Play it again! A Natural Experiment on Definitivity Avoidance[*]

Thomas Bassetti[†]
Stefano Bonini[‡]
Fausto Pacicco[§]
Filippo Pavesi[¶]

Abstract

In this paper we introduce *Definitivity Avoidance*, a behavioral bias that induces individuals to inefficiently shy away from choices that involve a final judgment. We model its effects and explore how the introduction of explicit exposure mechanisms can contribute to attenuate them. Using a unique natural experiment - the introduction of a technology assisted review system in professional tennis - we test the model predictions and confirm the relevance of this behavioral bias in a competitive setting. Possible instances of definitivity avoidance can be identified in multiple contexts such as debt roll-over decisions, inefficient asset allocations, court rulings, and child adoptions. The broad applicability of our model carries relevant policy implications as it provides a conceptual framework for the design of institutions to alleviate the welfare costs associated with definitivity avoidance.

JEL Classification Code: D81, D91, C93
Keyword: Prospect Theory; Natural experiment; Uncertainty, Decision Avoidance, Behavioral Economics.

[*] Stefano Bonini acknowledges financial support from the Stevens Institute of Technology School of Business. This paper was initiated while Filippo Pavesi was Visiting Associate Professor of Economics at the Stevens Institute of Technology School of Business. We thank Etan Green, John V. Duca, Suman Banerjee, Anand Goel, Alexander Rodivilov, Luca Corazzini for their helpful suggestions. We are also grateful to participants at the 2019 LACEA-LAMES Conference in Mexico City, the 2019 Academy of Behavioral Finance and Economics annual conference, the BEEN 2020 Conference in Verona and seminar participants at Stevens Institute of Technology.

[†] University of Padua, Department of Economics and Management "Marco Fanno", Via del Santo 33, 35123 Padova, Italy, thomas.bassetti@unipd.it

[‡] Stevens Institute of Technology. Address: School of Business, Stevens Institute of Technology, 1 Castle Point on Hudson, Hoboken, NJ 07030, USA, sbonini@stevens.edu

[§] LIUC Università Carlo Cattaneo, C.so Matteotti, 22, 21053 Castellanza (VA), Italy, fpacicco@liuc.it

[¶] LIUC Università Carlo Cattaneo, C.so Matteotti, 22, 21053 Castellanza (VA), Italy and Stevens Institute of Technology, fpavesi@liuc.it

1. Introduction

A number of behavioral explanations have been identified to account for the empirical observation that individuals tend to inefficiently shy away from making certain decisions. We focus on a bias that we refer to as definitivity avoidance that denotes situations in which an agent prefers to avoid making definitive decisions (i.e., those that involve a final judgment). This behavior may play an essential role in several applications in economics and finance. For instance, a bank's choice of rolling over debt for a poorly performing firm involves postponing a final decision, as opposed to forcing the firm to file for bankruptcy. Indeed, numerous finance papers (e.g., Bernanke, 1989; Carey et al., 2012) have identified inefficiently high rates of debt rollover which could be hinting at the existence of a bias. Similarly, closing an investment position even in the presence of noisy signals definitively precludes capturing a possible upside, as opposed to leaving it open and closing it when uncertainty might be reduced.[1] Also, in judicial systems with 'double jeopardy' provisions, acquittals are more definitive as they cannot be appealed. Consistent with a possible bias, anecdotal evidence by legal scholars (e.g., Leipold, 2005) highlights surprisingly high conviction rates in jury trials. Similarly, Westman (1991) has highlighted the social costs of the puzzling length of foster care for children arguing that this is due to the reluctance of court official to make "*definitive decisions*".[2]

In light of the negative welfare effects that these behaviors may determine, in isolation or in conjunction with other determinants, gathering a better understanding of the instances in which they can arise, may also provide guidance for the design of policies that can attenuate their negative consequences.

We show that, if decision making is characterized by definitivity avoidance, introducing a review system that allows interested parties to call for an *ex-post* verification of the correctness of the evaluation, may attenuate (or eliminate) the effect of the bias leading to tangible welfare

[1] The widespread exposure to ABS securities during the great financial crisis offers possible support to this implication.

[2] See Westman (1991), p 47.

gains. So, for example, allowing individuals that were judged guilty to invoke a formal review process of judges' decisions with the potential to undo haphazard decisions may lead to more impartial judgments. In a similar spirit, a stronger empowerment of independence and scope of internal audit committees in financial institutions may prevent excessive risk taking.[3]

The literature has identified four main sources of decision avoidance: status quo bias, omission bias, inaction inertia, and choice deferral (Anderson, 2003). Status quo bias (Samuelson and Zeeckhauser, 1998) is based on the idea that agents may suffer a cost of change. Aversion to action, instead, may be at the root of either omission bias (Ritov and Baron, 1998) or inaction inertia (Tykocinski et al., 1995), as well as possibly affecting the cost of making timely decisions under uncertainty, which may result in choice deferral or procrastination. With respect to these documented sources of decision avoidance, definitivity avoidance is based on the distinctive feature that decision-makers have a strict preference for choices that are non-definitive with respect to those that are definitive. By definitive decisions, we intend those that involve making a final judgment, while non-definitive choices imply a suspension of the final judgment. Our work is related with Gilbert and Ebert (2002) that show that individuals expect more satisfaction from reversible decisions than they do from irreversible ones.[4] Yet, there is a subtle distinction between definitive and reversible decisions. Namely, in the absence of reversibility, some decisions may be more definitive than others, while when decisions are reversible, by definition, they are no longer definitive. This is precisely the feature that we exploit to identify the existence of definitivity avoidance through the introduction of a decision review system.

