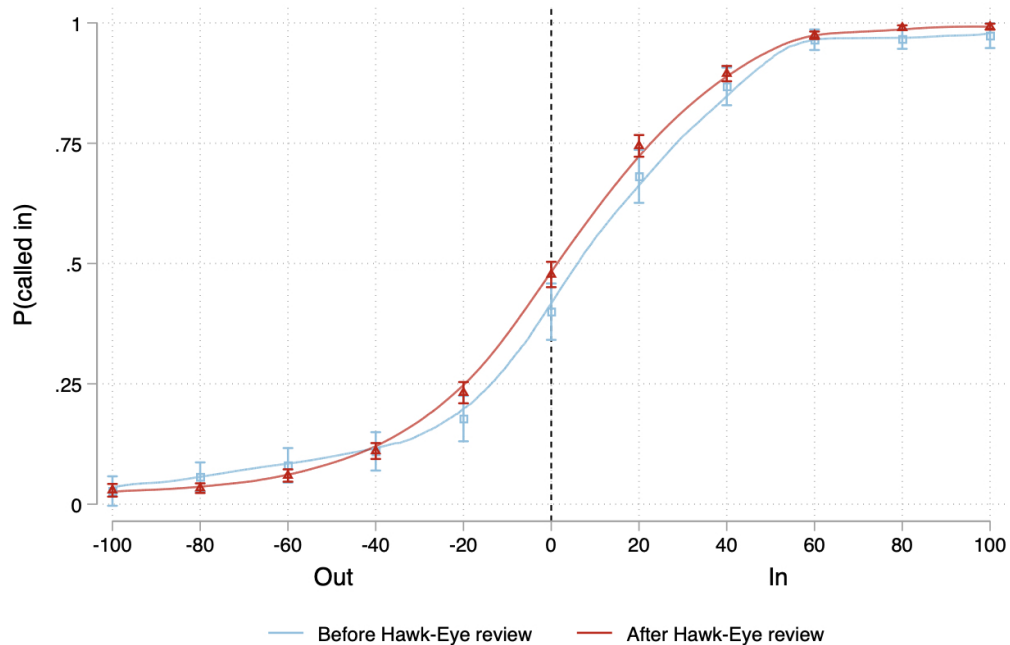


Figure 1: The figure plots the probability that a ball is initially called “in” against its signed distance from the line. Negative distances indicate balls landing out and positive distances balls landing in. Each marker reports the mean call-in rate for observations within 10 mm of the indicated distance. For example, the marker at zero includes balls landing between -10 and 10 mm from the line. Bars denote 95% confidence intervals. Curves are locally weighted scatterplot smoothing (LOWESS) fits, estimated separately before and after the introduction of Hawk-Eye review. The vertical dashed line denotes the court boundary.

$\mathbb{1}(\textit{Incorrect Call})$ is an indicator equal to 1 if call i in point p of match m was incorrect. $\textit{PostHK}_m$ equals 1 if Hawk-Eye review was active in the tournament edition containing match m . D_{ib} identifies signed-distance bin b . The ten coefficients α_b estimate the adjusted change in mistakes within each bin. X_p comprises 20 mm signed-distance-bin fixed effects, an indicator for ball speed above the sample median, indicators for the server’s point score, game number, and set number, and a tie-break indicator.¹⁶ Y_m comprises indicators for the match round and tournament tier.¹⁷ We cluster standard errors by tournament edition, yielding 35 clusters. This permits arbitrary correlation across calls and matches within the same tournament edition. Clustering by match is a plausible alternative because each match has its own officiating crew. It yields qualitatively similar and more precise results.

Figure 2 reports the interaction coefficients for $\textit{PostHK}_m$ and indicators D_{ib} for each of the ten 20 mm distance bins, with 95% confidence intervals based on standard errors clustered at the tournament-edition level. In the raw data, the mistake rate in the closest outside bin was 8.5 percentage points higher under review than the no-review baseline of approximately 25.0%, a 34% relative increase. After adjustment for point and match characteristics, the estimated difference was 8.6 percentage points, while the mistake rate was lower for comparably close balls inside. Consistent with a shift in the decision threshold, the estimated change in the closest outside bin exceeds the average change in the first two inside bins by approximately 14.5 percentage points. This contrast is statistically significant using wild-cluster-bootstrap inference at the tournament-edition level ($p = 0.034$).¹⁸

We examine the closest calls more directly by studying call-in rates for balls within 20 mm of the line, regardless of which side they landed on. The regression-adjusted call-in rate is 42.8% without Hawk-Eye review and 49.0% with review (see Figure C.2).¹⁹ The difference also appears gradual. The fitted monthly trend is -0.62 percentage points before review and 0.61 percentage points afterward, and the estimated difference in slopes is 1.23 percentage points per month ($p = 0.002$).

Taken together, the raw and regression-adjusted estimates indicate a reallocation of initial errors near the boundary under Hawk-Eye review. Whether that reallocation improves system performance also depends on which errors are corrected and the burden of their remedies. We return to these outcomes in Section 5. We first state the identifying assumption underlying the empirical comparison and assess the evidence supporting it.

¹⁶A tie-break is a one-off game held to decide the winner of a set when two players are locked at 6-6.

¹⁷Because the period without review has no ATP 500 tournament editions, we combine the ATP 250 and ATP 500 tiers, so the tier control distinguishes Masters 1000 tournaments from ATP 250 and ATP 500 tournaments.

¹⁸Table C.2 reports wild-cluster-bootstrap inference for the individual distance-bin estimates and cross-boundary contrasts.

¹⁹The fact that balls landing close to the line are more often called out is consistent with evidence from the visual-perception literature on moving objects (Whitney et al. 2008).

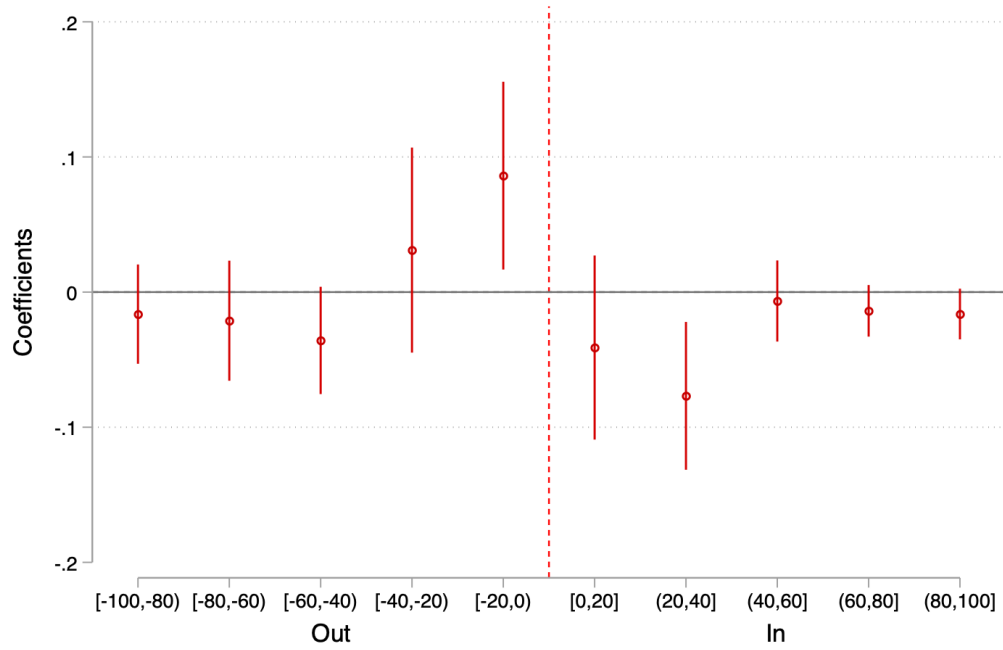


Figure 2: Estimated change in the incorrect-call rate after the introduction of Hawk-Eye review, by proximity to the line. Each dot is the coefficient on the interaction between a distance-bin indicator and $PostHK$, which equals 1 when Hawk-Eye review is active. The regressions include point and match controls. Standard errors are clustered at the tournament-edition level.

3.2 Identification and Supporting Evidence

Our estimates compare tournament editions with and without Hawk-Eye review. A causal interpretation requires that the conditional relationship between signed distance and the initial umpire call would not have changed systematically across these editions without review. Our central result on error reallocation requires a more specific assumption. No contemporaneous change in tournament composition, shot difficulty, or officiating would independently have shifted close calls toward “in” and thereby increased errors just outside the line while reducing errors just inside it. This assumption cannot be tested directly because the editions without review precede nearly all editions with review, tournament composition differs across regimes, and umpire identities are unavailable.

We nevertheless examine observable implications of the assumption and the sensitivity of the result to alternative specifications and samples. The distributions of signed distance and speed are similar across regimes (Figure C.3). Controlling for speed, score, game, set, tie-break status, round, and tournament tier changes the estimated increase in errors for the closest outside bin only from 8.5 to 8.6 percentage points. The sign reversal around the line is also reproduced in separate distance-bin regressions and remains visible when the review period is divided into early and later subperiods (Figure C.4).

Restricting the sample to named tournaments observed under both review regimes and including named-event fixed effects reduces the estimated cross-boundary contrast from 14.5 to 10.2 percentage points. The estimate remains economically sizable and of the same sign, although it is imprecisely estimated in the substantially smaller sample (wild-cluster-bootstrap $p = 0.152$). No single named event drives the result. Estimates that omit each event in turn range from 7.7 to 12.5 percentage points (Table C.3).

The cross-boundary contrast is also stable to richer controls for shot composition and difficulty. Table C.4 shows robustness to serve status, flexible speed controls, and player fixed effects. Following Goldsmith-Pinkham, Hull, and Kolesár (2024), we also estimate one-treatment-at-a-time regressions to address possible contamination bias (see Figure C.5).

The reported controls and sample restrictions preserve the central pattern, although they cannot rule out unobserved changes in officiating or tournament conditions. We next consider suggestive within-match evidence about how umpires adapted following public review of their decisions.