Our paper is also related to the literature that exploits sports markets as ideal quasi-experimental settings to study decision making behavior (Garicano et al., 2005; Romer, 2006; Bar-Eli et al., 2007; Massey and Thaler, 2013; Pope and Schweitzer, 2011; Green and

[3] Interestingly, this is one provision recommended by the Dodd-Frank act that however has not been implemented Dodd-Frank (2010).

[4] It is worth mentioning that the psychology literature has shown that reversible decisions are also associated with less anticipated regret (Zeelenberg et al., 1996; Tsiros and Mittal, 2000) Also, Gilbert and Ebert (2002) show that individuals display a dynamic inconsistency in that while anticipated satisfaction is greater for changeable decisions, *ex-post* satisfaction actually tends to be lower.

Daniels, 2018). In this respect, the present paper is closely related to Pope and Schweitzer (2011) that provide evidence that loss aversion persists even in contests characterized by high stakes such as professional golf tournaments.[5] However, our analysis is characterized by the following distinctive features. First, we introduce a novel bias that, to the best of our knowledge, has not been previously documented in the extant literature and show that it is consistent with the main features of prospect theory (Kahneman and Tversky, 1979). We then devise an accurate quasi-experimental design that allows us to empirically identify the bias. Finally, we provide welfare implications by analyzing in which cases introducing a review system that allows the interested parties to call on a third party to review decisions, can improve efficiency by attenuating the negative effects of definitivity avoidance.

To develop our claim, we introduce a simple formal model that delivers clear empirical predictions on the effect that the introduction of a review system will have on those decisions that we denote as definitive. The model involves a decision maker that must make a sequence of binary choices, each of which produces a loss for one agent (and may produce gains for a non-empty set of agents). Although only one of the two choices is correct, each decision contains an asymmetry since one of the two alternatives, that we refer to as the definitive choice, involves making a final judgment while opting for the non-definitive choice involves suspending the final judgment. Regardless of whether decisions are definitive or not, they cannot be undone or reversed in case of error unless a specific procedure is in place. Such a procedure, which we refer to as a review system, allows the agent that suffers a loss from the judgment to challenge the decision-maker's ruling. A challenged decision is then reviewed by an impartial third party and possibly overturned if found to be incorrect. The model delivers a positive result allowing us to state that the introduction of a review system will lead to a significant variation in the expected share of definitive decisions, if and only if, decision makers are characterized by definitivity avoidance.

As a second step, we then test the model's predictions by exploiting a natural experiment

[5] Although the paper tackles a more general question, our analysis is also related to the sports literature that addresses potentially biased behavior by judges or referees (Sacheti et al., 2015; Kovalchik et al., 2017).

provided by the introduction of a decision review system in professional tennis tournaments. This system is based on a new rule that allows players to challenge an official's decision and to verify the correctness of a call through the use of a ball-bounce tracking technology known as Hawk-Eye.[6] We consider the judges' decisions before and after the introduction of the Hawk-Eye technology, which occurred in three of the four major professional tennis tournaments (i.e., the Grand Slam tournaments): the US Open (since 2006), the Australian Open and the Wimbledon Championships (since 2007). As part of our identification strategy, we choose as our unit of analysis points that are commonly referred to in tennis jargon as "aces" (i.e., a valid serve that is not touched by the receiver and therefore attributes a point to the server unless called out by the officiating umpires). This choice is motivated by a distinctive feature of these points that allows us to clearly classify the choice as either definitive, when the serve is judged valid and a point attributed to the serving player, or non-definitive if the ball is called out leading the point to be replayed through a second serve. This decision is permanent unless a review system is in place.

The intuition is that when calls are close, umpires that suffer from definitivity avoidance will tend to refrain from reporting what they saw (i.e., making a call that is more likely of being correct), if doing so definitively assigns a point to one player or the other. If this is the case, according to our model, the introduction of the decision review system should lead the share of overturned calls to be skewed in one direction or the other. Indeed, our empirical analysis provides robust evidence in favor of the fact that the introduction of such a system increased the share of decisive calls. This allows us to claim that judges are subject to definitivity avoidance, as implied by the model.

Exploiting the results of the empirical analysis subsequently permits us to derive welfare implications. In particular, we show that a review system that increases the share of definitive decisions always leads to a welfare improvement, unless it more than offsets a decision maker's initial definitivity avoidance by inducing her to become excessively definitivity loving (i.e.,

[6] Throughout the paper we use the terms challenge rule and Hawk-Eye technology as synonyms to refer to the decision review system.

having a strict preference for definitive decisions over non-definitive ones). Even in this latter case, welfare can improve as long as the precision of the information received by the agents that are affected by the decision is sufficiently high, and the cost of invoking a challenge is sufficiently low in relation to the accuracy of the review technology.

The remainder of the paper is organized as follows: Section 2 presents the model and the testable prediction; Section 3 outlines the empirical setting; Section 4 describes the data; Section 5 introduces the empirical methodology; Section 6 presents the results and Section 7 addresses welfare implications. Section 8 provides a discussion of the results and Section 9 concludes.