3.3 Within-Match Responses to Algorithmic Review Outcomes

The bundled review regime creates several possible mechanisms that our design cannot separate. Institutional guidance or supervisory pressure could have encouraged a new default, al-

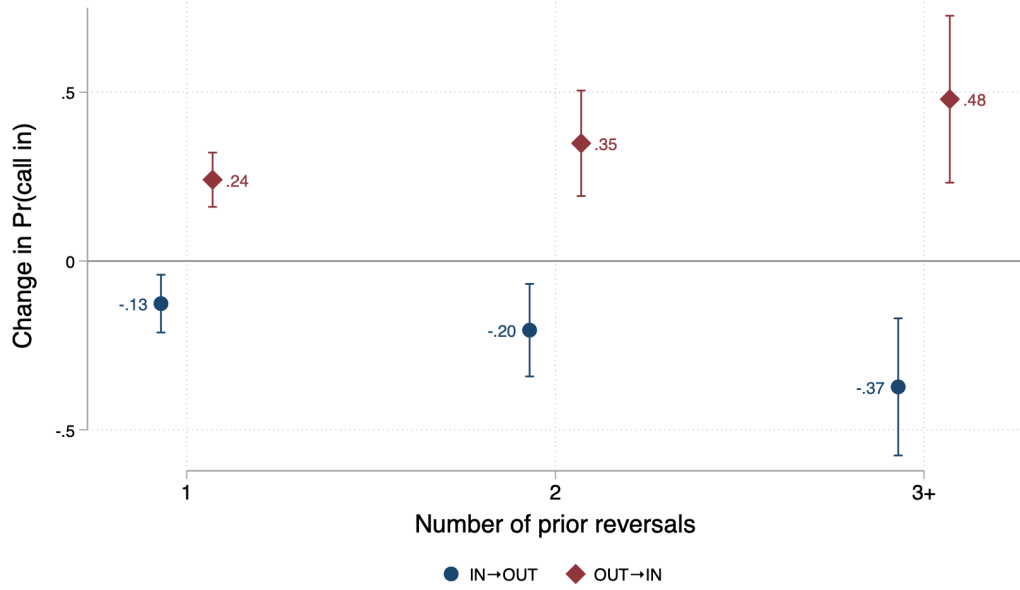
though the leading official we contacted confirmed that umpires were not instructed to make more in calls. Strategic adaptation to challenge exposure is another possibility, although challenge rates are similar across error types for close calls. Hawk-Eye could also facilitate learning about call accuracy. Finally, immediate public reversal could create reputational concerns, shame or embarrassment, or a desire to avoid the more discretionary remedy that follows an incorrect out call. Broadcast coverage made errors visible before stadium review, but Hawk-Eye changed the immediacy, publicity, and consequences of correction at the same time.

To provide evidence more closely tied to adaptation after public correction, we examine whether reversals are followed by changes in subsequent calls within the same match. For balls landing within 20 mm of the line, we estimate whether the probability of an initial in call varies with the number of earlier Hawk-Eye reversals. An in-to-out reversal shows that an initial in call was out. An out-to-in reversal shows that an initial out call was in. The regressions include match fixed effects to hold the officiating environment constant within each match.

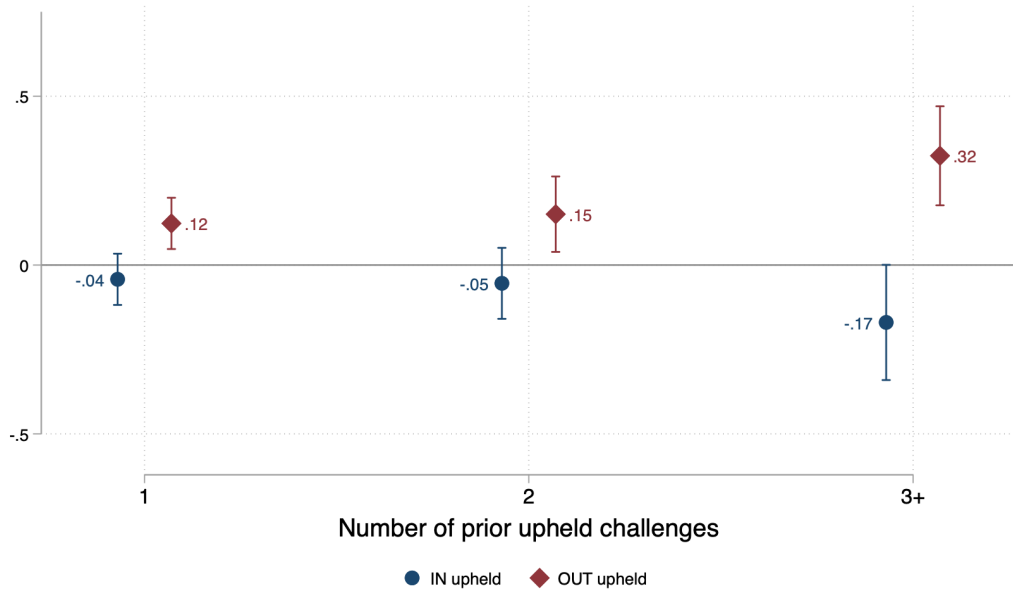
Because challenges and reversals are endogenous, this analysis is suggestive and should not be read as a stand-alone causal test. Figure 3 shows that subsequent calls move in the direction that would avoid repeating the previously corrected error. Following in-to-out reversals, umpires become less likely to call later close balls in. Following out-to-in reversals, they become more likely to do so. The adjustment is also asymmetric. In a specification entering the cumulative number of each reversal type linearly, the response to an additional out-to-in reversal exceeds the absolute response to an additional in-to-out reversal by 10.3 percentage points ($p = 0.011$). More flexible estimates after one, two, and three or more prior reversals yield asymmetries of 12.5, 17.8, and 15.6 percentage points, respectively, although these differences are jointly imprecisely estimated ($p = 0.151$).

Compared with an upheld challenge of the same initial-call direction, an in-to-out reversal is associated with an additional 6.7 percentage point shift toward out ($p = 0.078$), and an out-to-in reversal with an additional 11.5 percentage point shift toward in ($p = 0.004$). The corresponding coefficient differences are jointly significant ($p = 0.004$).

These patterns are consistent with adaptation to public correction and with a stronger response when the original error incorrectly stopped play. We next ask whether the reallocation was accompanied by a deterioration in average initial-call accuracy, as might occur if umpires treated review as a substitute for effort.



(a) Successful challenges (reversals)



(b) Unsuccessful challenges (upheld calls)

Figure 3: The figure reports coefficients from a linear probability model estimated on post-Hawk-Eye calls for balls landing within 20 mm of the line. The dependent variable equals one if the ball is initially called in. Panel (a) reports responses to prior successful challenges that reversed the initial call. Panel (b) reports responses to prior unsuccessful challenges that upheld the initial call. The four cumulative feedback measures are included jointly in the specification. For each series, the omitted category is zero prior events of that type within the match. Counts are capped at three, with the final category pooling three or more events. The specification includes match fixed effects and point-level controls. Standard errors are clustered at the match level. Bars denote 95% confidence intervals.

3.4 Overall Initial-Call Accuracy

A common concern about review and correction systems is that decision makers may treat the technology as a safety net and free ride on its ability to correct their mistakes, reducing effort and becoming less accurate before review. We examine this possibility using an analogous pooled specification in which the ten $PostHK_m \times D_{ib}$ interactions in equation 1 are replaced by a single $PostHK_m$ indicator.

For calls within 100 mm of the line, the controlled estimate is a 1.1 percentage point reduction in the initial mistake rate, about 8% of the baseline mean ($SE = 0.008$), although it is not conventionally statistically significant. Table C.7 reports the full regression results. Adding point and match controls changes the estimate from -1.4 to -1.1 percentage points. Raw mistake rates are lower in eight of the ten distance bins, but these comparisons do not measure effort or distinguish changes in effort from changes in the decision threshold.

Because this average combines potentially different responses across tasks, Section 4.2 examines serves and non-serves separately.

4 Examining Heterogeneity

We examine heterogeneity by tournament stage, tournament tier, and shot type.

4.1 Tournament Stage and Tier

Umpires are rewarded for good performance with assignments to advanced-stage matches, as the pool of officials needed becomes smaller and organizers can be more selective as the tournament progresses. These selection dynamics within tournaments are likely to create variation in umpiring skill across tournament stages.

To examine whether the change following Hawk-Eye review differs by tournament stage, we split our sample into two groups. The first includes Final, Semifinal, and Quarterfinal matches where we expect assignments to favor more skilled umpires. The second contains the remaining matches. Figure 4 reports the estimates separately for these groups. The estimated change in error composition near the line appears more pronounced for matches in earlier stages (magenta markers). The estimates for advanced-stage matches (green markers) are not statistically significant in any bin, but these differences in significance do not establish a difference between the groups.

In a study on radiologists, Chan, Gentzkow, and Yu (2022) found that aversion to false negatives tends to be negatively related to radiologist skill. Specifically, lower-skilled radiol-

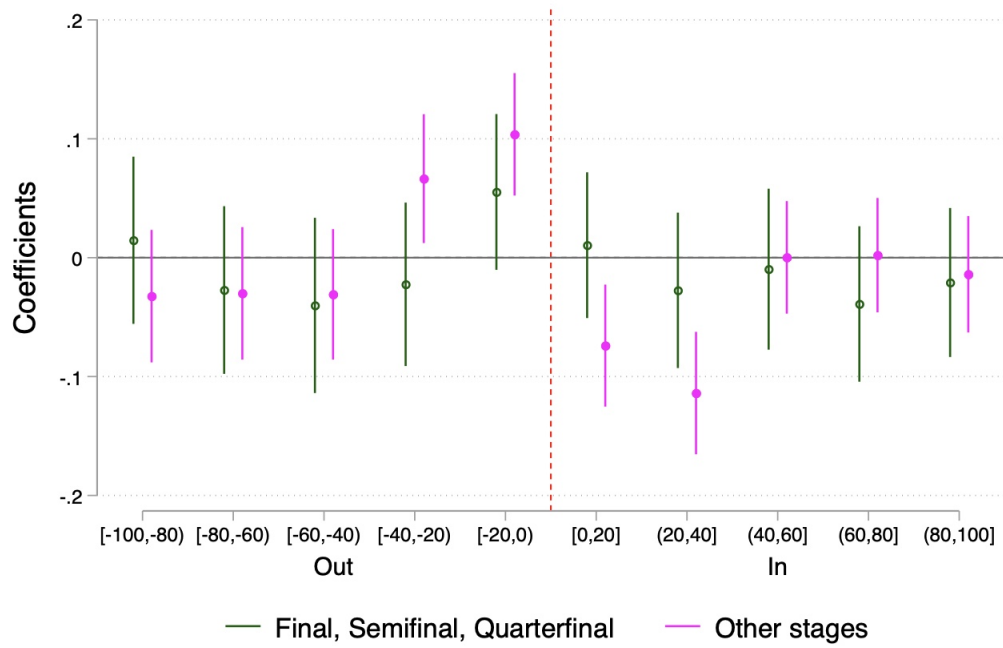


Figure 4: Estimated change in the incorrect-call rate after the introduction of Hawk-Eye review, by proximity to the line and tournament stage. Matches at the final, semifinal, and quarterfinal stages are analyzed separately from earlier-stage matches. Each dot represents the coefficient on the interaction between a distance-bin indicator and *PostHK*.

ogists are more likely to falsely diagnose a healthy patient with pneumonia (false positive) because this error is less costly than missing a diagnosis. This provides a possible interpretation of the tournament-stage pattern, but we cannot attribute it to officiating skill because stage is an imperfect proxy and umpire identities and skill are not directly observed.

We also compare higher-tier tournaments (Masters 1000) and lower-tier tournaments (ATP 500 and 250). The official we consulted reported substantial overlap in officials across tiers, although assignments also depend on geography, budget, availability, and logistics.

Figure C.6 does not show a comparable pattern when contrasting higher- and lower-tier tournaments. Tier and stage nevertheless capture different dimensions of stakes and assignment, so this comparison does not isolate skill from stakes or rule out either explanation for the tournament-stage pattern.

4.2 Serves and Non-Serves

There is value in distinguishing between serves and non-serves,²⁰ as we can think of them as distinct tasks for the umpires. For umpires, the primary differences between serves and subsequent hits are speed and anticipation. Serves may be more challenging because they are much faster. Their average speed in our sample is 150 km/h compared to 83 km/h for non-serves.²¹ Serves also allow anticipation because umpires know the ball is about to be served and will most likely bounce in the serving box. This appears to matter when we compare the slowest quartile of serves (132 km/h) with the fastest quartile of non-serves (97 km/h). Error rates are lower for serves despite their greater speed. The average incorrect-call rate during the period without Hawk-Eye review for the first quartile in serves and the fourth quartile in non-serves is 11.6% versus 15.6% for balls within 100 mm of the line, and 25.7% versus 40% within 20 mm of the line. These comparisons are consistent with anticipation helping umpires during serves, although differences in placement and other shot characteristics could also contribute.

Figure C.8 shows changes in error composition for both serves and non-serves. For serves, the main pattern is a reallocation of errors. For non-serves, it is a broader decline in mistakes.

Table 2 reports the pooled specification described in Section 3.4, estimated separately for serves and non-serves within 100 mm of the line. Although separate significance levels do not establish a difference between tasks, the estimates suggest different patterns for serves and non-serves. We do not find evidence of a change in average performance for serves. For non-serves, the estimated mistake rate declines by 2.3 percentage points, a 17% reduction relative to the baseline mean of 0.136. With tournament-edition clustering, its standard

²⁰The serve starts off every point in tennis, with players alternating serving each game.

²¹Figure C.7 shows the speed distribution for both types.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Incorrect call (Serves)				Incorrect call (Non-serves)			
PostHK	-0.006 (0.010)	-0.002 (0.009)	0.000 (0.009)	0.000 (0.009)	-0.024** (0.010)	-0.021** (0.010)	-0.023** (0.010)	-0.023* (0.012)
Point controls		X	X	X		X	X	X
Match controls			X	X			X	X
Cluster level				Tournament				Tournament
N	9,253	9,253	8,990	8,990	7,559	7,559	7,345	7,345
Baseline mean	.139	.139	.137	.137	.138	.138	.136	.136

Standard errors in parentheses

* $p < .10$, ** $p < .05$, *** $p < .01$

Table 2: OLS regressions of umpire mistakes for balls bouncing within 100 mm of the line separately for serves and non-serves. $PostHK = 1$ if Hawk-Eye review is active. Point controls include distance fixed effects (by 20 mm bins), speed (whether it is below or above the median), score, game, set, and an indicator if the point played is in the tie-break stage. Match controls include round and tournament tier.

error is 1.2 percentage points.

If we restrict attention to balls landing between 20 and 100 mm from the line, Table 3a shows statistically significant reductions in mistakes for both serves and non-serves. The closest calls look different. Table 3b restricts the sample to bounces within 20 mm, where perceptual difficulty is greatest. For serves in this range, the estimated mistake rate increases by 7.3 percentage points, or 22.9% relative to baseline. Although this estimate concerns a narrow subset of calls, it identifies where the review regime is associated with worse initial decisions.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Incorrect call (Serves)				Incorrect call (Non-serves)			
PostHK	-0.020** (0.008)	-0.018** (0.008)	-0.017** (0.008)	-0.017** (0.009)	-0.012 (0.009)	-0.017* (0.009)	-0.020** (0.009)	-0.020** (0.009)
Point controls		X	X	X		X	X	X
Match controls			X	X			X	X
Cluster level	Tournament				Tournament			
N	7,499	7,499	7,238	7,238	6,047	6,047	5,785	5,785
Baseline mean	.092	.092	.091	.091	.082	.082	.080	.080

Standard errors in parentheses

* $p < .10$, ** $p < .05$, *** $p < .01$

(a) For balls bouncing 20-100 mm away from the line.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Incorrect call (Serves)				Incorrect call (Non-serves)			
PostHK	0.063** (0.031)	0.063** (0.031)	0.073** (0.032)	0.073** (0.021)	-0.035 (0.031)	-0.037 (0.031)	-0.031 (0.033)	-0.031 (0.032)
Point controls		X	X	X		X	X	X
Match controls			X	X			X	X
Cluster level	Tournament				Tournament			
N	1,804	1,804	1,752	1,752	1,512	1,512	1,470	1,470
Baseline mean	.325	.325	.319	.319	.333	.333	.325	.325

Standard errors in parentheses

* $p < .10$, ** $p < .05$, *** $p < .01$

(b) For balls bouncing within 20 mm of the line.

Table 3: OLS regressions of umpire mistakes (separately for serves and non-serves). $PostHK = 1$ if the Hawk-Eye review system is active. Point controls include distance fixed effects (by 20 mm bins), speed (whether it is below or above the median), score, game, set, and an indicator if the point played is in the tie-break stage. Match controls include round and tournament tier.

5 Error Composition and Review Performance

Initial accuracy alone is insufficient to evaluate review. In our descriptive comparison, remaining errors decline for both error types, but correction offsets worse initial decisions for incorrect in calls and reinforces better initial decisions for incorrect out calls. These different paths imply different conditions under which review reduces organizational costs.

Preventing an incorrect out call avoids interrupting a rally. Correcting it restores the line classification but may still require a replay or a discretionary decision about who receives the point. Our measures of remaining errors concern final line classifications, not whether points were correctly allocated or uninterrupted play was restored.

5.1 A Simple Cost Comparison

For each error type $j \in \{\text{in}, \text{out}\}$, let E_j^0 and E_j^1 be the shares of all calls initially wrong without and with review. Incorrect in calls classify an out ball as in, while incorrect out calls classify an in ball as out. Let C_j be the share of all calls corrected under review. As elsewhere in the paper, we treat Hawk-Eye as ground truth and assume review is correctly applied to the relevant bounce, so it introduces no classification errors. The share left wrong under review is therefore $E_j^1 - C_j$. Without review, initial mistakes remain uncorrected by that system.

Consider an additive comparison in which w_j is the organizational loss from an error left uncorrected and k_j is the cost of remedying a corrected error. These costs can differ across error types. Let H be the additional cost per call of providing and conducting review relative to the no-review regime. This covers all reviews, including reviews that uphold calls but excluding remediation costs already counted in k_j . Using the same loss weights in both regimes, review reduces organizational costs when

$$\sum_j w_j(E_j^0 - E_j^1) + \sum_j C_j(w_j - k_j) > H. \quad (2)$$

The first term values differences in initial errors, including changes in their composition. The second values correction because each corrected mistake removes loss w_j but incurs remediation cost k_j . Their sum is the amount available to cover review costs. This comparison does not require a model of attention or a particular explanation of the human response. Umpires' private costs of reversal need not coincide with organizational losses.

To isolate the role of error composition, let $\eta_j = C_j/E_j^1$ be the fraction of initial mistakes of type j corrected under review. Excluding common review costs, the expected loss per initial mistake is

$$\ell_j = (1 - \eta_j)w_j + \eta_j k_j. \quad (3)$$

Holding total initial errors, correction fractions, unit costs, and common overhead fixed, replacing an incorrect out call with an incorrect in call reduces loss when $\ell_{\text{in}} < \ell_{\text{out}}$. Reallocation need not therefore be an organizational distortion. The comparison applies to both human and algorithmic review, although our evidence does not compare AI with human review or establish that review is preferable to full automation.

5.2 Errors, Correction, and Organizational Costs

Table 4 traces both error types from the initial call through correction for balls landing within 100 mm of the line. It uses initial and remaining error counts from the same sample, with corrected errors given by their difference. Unlike the regression estimates above, these are unadjusted descriptive shares among all calls in each regime.

Quantity	Incorrect in	Incorrect out	Total
Without review: initial errors, E_j^0	5.61	8.28	13.88
With review: initial errors, E_j^1	6.19	6.27	12.47
With review: corrected errors, C_j	2.09	1.35	3.45
With review: remaining errors, $E_j^1 - C_j$	4.10	4.92	9.02
Reduction in remaining errors, $E_j^0 - E_j^1 + C_j$	1.50	3.36	4.86
Initial mistakes corrected, C_j/E_j^1 (%)	33.8	21.6	27.6

Table 4: Error accounting for calls within 100 mm of the line.