2. Theoretical Framework

We develop a model in which a decision maker must make a sequence of binary choices, each of which affects the payoffs of a non-empty set of agents and produces negative consequences for one specific agent. Although only one of the two choices is correct, each decision contains an asymmetry since one of the two alternatives, that we denote as the definitive choice, involves making a final judgment while opting for the non-definitive choice involves suspending the final judgment. Regardless of whether decisions are definitive or not, they cannot be undone or reversed in case of error unless a specific procedure is in place. Such a procedure, which we refer to as a review system allows the negatively affected agent to challenge the decision (at a cost) A challenged decision is then reviewed by an impartial third party and possibly overturned if found to be incorrect.

Given the scope of our analysis, we concentrate on modeling the subset of dubious decisions that could have been challenged had the review system been in place. Dubious decisions are defined as those for which the decision maker and agents do not observe the state, but each player observes an independent signal, $s \in [0,1]$ that is imperfectly informative on the true state $\omega \in \{0,1\}$, before the binary decision $d \in \{0,1\}$ is made. Here $d = 0$ and $d = 1$

respectively represent the non-definitive and the definitive choices. Signals are distributed according to a continuous density function $f_\omega(s)$ with cumulative distribution function $F_\omega(s)$. This information structure represents a setting characterized by binary signals with different degrees of precision. To simplify exposition, we assume that the signals of the decision maker and the agents are independent and follow the same distribution, since assuming different distributions would make the notation cumbersome without affecting the results.

Before observing a signal on the state, the decision maker is assumed to have a fair prior, so that $\Pr(\omega = 1) = 1/2$.[7] We assume that the signal is informative on the state, meaning that it satisfies MLRP (Marginal Likelihood Ratio Property) so that $\frac{f_1(s)}{f_0(s)}$ is increasing in s. This implies that the higher (lower) is the signal, the more likely it is that the state of the world is higher (lower). Moreover, the signal structure is assumed to be symmetric, so that $f_1(s) = f_0(1 - s)$ for every s. By Bayes' rule we therefore have that:

$$Pr(\omega = 1 \mid s) = \frac{f_1(s)}{f_1(s) + f_0(s)}.$$

We introduce a bias parameter $b \in \{DA, NB\}$ where DA denotes definitivity avoidance meaning that the judge obtains a higher net benefit from providing a correct evaluation when $\omega = 0$ with respect to when $\omega = 1$, while in the absence of a bias (NB) the net benefit of a correct decision is equal in both states. We also introduce a regime variable $r \in [H, NH]$ that denotes whether a decision review system is present (H) or whether such a system is not in place (NH).

The review system is entirely characterized by two features. The first is the precision of the review technology $p \in (1/2, 1]$ which denotes the probability that the true state of the world is discovered once the review system is invoked. The second is the probability that a decision maker's decision is not challenged when an agent's private information contradicts that of the decision maker, that we denote with $c \in [0, 1)$, and is a reduced form to account for

[7] Considering the specific application to tennis officials, given that challenges occur mainly for balls bouncing close to the line (Mather, 2008), it is straightforward to assume that in these cases, the probability of the ball being in or out prior to observing a signal is close to 1/2.

the agent's cost of making a challenge. Given that the bias applies to those in the position to make decisions, we naturally assume that the agents are not subject to definitivity avoidance.

The utility of the decision maker of making decision d if the bias is b in regime r after receiving signal s is given by the following expression:

$$U(d, b, r \mid s) = Pr(\omega = 1 \mid s)v(d, 1, b, r) + Pr(\omega = 0 \mid s)v(d, 0, b, r),$$

where $v(d, w, b, r)$ represents the value function for the decision maker of choosing d when the state of the world is ω, the bias is b, and the regime is r.

It is relevant to point out that definitivity avoidance (DA) may be derived from the following standard properties of prospect theory (Kahneman and Tversky, 1979): 1) utility is defined in terms of gains and losses with respect to a reference point; 2) utility is steeper in losses than gains which implies loss aversion; 3) utility is convex in losses and concave in gains (i.e., the value function exhibits diminishing sensitivity).

INSERT FIGURE 1 HERE

We provide a description of the role of each of these properties, which are graphically represented in Figure 1. First, notice that the reference point for gains or losses is the single decision and not the complete set of decisions made by the decision maker over a longer time frame (or the course of her/his career), which is consistent with property 1.[8] Since previous decisions do not play a role, making a correct current choice naturally leads to a gain, while getting it wrong leads to a loss with respect to the reference point, which is zero before the decision is made. We represent this with the indicator function $x \in \{-1, 1\}$, where $x(d \neq \omega) = -1$ and $x(d = \omega) = 1$. Property 2 implies that $0 < v(d = \omega, b, r) < -v(d \neq \omega, b.r)$, in other words, the utility from a correct call is less than the disutility from an incorrect call. Now notice that, in the absence of a review system, getting it wrong when choosing

[8] In the tennis setting, the single decision represents the current point as opposed to the complete set of calls made throughout the match. Although within a match some crucial points may be more salient than others, we abstract from this heterogeneity. Indeed the impact of definitivity avoidance should be more pronounced if we were to consider only these salient points.