Notes: Entries are errors per 100 calls, except the last row, which reports the percentage of initial mistakes corrected. The sample contains 2,622 calls without review and 14,190 with review, with each call given equal weight and calls exactly 100 mm from the line excluded. Initial calls precede any review. Corrected errors equal initial minus remaining errors under the ground-truth benchmark. All error shares use all calls in the corresponding regime as their denominator. The correction fractions instead use initial mistakes of the indicated type, with their total weighted by initial error shares. Entries are rounded separately, and cost thresholds use unrounded counts.

Remaining errors fall for both types, but initial incorrect in calls increase while initial incorrect out calls decrease. The reduction in remaining errors combines these initial differences with subsequent correction.

We can now translate these outcomes into costs. Applying Table 4 to equation 2, the condition per 100 calls is approximately

$$1.50w_{\text{in}} + 3.36w_{\text{out}} - 2.09k_{\text{in}} - 1.35k_{\text{out}} > 100H. \quad (4)$$

The first two terms value fewer remaining errors, while the next two subtract remediation costs. The different paths to lower remaining errors imply different conditions for each category to contribute a saving.

For incorrect in calls, correction must compensate for the deterioration in initial decisions. Review corrects 2.09 mistakes but leaves only 1.50 fewer errors per 100 calls, giving a contribution of $1.50w_{\text{in}} - 2.09k_{\text{in}}$. Using the unrounded counts, this contribution is positive before common review costs when $k_{\text{in}} < 0.72w_{\text{in}}$. Remediation must therefore cost less than about 72% of the loss from leaving an incorrect in call uncorrected for this category to generate savings.

For incorrect out calls, fewer initial mistakes and subsequent correction reinforce one another. The 1.35 corrections accompany 3.36 fewer remaining errors, giving a contribution of $3.36w_{\text{out}} - 1.35k_{\text{out}}$. This contribution is positive before common review costs when $k_{\text{out}} < 2.48w_{\text{out}}$. This does not mean that an individual correction is worthwhile when its remediation cost exceeds the loss it removes. The condition combines the descriptive decline in initial mistakes, which requires no subsequent correction, with the value of correcting other mistakes.

Neither error category must pay for itself separately. A saving in one can offset a loss in the other, but the combined saving must cover $100H$. If loss and remediation costs are each one unit for both error types, the contribution from incorrect in calls is negative, but the saving from incorrect out calls more than offsets it. This leaves about 1.42 loss units per 100 calls before common review costs.

Using the unrounded counts, initial errors decline by 1.42 per 100 calls and correction removes another 3.45, giving 4.86 fewer remaining errors. The difference in initial errors accounts arithmetically for about 29% of this unadjusted reduction, but this decomposition does not identify an attention effect or separate causal effects of changed initial decisions and correction.

5.3 An Illustration Using System Costs

For the cost illustration, suppose loss w per uncorrected error and remediation cost k are common across error types. Let K be the system charge over N calls within 100 mm of the line. Considering that charge alone sets $H = K/N$. Equation 4 gives the minimum loss per uncorrected error needed to cover it:

$$w_{\min} \approx \frac{K/N + 0.03446k}{0.04862}. \quad (5)$$

Review generates a net saving when $w > w_{\min}$. The loss w can include costs from disputes or reduced participation, rather than simply prize money transferred between players.

Free remediation ($k = 0$) allows savings from the full reduction in remaining errors. If remediation costs as much as an uncorrected error ($k = w$), correction benefits and remediation costs cancel. Savings then depend entirely on the descriptive decline in initial

errors. The threshold is approximately $w > K/(0.01416N)$. If that decline were zero, review could not cover a positive system charge under this benchmark.

A United States Tennis Association (USTA) report gives an annual PlayReplay charge of approximately \$3,200 per court at Edmond Center Court (Schluter 2026). For Hawk-Eye during our study period, Fischetti (2007) reports \$40,000–\$50,000 per court-week for setup and operation, in 2007 dollars. These are court-level charges, not prices per review.

Table 5 holds the observed error changes fixed and assumes 25 calls within 100 mm per match. Increasing annual use from 200 to 1,000 matches spreads the same PlayReplay charge over five times as many calls. This reduces the required loss per error fivefold.

Illustrative utilization	Minimum w if $k = 0$	Minimum w if $k = w$
PlayReplay: 200 matches per court-year	\$13.16	\$45.20
PlayReplay: 1,000 matches per court-year	\$2.63	\$9.04
Historical Hawk-Eye: 25 matches per court-week	\$1,316–\$1,645	\$4,520–\$5,650

Table 5: Error-loss thresholds for covering reported system charges.

Notes: Thresholds are dollars per remaining error, calculated from the unrounded counts in Table 4. Assumed utilization gives $N = 5,000$, 25,000, or 625 calls. The table covers the reported system bill alone, excluding additional review costs and other benefits. Hawk-Eye prices remain in 2007 dollars. Applying professional umpires’ error changes to another technology or to amateur players is an extrapolation, not an estimated adoption effect.

At 200 matches per court-year, the illustration gives about 243 fewer remaining errors annually. With free remediation, a loss of about \$13.16 per remaining error covers the \$3,200 charge. When $k = w$, only the roughly 71 fewer initial errors generate net savings. The threshold therefore rises to \$45.20. Additional review costs raise these thresholds, while staffing savings or other benefits lower them.

Fewer remaining errors do not necessarily imply lower organizational costs. In our comparison, correction must compensate for an increase in initial incorrect in calls, so this category contributes savings before common review costs only when remediation costs less than 0.72 times the loss from an uncorrected error. For incorrect out calls, fewer initial mistakes create additional savings, raising the corresponding threshold to 2.48. The economic distinction is between correction that offsets worse initial decisions and correction that accompanies better ones. The two contributions must jointly cover the cost of providing review. Greater utilization spreads fixed system costs across more calls, but cannot make review worthwhile if those combined contributions are negative. Evaluating review therefore requires valuing the changes in error composition and the consequences of correction, not simply counting successful reversals.

6 Conclusion

Algorithmic review can change the decisions it is intended to correct. After Hawk-Eye review began, umpires became more likely to call marginal balls in. Errors were lower just inside the line and higher just outside. For the closest outside bin, the mistake rate was 8.6 percentage points higher under review. This exceeds the average difference in the first two inside bins by 14.5 percentage points. The main pattern associated with review is therefore a reallocation of errors, rather than a uniform improvement or deterioration in initial accuracy.

The within-match estimates are consistent with adaptation to correction, but endogenous challenges and reversals prevent them from identifying the mechanism behind the adjustments. Following an in-to-out reversal, subsequent close calls shift toward out. Following an out-to-in reversal, they shift toward in. In the linear specification, responses are larger after out-to-in reversals, where the initial call stopped play and correction can require a discretionary remedy. Reversals are followed by greater adjustment than upheld challenges of the same initial-call direction, with joint equality rejected ($p = 0.004$).

The organizational value of these patterns depends on both changes in initial errors and the costs of correction. Under our cost comparison, incorrect in and incorrect out calls contribute savings before common review costs when remediation costs are below about 0.72 and 2.48 times the loss from an uncorrected error of the corresponding type, respectively. The second threshold reflects fewer initial mistakes as well as correction, not a benefit from correcting any one error at a cost greater than the loss it removes.

The combined savings must cover the costs of providing review, but neither category needs to pay for itself separately. Greater utilization spreads system costs over more calls in our illustration. Yet utilization and final accuracy alone do not determine whether review is worthwhile. Evaluation also requires measuring how initial errors change, which errors are corrected, and what costs remain after correction.

References

- Abramitzky, Ran, Liran Einav, Shimon Kolkowitz, and Roy Mill (2012). “On the Optimality of Line Call Challenges in Professional Tennis”. *International Economic Review* 53.3, pp. 939–964.
- Agarwal, Nikhil, Alex Moehring, Pranav Rajpurkar, and Tobias Salz (2023). “Combining Human Expertise with Artificial Intelligence: Experimental Evidence from Radiology”. *Working Paper 31422, National Bureau of Economic Research*.
- Agarwal, Nikhil, Alex Moehring, and Alexander Wolitzky (2025). *Designing Human-AI Collaboration: A Sufficient-Statistic Approach*. Working Paper 33949. Revised August 2026. National Bureau of Economic Research. DOI: [10.3386/w33949](https://doi.org/10.3386/w33949).
- Albright, Alex (2024). “The Hidden Effects of Algorithmic Recommendations”. *Working paper*.
- Almog, David (2025a). “AI Recommendations and Non-instrumental Image Concerns”. *SSRN working paper 5232232*.
- (2025b). “Barriers to AI Adoption: Image Concerns at Work”. *arXiv: 2511.18582*.
- Almog, David, Lucas Lippman, and Daniel Martin (2026). “When an AI Judges Your Work: The Hidden Costs of Algorithmic Assessment”. *arXiv: 2603.02076*.
- Athey, Susan, Kevin Bryan, and Joshua Gans (2020). “The Allocation of Decision Authority to Human and Artificial Intelligence”. *AEA Papers and Proceedings* 110, pp. 80–84.
- Avery, Mallory, Andreas Leibbrandt, and Joseph Vecchi (2023). “Does Artificial Intelligence Help or Hurt Gender Diversity? Evidence from Two Field Experiments on Recruitment in Tech”. *Working paper*.
- Balakrishnan, Maya, Kris Johnson Ferreira, and Jordan Tong (2026). “Human-Algorithm Collaboration with Private Information: Naïve Advice-Weighting Behavior and Mitigation”. *Management Science* 72.1, pp. 265–284.
- Bursztyn, Leonardo and Robert Jensen (2017). “Social Image and Economic Behavior in the Field: Identifying, Understanding, and Shaping Social Pressure”. *Annual Review of Economics* 9.1, pp. 131–53.
- Caplin, Andrew, David Deming, Shangwen Li, Daniel Martin, Philip Marx, Ben Weidmann, and Kadachi Ye (2025). “The ABC’s of Who Benefits from Working with AI: Ability, Beliefs, and Calibration”. *Management Science*.
- Chan, David C., Matthew Gentzkow, and Chuan Yu (2022). “Selection with Variation in Diagnostic Skill: Evidence from Radiologists”. *The Quarterly Journal of Economics* 137.2, pp. 729–783.
- Cho, Sungwoo (2024). “The Effect of Robot Assistance on Skills”. *Working paper*.
- Ely, Jeffrey, Romain Gauriot, and Lionel Page (2017). “Do Agents Maximise? Risk Taking on First and Second Serves in Tennis”. *Journal of Economic Psychology* 63, pp. 135–142.

- Fischetti, Mark (July 1, 2007). *Working Knowledge: Tennis Hawkeye—In or Out?* Scientific American. URL: <https://www.scientificamerican.com/article/in-or-out/> (visited on 10/15/2026).
- Gauriot, Romain and Lionel Page (2019). “Does Success Breed Success? a Quasi-Experiment on Strategic Momentum in Dynamic Contests”. *The Economic Journal* 129.624, pp. 3107–3136.
- Gauriot, Romain, Lionel Page, and John Wooders (2023). “Expertise, Gender, and Equilibrium Play”. *Quantitative Economics* 14.3, pp. 981–1020.
- Goldsmith-Pinkham, Paul, Peter Hull, and Michal Kolesár (2024). “Contamination Bias in Linear Regressions”. *American Economic Review* 114.12, pp. 4015–51.
- Gosnell, Greer K., John A. List, and Robert D. Metcalfe (2020). “The Impact of Management Practices on Employee Productivity: A Field Experiment with Airline Captains”. *Journal of Political Economy* 128.4, pp. 1195–1233.
- Grimon, Marie-Pascale and Christopher Mills (2025). “Better Together? A Field Experiment on Human-Algorithm Interaction in Child Protection”. *Working paper*.
- Hoffman, Mitchell, Lisa Kahn, and Danielle Li (2018). “Discretion in Hiring”. *The Quarterly Journal of Economics* 133.2, pp. 765–800.
- Hoong, Ruru and Bnaya Dreyfuss (2026). “Calibrated Coarsening: Designing Information for AI-Assisted Decisions”. *MIT Sloan Working Paper No. 7407-25*.
- Huelsen, Tobias, Vitalijs Jascisens, Lars Roemheld, and Tobias Werner (2024). “Human-Machine Interactions in Pricing: Evidence from Two Large-Scale Field Experiments”. *Working paper*.
- Ide, Enrique and Eduard Talamàs (2025). “Artificial Intelligence in the Knowledge Economy”. *Journal of Political Economy* 133.12, pp. 3762–3800.
- Kawaguchi, Kohei (2020). “When Will Workers Follow an Algorithm? A Field Experiment with a Retail Business”. *Management Science* 67.3, pp. 1670–1695.
- Kleinberg, Jon, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan (2018). “Human Decisions and Machine Predictions”. *The Quarterly Journal of Economics* 133.1, pp. 237–293.
- Kovach, Matthew, Daniel Martin, and Gerelt Tserenjigmid (2026). “Learning from an Unknown DGP: Experimental Evidence on Belief Updating with AI Recommendations”. *Working paper*.
- Lee, Kichang, Gyeongmin Han, Sungmin Lee, and JeongGil Ko (2026). “When the Strike Zone Becomes Algorithmic: Umpire Judgment and Player Challenge Decisions under AI Review”. *arXiv:2609.25525*.
- List, John A. (2020). “Non est Disputandum de Generalizability? A Glimpse into The External Validity Trial”. *NBER Working Paper No. 27535*.
- McLaughlin, Bryce and Jann Spiess (2024). “Algorithmic Assistance with Recommendation-Dependent Preferences”. *arXiv Preprint arXiv:2208.07626*.

- Michopoulos, Constantine, Peter Newberry, Seth Smith, Hannah Waldfogel, and Sarah Waldfogel (2026). “Avoiding Being Caught: Strike Calling in the ABS Era”. *SSRN Working Paper 7242822*.
- Nagin, Daniel S., James B. Rebitzer, Seth Sanders, and Lowell J. Taylor (2002). “Monitoring, Motivation, and Management: The Determinants of Opportunistic Behavior in a Field Experiment”. *American Economic Review* 92.4, pp. 850–873.
- Olken, Benjamin A. (2007). “Monitoring Corruption: Evidence from a Field Experiment in Indonesia”. *Journal of Political Economy* 115.2, pp. 200–249.
- Raghu, Maithra, Katy Blumer, Greg Corrado, Jon Kleinberg, Ziad Obermeyer, and Sendhil Mullainathan (2019). “The Algorithmic Automation Problem: Prediction, Triage, and Human Effort”. *Working paper*.
- Rambachan, Ashesh (2024). “Identifying Prediction Mistakes in Observational Data”. *The Quarterly Journal of Economics* 139.3, pp. 1665–1711.
- Schluter, Blair (Mar. 25, 2026). *PlayReplay Electronic Line Calling Brings Precision, Innovation to Edmond Center Court*. United States Tennis Association. URL: <https://www.usta.com/en/home/stay-current/missourivalley/oklahoma/edmond-center-court-playreplay-electronic-line-calling-system.html> (visited on 10/15/2026).
- Whitney, David, Nicole Wurnitsch, Byron Hontiveros, and Elizabeth Louie (2008). “Perceptual Mislocalization of Bouncing Balls by Professional Tennis Referees”. *Current Biology* 18, R947–9.
- Zou, Eric Yongchen (2021). “Unwatched Pollution: The Effect of Intermittent Monitoring on Air Quality”. *American Economic Review* 111.7, pp. 2101–2126.

A Data Management

A.1 Exclusion of Clay Tournaments

Clay surfaces present some particular difficulties for the Hawk-Eye system, so Hawk-Eye review has not been adopted in clay tournaments. One issue is that the surface is continuously changing during a match, so the accuracy of Hawk-Eye decreases and requires constant re-calibration to operate at its best.²² Also, Hawk-Eye review on clay courts might introduce controversies where the ball’s mark and Hawk-Eye might disagree. As noted in The Guardian, Lars Graf, one of the ATP’s most experienced officials, explained the decision not to use Hawk-Eye on clay: “We decided not to use Hawk-Eye on clay because it might not agree with the mark the umpire is pointing at.”²³

Because Hawk-Eye was not adopted in clay tournaments during our sample period, they may appear as an ideal control group for comparing umpire reactions in tournaments with Hawk-Eye review versus those on clay. However, there are several factors that led us to exclude clay tournaments from our analysis. For one, we are unable to determine from our data set which calls were corrected after examining the clay marks. In addition, before the introduction of Hawk-Eye, players were allowed to request the chair umpire to review the ball bounce marks on the surface, and changing the call based on that is permitted, so a form of review system was available.

A.2 Merging Algorithm

Our merging algorithm was designed to identify the highest number of the 144 challenges (derived from the 43 matches with video replays) within the Hawk-Eye Base data set, while keeping the number of false positives (challenges matched into the wrong point) as low as possible. We used standard variables, such as set, game, score, distance difference, player hitting the ball, and whether it was a tie-break, in order to minimize overfitting.

Our final merging algorithm comprises eight iterations, where we start using stricter criteria and gradually relax them in subsequent iterations. Some iterations, like 5 and 6, focus on very specific situations of the match (tie-breaks), and while they do not seem useful in the testing sample, they might prove useful in a bigger sample.

The algorithm merged 143 out of the 144 challenges, achieving a merge rate of 99.3%. Out of these, 136 were merged to the correct point, as validated through video auditing replays, resulting in an accuracy rate of 95.1%. We will now elaborate on the specific criteria

²²<https://www.perfect-tennis.com/why-there-is-no-hawk-eye-on-clay/>

²³<https://www.theguardian.com/lifeandstyle/2009/jun/27/tennis-hawk-eye>

Iteration	Correct Merge	False Positives	Efficacy	Challenges Left
1	98	2	98%	44
2	25	2	93%	17
3	5	0	100%	12
4	1	0	100%	11
5	0	0	—	11
6	0	0	—	11
7	5	0	100%	6
8	2	3	40%	1

Table A.1: Performance for each iteration of the merging algorithm. Out of the 144 challenges audited, only one remained unmerged.

used in each of the eight iterations, followed by Table A.1, which summarizes the number of challenges successfully merged (and false positives) achieved in each iteration.

Merging criteria for each iteration:

Iteration 1. Same set, game, score, and player (with distance difference <35 mm).

Iteration 2. Same set, game, and player (with distance difference <15 mm).

Iteration 3. Same set, score, and player (with distance difference <15 mm).

Iteration 4. Same game, score, and player (with distance difference <10 mm).

Iteration 5. For tie-breaks: Same game and score (with distance difference <35 mm).

Iteration 6. For tie-breaks: Same game. If multiple, pick closest in distance (with distance difference <15 mm).

Iteration 7. Same set, game and player. If multiple, pick closest point in score and then in distance (with distance difference <35 mm).

Iteration 8. Same set and player. If multiple, pick closest point in distance (with distance difference <35 mm).

A.3 Identification of Incorrect Calls

The Hawk-Eye Base dataset records the precise location of every ball bounce and identifies the winner of each point, but does not directly report the umpires’ calls. We therefore

reconstruct the original calls using four criteria. The first two identify points at which the umpire incorrectly called a ball in, while the final two identify points at which the umpire incorrectly called a ball out. During the period when players could challenge calls, we first recover the umpire’s original call for each successful challenge and then assess whether any of the four criteria are met.

Criterion 1. A player’s stroke is out and wins the point.

We selected strokes where the first player to hit a ball out in a given point also won the point.

Criterion 2. The point continued after an out.

To complement Criterion 1, we also identify points in which a player hit a ball out, play continued, and that same player ultimately lost the point. Specifically, we examine instances in which at least three additional strokes were recorded after the ball landed out.²⁴

Criterion 3. All the strokes of a point are in, and the player hitting last loses.

We identify strokes where the last player to hit the ball in did not win the point.

Criterion 4. Observing a second serve after the first serve was in.