$d = 1$ is a definitive mistake, meaning that it involves assigning a negative payoff to a specific agent. By property 3, the convexity of the value function in the negative domain implies that making a mistake when choosing $d = 0$ is strictly better than when $d = 1$, because the final decision is suspended and therefore equivalent to a lottery in which the loss is not certain. The opposite holds for the positive domain, since a sure gain is always preferable to an uncertain one in the presence of risk aversion. These considerations lead us to define the following relations relative to the decision maker's value function in the different states in the presence of definitivity avoidance and in the absence of a review system:

$$0 < v(1, 1, DA, NH) - v(0, 0, DA, NH) < v(0, 1, DA, NH) - v(1, 0, DA, NH). \qquad (1)$$

When a review system is introduced, it will make the definitiveness of the decision weakly less relevant, making correctness the salient attribute. More specifically, when an incorrect decision is made, since it becomes reversible, the disutility of the choice will depend (weakly) less on which was made. On the other hand, the greater public exposure provided by the review system makes the correctness of the decision (weakly) more salient even if one is less definitive than the other. These properties are more formally represented by the following two expressions:

$$0 \leq v(1, 1, DA, H) - v(0, 0, DA, H) \leq v(1, 1, DA, NH) - v(0, 0, DA, NH), \qquad (2)$$

and

$$0 \leq v(0, 1, DA, H) - v(1, 0, DA, H) \leq v(0, 1, DA, NH) - v(1, 0, DA, NH). \qquad (3)$$

In other words, in the presence of a bias, the introduction of the review system weakly reduces the distance between the gains (losses) of providing a correct (incorrect) evaluation with $d = 1$ with respect $d = 0$. Naturally, if the decision maker is unbiased, the introduction of the review system should have no impact on the value assigned to correct versus incorrect decisions since definitivity does not play a role. We, therefore, have that:

$$0 = v(1, 1, NB, r) - v(0, 0, NB, r) = v(0, 1, NB, r) - v(1, 0, NB, r). \tag{4}$$

We make the standard assumption that information has an impact on decisions, which implies that signals are persuasive regardless of the bias. In other words, the utility functions of decision makers and the informativeness of signals are such that there always exists a threshold $s^* \in (0, 1)$, for which a decision maker will set $d = 1(d = 0)$ for $s > s^*$ $(s < s^*)$. This threshold value is defined by the value of s for which the decision maker is indifferent between taking either action, implying that $U(1, b, r \mid s^*) = U(0, b, r \mid s^*)$. We therefore have that:

$$\frac{Pr(\omega = 1 \mid s^*)}{[1 - Pr(\omega = 1 \mid s^*)]} = \frac{v(0, 0, b, r) - v(1, 0, b, r)}{v(1, 1, b, r) - v(0, 1, b, r)}.$$

We denote s^*_r as the threshold value of s^* in regime r, and s^*_{NB} as the corresponding benchmark threshold in the absence of a bias. Given relations (1), (2), (3) and (4), then the following proposition follows directly:

Proposition 1. *If $b = NB$ then $s^*_H = s^*_{NH} = s^*_{NB}$, while if $b = DA$ then $s^*_{NH} > s^*_{NB}$ and $s^*_H \lesseqgtr s^*_{NB}$.*

Intuitively, whenever there is DA, expressions (1) and (4) imply that $s^*_{NH} > s^*_{NB}$, while in the absence of a bias (4) implies that $s^*_H = s^*_{NH} = s^*_{NB}$. However, in the presence of DA the introduction of the review system could lead to a share of definitive decisions that is either above, below (or equal) to the unbiased share, based on whether the distance between the gains of providing a correct evaluation with $d = 1$ with respect $d = 0$ is greater that the loss of providing an incorrect evaluation with $d = 1$ with respect to $d = 0$. More formally, if $v(1, 1, DA, H) - v(0, 0, DA, H) < (\geq) v(0, 1, DA, H) - v(1, 0, DA, H)$ then $s^*_H > (\leq) s^*_{NB}$.

Notice that the symmetric signal structure implies that $(1 - F_1(s^*_{NB})) = F_0(s^*_{NB})$, so that in the absence of a reporting bias the probability of providing a correct evaluation is equal in both states.

In terms of agent behavior, based on the signal observed, a given agent believes state $\omega = 1 (\omega = 0)$ is more likely to be the true state whenever $Pr(\omega = 1 \mid s) > 0 (< 0)$. We assume that in the presence of the review system, the decision is challenged with probability $(1 - c)$ by the agent that incurs in a loss from the decision and disagrees with the decision maker, meaning that the decision maker's decision does not correspond to the state the agent believes is more likely to be true.[9] Whenever a judgment is challenged, with probability p, which defines the precision of the review technology, the decision is overturned if it does not match the true state.

We denote the decision maker's decision conditional on the bias, the regime and the signal observed with $d(b, r, s) \in [0, 1]$ and the expected decision before observing the signal with $E[d(b, r)]$. The difference between the expected decision before and after the introduction of the review system when the bias is b is given by $\Delta d(b) \equiv E[d(b, H)] - E[d(b, NH)]$, which is equal to[10]:

$$\Delta d(b) = 1/2[\sum (F_\omega(s^*_{NH}) - F_\omega(s^*_H))] +$$

$$+ \left[(F_0(s^*_H) + F_1(s^*_H) - 1)(1 - c)(F_1(s^*_{NB})(1 - p) + (F_0(s^*_{NB})p) \right].$$

Considering Proposition 1, it can be shown that $\Delta d(NB) = 0$ and $\Delta d(DA) \gtreqless 0$, which leads to the following empirical prediction:

Prediction 1. *The introduction of the review system leads to a significant variation in the expected share of definitive decisions if and only if decision makers are subject to definitivity avoidance* (DA) *(proof in the appendix).*

[9] In order to simplify the analysis, we abstract from strategic challenges that may arise even when an agent's private information does not contradict that of the decision maker.

[10] A formal derivation of $\Delta d(b)$ is provided in Appendix A.