This criterion identifies first serves that landed in but were incorrectly called out, leaving the point without a recorded winner and prompting a second serve.²⁵

²⁴Video validation indicated that a threshold of three additional strokes is necessary. A threshold of two is too lenient and captures multiple cases in which the call was correct but the players continued hitting the ball after the point had ended.

²⁵For this criterion, we restrict the analysis to serves landing within 40 mm of the line because the criterion becomes less accurate beyond this threshold. This interval nevertheless captures the cases most relevant to the paper’s main results.

B External Validity

To provide a more complete assessment of external validity, we build around the four transparency conditions given as an *onus probandi* in List (2020).

B.1 Selection

This condition concerns the representativeness of the studied group compared to the underlying population, in particular in “observables that might impact preferences, beliefs, or individual constraints.” The evidence concerns experienced officials making rapid perceptual judgments under public review, but does not establish how responses vary with expertise or professional identity.

B.2 Attrition

This condition asks whether motivational or incentive differences affect continued participation. A leading official at the time of Hawk-Eye’s introduction confirmed that umpire training and assessment, the pool of umpires, and the positioning and instructions given to line judges remained stable during our study period. In other settings, introducing algorithmic review and public correction could cause selective attrition. Workers who dislike correction most may be especially likely to leave, so estimates among those who remain could understate the response in the original workforce.

B.3 Naturalness

This condition addresses the artificiality of a setting, such as the nature of the choice task. Our study observes routine officiating decisions with consequences for live matches. Corrections are visible to players, spectators, and television audiences, but the design does not establish how responses vary with the visibility of correction.

A final factor is that Hawk-Eye is nearly perfect at the task we study. Our results are therefore most likely to extend to settings where the reviewing algorithm is highly accurate when it intervenes. Workers may respond differently to less reliable technology because they can reasonably contest its judgment or attribute reversals to algorithmic error, but whether reliability moderates responses to algorithmic correction remains an open question.

B.4 Scaling

This condition concerns whether responses persist as an intervention scales. During our study period, review was implemented at top tournaments, so umpires and their professional peers were broadly exposed to the same regime. Greater familiarity could normalize correction and weaken its salience, or repeated exposure could strengthen adaptation to the protocol. The larger threshold-shift pattern in the later subperiod is consistent with the latter possibility, but the subperiod analysis cannot separate exposure from other changes over time.

C Additional Tables and Figures

Tournament	Start Month	Category	Court Type	Number of Matches			
				2005	2006	2007	2008
Marseille	Feb	International (250)	Hard				25
Rotterdam	Feb	International Gold (500)	Hard			15	26
Dubai	Mar	International Gold (500)	Hard				19
Las Vegas	Mar	International (250)	Hard				19
Indian Wells	Mar	Masters (1000)	Hard	17		27	
Miami	Mar	Masters (1000)	Hard	15	18	26	
Queen’s Club	Jun	International (250)	Grass	13	19	18	
UCLA	Jul	International (250)	Hard			23	
Indianapolis	Jul	International (250)	Hard		7	23	
Washington	Jul	International (250)	Hard			21	
Montreal	Aug	Masters (1000)	Hard	20		27	
Toronto	Aug	Masters (1000)	Hard		25		
Cincinnati	Aug	Masters (1000)	Hard	11	19	24	
New Haven	Aug	International (250)	Hard			18	
Beijing	Sep	International (250)	Hard			20	
Kremlin Cup	Oct	International (250)	Carpet06 Hard07		8	15	
Madrid	Oct	Masters (1000)	Hard		30	30	
Basel	Oct	International (250)	Hard			12	
Paris Masters	Oct	Masters (1000)	Carpet06 Hard07		31	33	
Shanghai	Nov	Masters (1000)	Carpet05 Hard06-07	14	15	15	

Table C.1: Information on the 35 tournament editions in the consolidated data set. The numbers in each cell indicate the matches available in our data set. The color represents the review regime, with blue indicating tournament editions without Hawk-Eye review and red indicating tournament editions with Hawk-Eye review.

	Estimate	Clustered SE	Wild-bootstrap p	95% CI
Closest outside bin $[-20, 0)$	0.086	0.034	0.052	$[-0.001, 0.163]$
First inside bin $[0, 20)$	-0.041	0.034	0.256	$[-0.114, 0.040]$
Second inside bin $[20, 40)$	-0.077	0.027	0.028	$[-0.136, -0.013]$
<i>Cross-boundary contrasts</i>				
Closest outside – first inside	0.127	0.060	0.076	$[-0.019, 0.259]$
Closest outside – second inside	0.163	0.049	0.018	$[0.048, 0.270]$
Closest outside – average of first two inside	0.145	0.052	0.034	$[0.017, 0.259]$

Table C.2: Wild-Cluster-Bootstrap Inference for the Main Distance-Bin Estimates. The first three rows report estimated changes in the incorrect-call rate after the introduction of Hawk-Eye review for the indicated 20 mm distance bins. The remaining rows report contrasts between these estimates. Standard errors are clustered at the tournament-edition level. Wild-cluster-bootstrap inference is based on 9,999 replications clustered at the tournament-edition level. For the final row, the reported estimate compares the closest outside bin with the average of the first two inside bins. The corresponding wild-bootstrap test is implemented as $2\beta_{[-20,0)} - \beta_{[0,20)} - \beta_{[20,40)} = 0$. Its confidence set is divided by two so that it is expressed in the same units as the reported contrast.

Specification	Cross-boundary contrast	Clustered SE	Wild p	Tournament editions
<i>Panel A. Same-event specification</i>				
Full sample	0.145	0.052	0.034	35
Same events + event fixed effects	0.102	0.063	0.152	17
<i>Panel B. Leave-one-event-out estimates</i>				
Omitted event	Contrast	Clustered SE	95% CI	
Canada Open	0.079	0.081	-0.097	0.255
Cincinnati	0.110	0.064	-0.030	0.250
Indian Wells	0.125	0.068	-0.022	0.272
Miami	0.099	0.072	-0.058	0.257
Queen's Club	0.103	0.059	-0.026	0.232
Shanghai	0.077	0.067	-0.069	0.222

Table C.3: Same-Event and Leave-One-Event-Out Robustness. The cross-boundary contrast is defined as the estimated change in the incorrect-call rate for the closest outside bin, $[-20, 0)$, minus the average estimated change for the first two inside bins, $[0, 20)$ and $[20, 40)$. The same-event specification restricts the sample to named events observed under both review regimes and includes named-event fixed effects. Toronto and Montreal are treated as a single event, the Canada Open, because the tournament alternates between the two host cities. Standard errors are clustered at the tournament-edition level. Wild-cluster-bootstrap inference is based on 9,999 replications clustered at the tournament-edition level. Panel B re-estimates the same-event specification after omitting each named event in turn. The leave-one-event-out estimates are reported to assess sensitivity to individual events rather than for separate hypothesis testing.

Specification	Cross-boundary contrast	Clustered SE	95% CI
Baseline	0.145	0.052	[0.040, 0.251]
+ Serve/non-serve control	0.144	0.052	[0.038, 0.250]
+ Flexible speed controls	0.144	0.052	[0.039, 0.250]
+ Hitter and receiver fixed effects	0.136	0.053	[0.028, 0.244]

Table C.4: Robustness to Shot and Player Composition. The cross-boundary contrast is defined as the estimated change in the incorrect-call rate for the closest outside bin, $[-20, 0)$, minus the average estimated change for the first two inside bins, $[0, 20)$ and $[20, 40)$. The baseline specification includes 20 mm distance-bin fixed effects, an above-median-speed indicator, score, game, set, tie-break status, round, and tournament tier. Controls are added sequentially across rows: the second row adds serve status, the third replaces the above-median-speed indicator with flexible controls for ball speed, and the fourth additionally includes hitter and receiver player fixed effects. Standard errors are clustered at the tournament-edition level.

Specification	Asymmetry	Clustered SE	<i>p</i> -value
<i>Panel A. Flexible effects of prior reversals</i>			
One prior reversal	0.125	0.061	0.040
Two prior reversals	0.178	0.102	0.081
Three or more prior reversals	0.156	0.158	0.324
Joint symmetry test	—	—	0.151
<i>Panel B. Linear cumulative-reversal specification</i>			
Per additional prior reversal	0.103	0.041	0.011

Table C.5: Asymmetry in Responses to Prior Reversals. The table tests whether the behavioral response following an out-to-in reversal is larger in absolute magnitude than the response following an in-to-out reversal. The asymmetry estimate is the sum of the coefficients on the corresponding out-to-in and in-to-out reversal variables. A positive estimate therefore indicates a larger response following an out-to-in reversal. Panel A allows separate effects after one, two, and three-or-more prior reversals. The joint symmetry test evaluates the null that the three corresponding asymmetry estimates are jointly equal to zero ($F(3, 567) = 1.78$). Panel B enters the cumulative number of reversals of each type linearly and reports the difference in response magnitude associated with one additional prior reversal. Standard errors are clustered at the match level.

Comparison	Estimate	Clustered SE	<i>p</i> -value
IN→OUT reversal minus upheld IN	-0.067	0.038	0.078
OUT→IN reversal minus upheld OUT	0.115	0.040	0.004
Joint equality with upheld challenges	—	—	0.004
Difference in absolute incremental effects	0.049	0.053	0.363

Table C.6: The table compares estimated associations of subsequent calls with prior reversals and with unsuccessful challenges involving the same initial-call direction. A negative estimate in the first row indicates an additional shift toward calling balls out following an IN-to-OUT reversal. A positive estimate in the second row indicates an additional shift toward calling balls in following an OUT-to-IN reversal. All four cumulative feedback variables are included simultaneously in the same within-match specification used in Section 3.3. Standard errors are clustered at the match level. The joint test evaluates whether the two reversal coefficients equal their corresponding upheld-challenge coefficients. The final row tests whether the incremental associations are equal in absolute magnitude.