3. Empirical Setting

3.1. Motivation

In order to test our model implications, we need to identify a setting characterized by the existence of decisions that can have definitive effects on the outcome, the possibility to observe the results of a review system, and a sufficient number of observations to ensure robust inferences. We believe that professional tennis matches represent an ideal setting for the following reasons. First, in most played points, a judge decision is required to determine the validity of the shot. These decisions can be definitive because they may result in the attribution of a point to one of the two players. Second, in 2006, professional tennis tournaments started introducing a review system called Hawk-Eye, whereby players have the opportunity of "challenging" a judge's call if they have reason to believe that it is incorrect. Crucially, the review is done by a mechanical tool that does not involve human intervention and ensures "fair" decisions. Finally, the staggered introduction of the review system across courts allows employing a set of DD estimators to precisely identify the effect of the bias, if any.

3.2. Tennis game features and structural break

In professional tennis, officials can be on or off-court. Off-court officials are responsible for ensuring that the rules of tennis are correctly enforced and act as the final authority on all questions related to tennis norms. On-court officials decide on all issues during the match. A team of on-court officials consists of a chair umpire and some line judges. The chair umpire has the last word on all questions relating to on-court facts, for example, whether a ball was "in" or "out," a service touched the net, a player had committed foot fault, etc. Line judges call all shots related to their assigned line and help the chair umpire in guaranteeing a fair match. On-court officials must be in good physical condition with a natural or corrected vision of 20-20 and normal hearing. International chair umpires must submit a completed eye test form each year to ITF Officiating, while all other certified officials must submit a completed eye test form every three years. The chair umpire may overrule a line judge only in the case of a clear mistake (i.e., beyond any reasonable doubt) by the line judge and only if the overrule is made promptly (i.e., almost simultaneously) after the error is made. A full line team consists of ten line judges, but other configurations are possible.[11] However, the improved physical performance of players together with the evolution of equipment have substantially increased the speed of the game, thus making judges' calls increasingly contested. To address this issue, in 2006 the Association of Tennis Professionals (ATP) introduced at the US Open a rule allowing players to challenge a decision made by the officials, invoking *ex-post* verification of the correctness of a call through a technology known as Hawk-Eye. This rule was first extended to the Australian Open and the Wimbledon Championships in 2007 and then gradually rolled out to the other competitions. The Hawk-Eye technology is a ball-tracking system used to reconstruct a four-dimensional position of the ball. This technology is based on six or more computer-linked cameras situated around the court. The videos from the cameras are triangulated and combined to create a three-dimensional representation of the ball's trajectory. Once a player challenges a line judge's call, the system accurately reconstructs the path of the ball and its landing point with high precision.[12]

Given its substantial cost, the Hawk-Eye system has been only gradually adopted by main tournaments. This allows us to adopt a difference-in-differences (DD) approach where the treatment group will be given by matches in courts that, at some point, introduced the system, while the control group will be characterized by courts in which the Hawk-Eye technology has not been implemented during the period of analysis.

A possible concern with our identification strategy is that judges may exhibit idiosyncratic

[11] For instance, at the Wimbledon Championships 2008, line teams worked on a timed rotation (75 minutes on, 75 minutes off), with nine line judges per team on the main four courts and seven line judges on the others.

[12] The Hawk-Eye Innovations website (https://www.hawkeyeinnovations.com/) reports that the ball position is exact within a 3.6 mm average margin of error. Since the standard diameter of a ball is 67 mm, the error is 5.37% of the ball diameter. According to the International Tennis Federation (ITF), this is an acceptable margin since the ball maximum stretch can be longer.

biases in their officiating, thus potentially affecting our results. However, several arguments moderate this concern. First, challenges are a relatively low-frequency event; therefore, the impact on the overall outcome of the game is limited. According to Mather (2008) and Whitney et al. (2008), the average number of challenges in the top three tournaments following the introduction of the system has been 6.85 for men, and 4.14 for women, with only 27% of these challenges that overturned the line judge decision. Second, as highlighted, judges rotate frequently during a match, thus minimizing the impact of any judge-specific noise on the calls. In light of these arguments, it would be implausible to attribute any significant result to the idiosyncratic effect of judge-specific behaviors or characteristics, thus ensuring a reliable setting for our study.

3.3. Unit of analysis

Challenges can be invoked by players on any point during the match, under the challenge quota constraint.[13] In this respect, players may engage in strategic behavior when choosing whether to challenge points. In order to minimize, if not altogether eliminate, this possible confounding effect in our identification strategy, we select as unit of analysis what are commonly denoted as "aces" in the tennis lingo.[14] Tennis matches are characterized by two players that alternate in initiating the play by "serving" the ball to the opponent. Players are given two chances to initiate the game with a valid serve. If the serving player fails both, a point is awarded to the opponent. An ace is a valid serve that is not touched by the receiver. Typically aces are scored in the first of the two allotted opportunities to initiate the game as the serving player can take more risks in serving more powerfully and/or seeking more extreme ball placements. This implies that, if the line judge does not intervene by calling the ball "out" of the service box, the serve will be an ace, and a point will be assigned to the server. In this respect, aces identify a situation where a third party decision may have

[13] A player can invoke a review a maximum of three incorrect challenges per set, after which they are not permitted to challenge again in the set. However, if a set goes to a tiebreak, this limit increases from three to four incorrect challenges for the set.

[14] By definition, an ace is a legal serve that is not touched by the receiver.

definitive effects: if a judge chooses to intervene s/he avoids assigning the point which would otherwise permanently affect the players' scores. Because the absolute number of aces may be affected by the match length, we standardize it computing an "ace ratio" variable given by the total number of aces over the total number of served points that we use as our dependent variable.

A possible confounding factor in our tests is the endogenous change in players' characteristics and in the equipment technology. Over time, tennis has become substantially more muscular and players' characteristics have changed significantly, also in response to the introduction of new materials, designs, and construction techniques for rackets. We mitigate this issue in several ways. First, we constrain the length of the estimation window to matches played between 2002 and 2010 (i.e., 4 years before and after the introduction of the challenge system). Second, if players changed their strategies because of the Hawk-Eye technology, they should rationally use the same strategies also on clay courts, where the ball leaves a mark on the surface that generally is accurate enough to establish whether the ball bounced in or out. We exploit this feature to estimate a triple difference model in which matches played on a clay court constitute a placebo control group. Third, the lack of experience on clay, which embeds this natural review system, should translate into a steeper learning curve for less clay-experienced players as they would need time to change both style and strategies. We, therefore, classify players according to their experience on clay courts and test whether the Hawk-Eye effect on the ace ratio is higher for matches with players having a lower experience. Finally, we further check the robustness of our results, carrying out a double robust treatment effect analysis in which each year is considered as a separate experiment.

4. Data

Our dataset is derived from the *tennis ATP* data published by Jeff Sackmann[15]. This dataset contains detailed statistics and results on most of the professional tennis matches

[15] https://github.com/JeffSackmann/tennis_atp

from the beginning of the *Open Era* (1968) until now, for both the Association of Tennis Professionals (ATP) and Women's Tennis Association (WTA). While for older tournaments the coverage is slightly less detailed, Sackmann validated these results to avoid the inclusion of wrong ones. The resulting dataset is recognized as highly accurate and reliable enough to have been used in several prior studies (e.g Rodenberg et al., 2016; Kovalchik et al., 2017; Cohen-Zada et al., 2018; Antoniou and Mavis, 2019). As discussed in Section 3, in order to mitigate possible concerns about endogenous changes in players and/or equipment characteristics, we constrain our data to a 9-year window centered around the 2006 first introduction of the HawkEye system. Despite the richness of the *tennis ATP* dataset, it does not include the court name on which the match was played: as the Hawk-Eye was initially adopted only on a selected number of courts, we need to unequivocally identify for each match and at any given point in time whether the court was treated (i.e., whether the review system was operative and officially used). We retrieve this information by consulting the archived versions of the official tournament websites available through Wayback Machine, scraping the court name for each match from the initial Hawk-Eye introduction. For the tournaments in our treatment group, the US Open, the Australian Open and Wimbledon, the US Open was the first Grand Slam tournament to adopt the Hawk-Eye technology in 2006, precisely on the Arthur Ashe and Louis Armstrong stadiums, followed by the adoption on two Wimbledon courts (Centre Court and Court 1) and the Rod Laver Arena in the Australian Open, both in 2007. Out of 508 matches played after 2006, we were able to identify the name of the court for 499 matches, or 98.2%, a result that allows us to confidently state that there is an absence of sampling bias.

Table 1 reports the number of matches played in treated and untreated courts before and after the introduction of the Hawk-Eye technology. The first part of the table refers to our baseline sample (i.e., matches played only on grass and hard surface), whereas the second part of the table also includes French Open matches, played on clay surfaces, a natural review system that, as indicated, we use as a control in robustness tests. Since our analysis considers couples of players that played at least two times, the number of matches played with and without Hawk-Eye technology increases because now they are coupled with matches on the clay surface. Notice that in 2006 the Hawk-Eye system was used only in 16 matches and was fully implemented in 2007 when all treated courts had the new monitoring system. The time distribution of treated matches supports the strategy of including data relative to 9 years centered around 2006 to identify systematic differences in referees' behaviors before and after the introduction of Hawk-Eye.[16] We restrict the sample to matches played by the same paired couple of players before and after the introduction of the Hawk-Eye system. The rationale is that such constraint allows us to minimize unobserved heterogeneity that might affect tests on the whole sample. Accordingly, we model the treatment variable as a dummy set to 1 if the Hawk-Eye system is used on the match court.

INSERT TABLE 1 HERE

Our dependent variable is the ratio of aces in a match (i.e., the total number of aces over the total number of served points). Figure 2 provides a box plot showing the distribution of ace ratios (in percentage points) for matches played in control and treatment courts where whiskers identify the minimum and maximum contiguous observations without outliers. As expected, since our experiment has a crossover design (i.e., the same couple of players may play in both types of courts every year), the two distributions tend to overlap, suggesting caution in the visual identification of a pattern.

INSERT FIGURE 2 HERE

We complement these variables with a set of time-varying covariates potentially affecting the ace ratio: players ages, ranking, and home-field advantage as well as the match length in minutes. Finally, we also include court, time and pair of players' fixed effects,

We present descriptive statistics in Table 2.

[16] Table B1 in Appendix B reports the number of matches played in each treated court over time (i.e., before and after the treatment).

INSERT TABLE 2 HERE

The total number of observations, including French Open matches, is 1,010, the average ace ratio is 7.529%, and the fraction of matches played with the Hawk-Eye system is 36.2%. In addition, 52.7% of the matches are played on hard surfaces (the US Open and the Australian Open), a 23.7% on clay (French Open) and a 23.6% on grass (Wimbledon). The total number of observations is almost equally divided into pre- and post-treatment period. On average, the favorite, (i.e., the player with the lowest rank in the match) is 17th in the World ranking, whereas the average rank of the challenger (i.e., the player with the highest rank) is about 66th. Looking at age, the favorite and the opponent do not differ significantly, both averaging at about 25-years of age. The fraction of matches in which at least one of the two players comes from the country organizing the tournament is 14.3%, while on average matches last 148.3 minutes. Finally, we also proxy players' clay experience (CE), accumulated in the four years before the introduction of the Hawk-Eye, with the average share of matches that a pair played on clay courts. On average, the CE is 0.26.