	(1)	(2)	(3)	(4)
	Incorrect call			
PostHK	-0.014** (0.007)	-0.011 (0.007)	-0.011 (0.007)	-0.011 (0.008)
Point controls		X	X	X
Match controls			X	X
Cluster level				Tournament
N	16,812	16,812	16,335	16,335
Baseline mean	.139	.139	.136	.136
Standard errors in parentheses				
* $p < .10$, ** $p < .05$, *** $p < .01$				

Table C.7: OLS regressions of umpire mistakes for balls bouncing within 100 mm of the line. $PostHK = 1$ if Hawk-Eye review is active. Point controls include 20 mm distance-bin fixed effects, an above-median-speed indicator, score, game, set, and a tie-break indicator. Match controls include round and tournament tier.

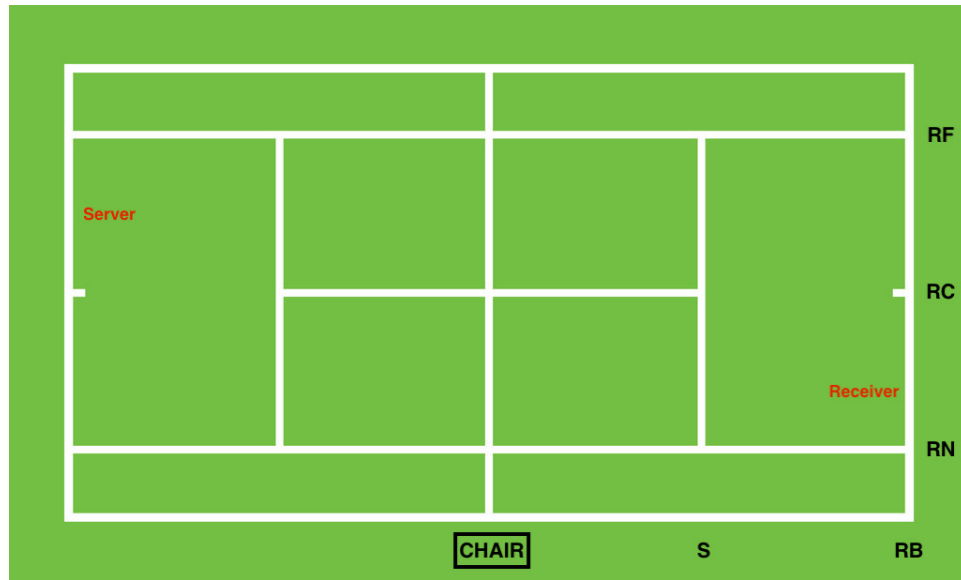


Figure C.1: The location of the umpiring crew: S = the serve-line umpire, RB = the right baseline umpire, RN = the right near long-line umpire, RC = the right center-line umpire, RF = the right far-side long-line umpire. The left side of the court has corresponding line umpires, except for the serve-line umpire, who moves to the left side when the right side player has the serve.

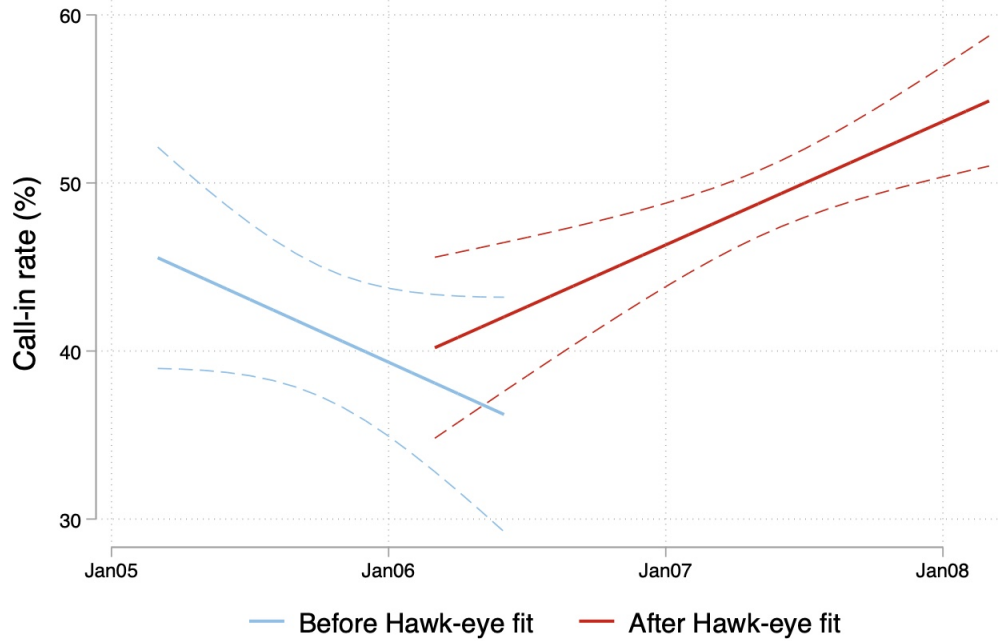


Figure C.2: Call-in rates before and after the introduction of Hawk-Eye review for balls landing within 20 mm of the line, regardless of which side of the line they actually bounced on. The solid lines report fitted values from a linear probability model in which the probability of calling a ball in is allowed to vary linearly with month, separately before and after Hawk-Eye review. The dashed lines indicate 95% confidence intervals. The regression controls for game, set, point, tiebreak, round, tournament tier, and shot speed. Standard errors are clustered at the match level.

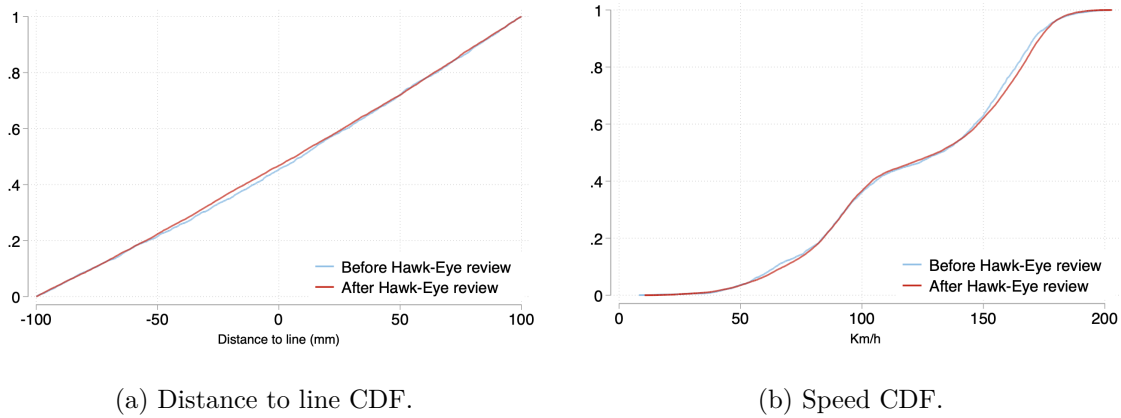


Figure C.3: Empirical distributions of signed distance from the line and ball speed before and after Hawk-Eye review.

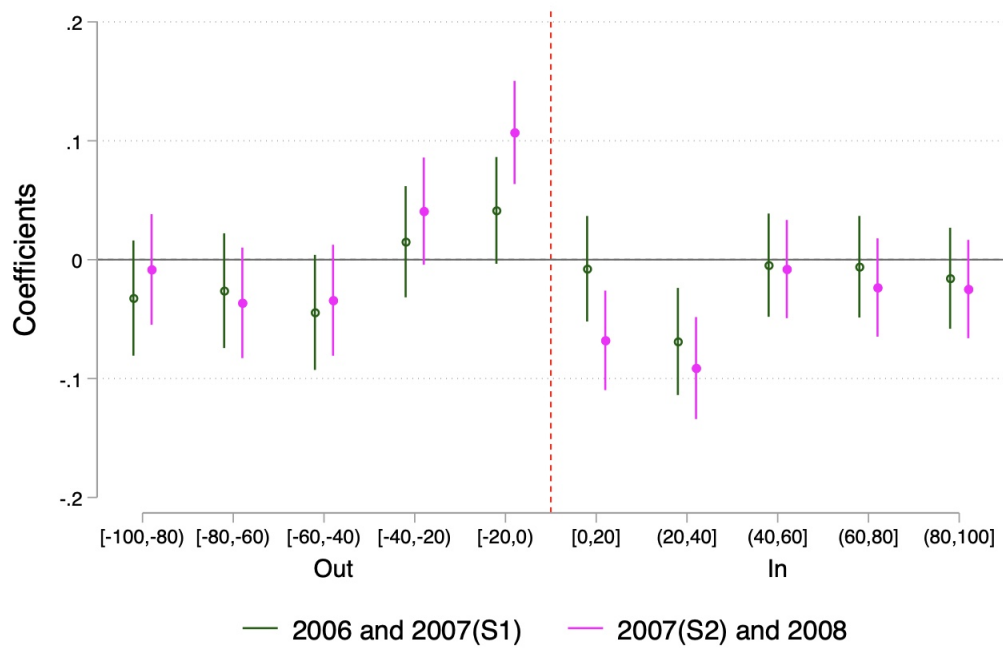


Figure C.4: Subperiod analysis of changes in the incorrect-call rate after the introduction of Hawk-Eye review, by proximity to the line. Matches with review are grouped into tournaments from 2006 through the first half of 2007 and tournaments from the second half of 2007 through 2008. Each dot represents the coefficient on the interaction between a distance-bin indicator and *PostHK*.