We present pairwise correlation coefficients for all our variables in Table 3.

INSERT TABLE 3 HERE

There is a positive unconditional correlation between the ace ratio and the matches disputed with the support of the Hawk-Eye technology. The ace ratio is also positively correlated with the challenger's age. Vice versa, there exists a negative correlation between the ace ratio and the total number of minutes, that is, the performance seems to decrease with the length of the match. As expected, home players and those with the lowest rank are also more likely to play with the Hawk-Eye technology. Finally, younger players tend to have better positions in the ranking, independently of whether they are favorites or challengers.

5. Methodology

To identify a line judge behavioral bias, we consider matches disputed with and without the Hawk-Eye technology. Although the ITF system is designed to avoid line judges' idiosyncratic effects, we must account for the fact that some pre-treatment variables might affect both the outcome variable and the probability of being treated. Therefore, we estimate the average treatment effect of the Hawk-Eye technology on the treated, using two different techniques: a fixed-effect difference-in-differences (FE-DD) estimator for the longitudinal analysis and a doubly robust estimator for cross-sectional studies. This approach is helpful in tackling the assumption that points are i.i.d. that is underlying our model and that, empirically, characterizes most of the articles on professional tennis (see, e.g., George, 1973; Gillman, 1985; Walker and Wooders, 2001; Klaassen and Magnus, 2009; Ely et al., 2017). Klaassen and Magnus (2001) provide a seminal discussion of this assumption. In their contribution, the authors conclude that points are neither independent nor identically distributed. However, they also show that, controlling for players' quality, the i.i.d. property is reasonably holding (i.e., the deviation from a perfect i.i.d. hypothesis is small). Additionally, in a subsequent article, the same authors assume an independent, identical distribution of points and justify their choice as follows: "we do not use the points themselves but summary statistics (averages) so that any possible harm caused by the wrong assumption is much reduced" (Klaassen and Magnus, 2009, p. 78). Since our analysis refers to averages and - as further illustrated below - we control for both time-varying players' quality (proxied with the current ranking) and time-invariant players' quality (absorbed by pairs fixed effects), we assume without loss of generality that points are i.i.d. as in previous studies.

With longitudinal data, a FE-DD estimator represents a natural choice to control for potential confounders (see, e.g., Arellano, 2003; Angrist and Pischke, 2008; Wooldridge, 2010; Hsiao et al., 2012). This is because a FE approach effectively restricts matches to within a pair of players, while pairs without a change in treatment status do not affect results (see, e.g., Wooldridge, 2010).

Formally, we proceed as follows: we first divide courts into those affected by the introduction of the Hawk-Eye system at some point in time ($Treated = 1$) and those that never experienced the Hawk-Eye technology during the sample period ($Treated = 0$). Then, we also distinguish time periods in terms of years before the introduction of the monitoring system ($Break = 0$) and years after the introduction of the new technology ($Break = 1$). Clearly, in a FE specification, the direct effects of these two dummy variables will be absorbed by the vector of fixed effects. However, we are interested in the interaction term between $Treated$ and $Break$. Using the notation adopted in the theoretical model, we have that: $H \equiv Treated \cdot Break$. This interaction term indicates whether a match is played with the support of the Hawk-Eye technology ($H = 1$) or without ($H = 0$) and captures the average treatment effect of the Hawk-Eye technology on the treated. Formally, we estimate the following FE-DD model:

$$Y_{pct} = \alpha_c + \alpha_t + \beta \cdot H_{ct} + \gamma \cdot X_{pct} + \delta_p + e_{pct}, \tag{5}$$

where Y_{pct} is the aces to points ratio measured for pair p, in court c at time t, α_c and α_t are court and time dummies absorbing the direct effects of the treatment group and the Hawk-Eye introduction period, H_{ct} is a dummy variable indicating whether the Hawk-Eye technology was available in court c at time t, X_{pct} is a matrix of time-court varying pair's characteristics, namely rank, age, match duration, and home advantage (whenever applicable). Finally δ_p are pairs fixed effects, and e_{pct} is the error term. We account for pairs time-series dependence using clustered-robust standard errors and use different time breaks to properly identify the treatment period.[17]

In Equation (5), the treatment is assumed to be strictly exogenous conditionally on observed and unobserved heterogeneity. This approach is equivalent to a matching estimator restricting matches to within a pair of players, where pairs without a change in the treatment

[17] Since we have a crossover design, pairs are not nested within courts and, therefore, we cannot cluster errors at the court level.

status do not affect estimation (see Wooldridge, 2010). Although the treatment indicator, H_{ct}, exhibits enough within variability and the treatment assignment is likely to depend on pairs' observable and unobservable characteristics, one may argue that our specification is rather demanding in terms of both FE vector and clustered standard errors. Therefore, we also estimate a difference-in-differences model with only court and time fixed effects. Assuming that estimators based on data from courts are approximately independent, unbiased, and Gaussian, but not necessarily of equal variance, we adopt the approach proposed in Ibragimov and Müller (2010). This methodology leads to robust inference even when the data is heterogeneous and correlated in a largely unknown way. Equation (5) becomes:

$$Y_{pct} = \alpha_c + \alpha_t + \beta \cdot H_{ct} + \gamma \cdot X_{pct} + e_{pct}, \tag{6}$$