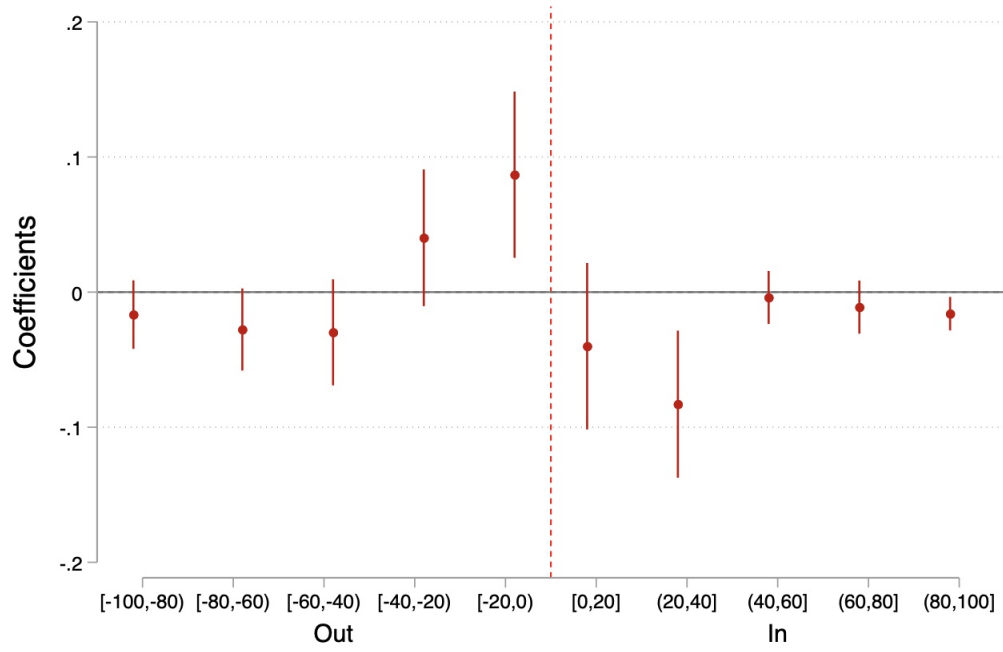


Figure C.5: An analog of Figure 2, using a one-treatment-at-a-time approach. We estimate the effect for each distance bin separately to avoid contamination across bins.

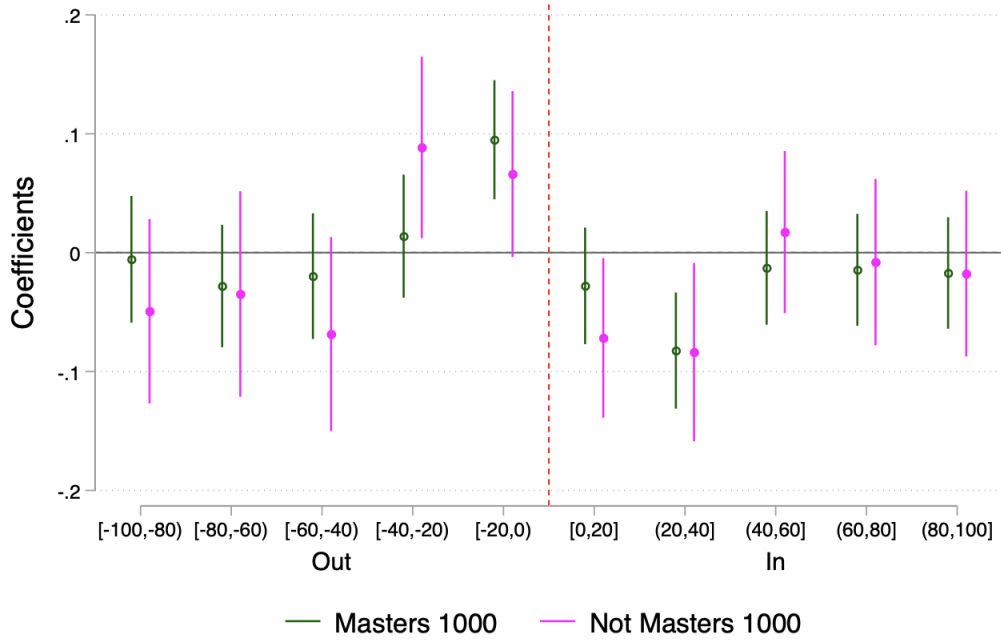


Figure C.6: Estimated change in umpire performance after the introduction of Hawk-Eye review, across 20 mm distance bins and by tournament category. Matches are grouped into higher-tier tournaments (Masters 1000) and lower-tier tournaments (ATP 500 and 250). Each dot represents the coefficient on the interaction between a distance-bin indicator and *PostHK*, estimated separately by tournament category.

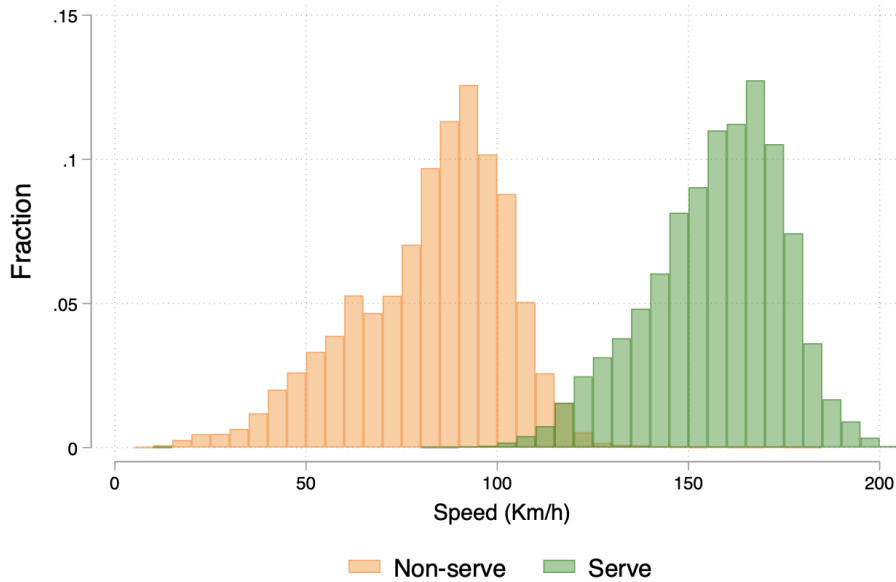
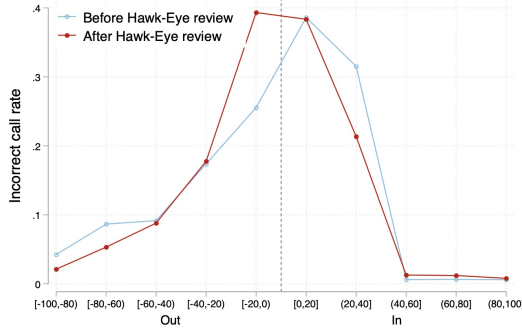
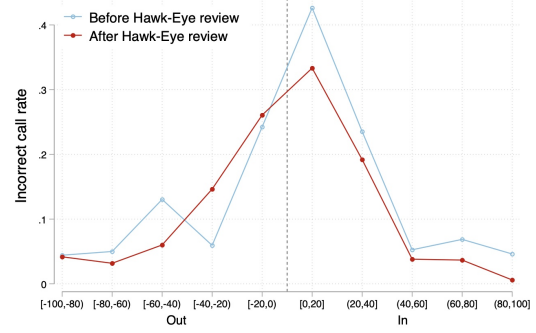


Figure C.7: Speed distribution for all the balls that bounced within 100 mm of the line, separated by non-serves and serves.



(a) Serves.



(b) Non-serves.

Figure C.8: Incorrect call rates by proximity to the line (separately for serves and non-serves). Each dot is the rate of incorrect calls for a bin of 20 mm, and dots to the left of the dashed line represent bins out of bounds, and those to the right of the dashed line represent bins in bounds.

D Pooled Call and Correction Statistics

This appendix reports the aggregate inputs needed to reconstruct initial-call probabilities and correction exposure for balls landing less than 100 mm from the line. The sample and weighting match Table 4. Every call receives equal weight, and calls exactly 100 mm from the line are excluded. Table D.1 crosses the ball’s actual location with the initial human call, before any algorithmic correction. The final column counts initially incorrect calls corrected by review, not all challenges or reviews.

Actual location	Initial call	Without review	With review	Corrected with review
In	In	1,216	6,664	0
In	Out	217	890	192
Out	In	147	879	297
Out	Out	1,042	5,757	0
Total		2,622	14,190	489

Table D.1: Initial calls by actual location and subsequent correction.

Notes: Entries are counts, not percentages. Actual location is determined by Hawk-Eye. Corrections are the reduction in mistakes between the initial and final line classifications. The zero entries for initially correct calls indicate that there is no initial error to correct, not that such calls are never reviewed. The table does not report reviews that uphold a call or their duration. Without algorithmic review, no initial mistakes are corrected by that system.

Table D.2 reports the resulting conditional probabilities. Let $r = 0$ denote no review and $r = 1$ review. The quantities p_r and q_r condition on the initial call, whereas m_r is the actual in share and λ_r is the initial call-in share. For example, $p_0 = 1,216/(1,216 + 147)$ and $q_0 = 217/(217 + 1,042)$. Correction exposure instead conditions on an initial mistake and its true state. Thus, $\eta_I = 192/890$ concerns balls that were in but called out, while $\eta_O = 297/879$ concerns balls that were out but called in. The subscripts I and O here index the ball’s true state, not the direction of the incorrect call used in Section 5.

Quantity and conditioning	Without review	With review
$p_r = \text{Pr}(\text{ball in} \mid \text{initial call in})$	0.89214967	0.88346812
$q_r = \text{Pr}(\text{ball in} \mid \text{initial call out})$	0.17235902	0.13389499
$m_r = \text{Pr}(\text{ball in})$	0.54652937	0.53234672
$\lambda_r = \text{Pr}(\text{initial call in})$	0.51983219	0.53157153
$\eta_I = \text{Pr}(\text{corrected} \mid \text{ball in, initial call out})$	0	0.21573034
$\eta_O = \text{Pr}(\text{corrected} \mid \text{ball out, initial call in})$	0	0.33788396

Table D.2: Pooled probabilities implied by Table D.1.

Notes: Probabilities are rounded to eight decimal places. Calculations should use the integer counts in Table D.1, rather than rounded entries. Under the ground-truth benchmark, correction exposure measures successful correction of an initially incorrect line classification, not the probability that an arbitrary call is challenged.

The integer counts reproduce the pooled moments without access to individual call records, but do not preserve dependence within matches or tournament editions. Clustered uncertainty therefore requires these counts separately by resampling unit, or the corresponding bootstrap moment draws. In particular, a tournament-edition bootstrap requires the four initial state–call counts and two corrected-error counts for each edition, together with its review status, rather than pooled totals alone.