The only difference between (5) and (6) is the lack of pair FEs in the second specification. Now, using a wild bootstrap-t, we can cluster standard errors at the court level (see Cameron et al., 2008).[18]

We cannot exclude *a priori* that, after the introduction of the Hawk-Eye, players have changed their strategies, acquiring specific skills such as a different way of serving or challenging the judges' calls. Similarly, officials might have learned how to umpire with the Hawk-Eye system. While this second case would not be a problem since a learning effect would represent an additional correction mechanism revealing the existence of a previous bias, a change in players' strategies constitutes a potential confounding factor. Now, if this change affects both courts, with and without the Hawk-Eye, time fixed effects will control for this effect; vice versa, if this change only happened in the presence of Hawk-Eye technology, it would weaken our identification strategy. We address this issue in three different ways.

[18] With respect to other residual, cluster-bootstrap techniques, this approach does not assume *a priori* that regression errors are i.i.d and we do not need a balanced data set where all clusters are equally represented. Indeed, in differences-in-differences analyses with few clusters, traditional bootstrap resampling methods may lead to inestimable coefficients. This happens because the regressors of interest are indicator variables, and some bootstrap samples may generate too little within variability. In general, as proved in Cameron et al. (2008), a wild cluster bootstrap-t procedure outperforms other bootstrap methods.

First, we estimate a triple difference model in which clay courts constitute a placebo control group. Our specification can be modified as follows:

$$Y_{pct} = \alpha_c + \alpha_t + \beta_1 \cdot H_{ct} + \beta_2 \cdot C_{ct} + \gamma \cdot X_{pct} + \delta_p + e_{pct}, \tag{7}$$

where C_{ct} indicates clay courts after the treatment period. In the presence of a definitivity bias, we expect the estimate of β_2 to be statistically insignificant. As before, we estimate Equation 7 with and without pair FE, δ_p. Second, we classify matches according to players' experience on clay courts. Because players with less experience on clay courts need more time to adapt their style and strategies to the new system, immediately after the introduction of the challenge rule, the Hawk-Eye effect should be higher for those players that are less used to play on clay courts. Accordingly, we modify Equation (5) as follows:

$$\begin{gathered} Y_{pct} = \alpha_c + \alpha_t + \beta_1 \cdot H_{ct} + \beta_2 \cdot Break \cdot CE_p + \beta_3 \cdot Treated \\ \cdot CE_p + \beta_4 \cdot H_{ct} \cdot CE_p + \gamma \cdot X_{pct} + \delta_p + e_{pct}, \end{gathered} \tag{8}$$

where CE_p denotes pairs' clay experience and is proxied by the average share of matches played by pair p on clay courts before the introduction of the Hawk-Eye. If our estimates of β_1 are statistically significant, this will be a convincing sign of treatment effect. Notice that the direct effect of CE_p is absorbed by δ_p.

Finally, we use a sequential approach in which every cross-section is considered as a separate experiment. This allows us to relax the parallel trend assumption embedded in DD models. Moreover, if the Hawk-Eye effect remains stable over time, then it would mean that players did not change their way of serving after the introduction of the Hawk Eye. In particular, we use an inverse-probability-weighted regression-adjustment (IPWRA) estimator. This methodology requires both a model for estimating the probability to be treated (propensity score) and a regression model for the outcome. The IPWRA estimator provides unbiased results of the treatment effect when either one or both models are correctly specified. In other words, it is a doubly robust estimator (see Wooldridge, 2007). We assess the association between the exposure to the Hawk-Eye and the outcome, controlling for a set of covariates.

In particular, for any $t = 2006, \ldots, 2010$, we first estimate a logistic selection model, and then we use the predicted probability scores to adjust our linear estimates. Formally, our selection equation is:

$$p(x_i) = Pr(H = 1 \Big| X_i) = \frac{1}{1 + e^{-\delta \cdot X_i}}, \tag{9}$$

where i denotes the i-th cross-sectional observation characterized by pair p playing in a specific tournament. Assuming $p(x_i) > 0$, $x_i \in X_i$, the expected average treatment effect on the treated (ATT) is simply

$$\tau_{ATT} = \sum_{i=1}^{N} \cdot H_i \cdot Y_i - \sum_{i=1}^{N} \frac{p(x_i)}{1 - p(x_i)} \cdot (1 - H_i) \cdot Y_i. \tag{10}$$

where $\frac{p(x_i)}{1-p(x_i)}$ is the corresponding odds, reflecting the likelihood that a pair will be assigned to a court equipped with the Hawk-Eye technology. An alternative double-robust estimator would be an augmented inverse-probability weighting (AIPW) estimator. However, the AIPW approach is sensitive to extreme values of the propensity score and can produce unreliable estimates when the outcome is bounded in some way (see, e.g., Kang and Schafer, 2007; Robins et al., 2007; Słoczyński and Wooldridge, 2018). Moreover, as shown in Wooldridge (2007), an IPWRA approach is particularly suitable when the propensity score model is more likely to be correctly specified. Indeed, in our case, the informal rules to assign a match to the main courts are fairly standard and straightforward: the assignment is inversely related to the rank of the players and home players are generally favored.

Finally, since tennis players usually prefer to play on specific surfaces and in specific tournaments, as additional robustness check, we re-estimate Equations (5) and (7) taking into account pairs-tournament specific effects.

6. Results

6.1. Main results

Table 4 shows the estimates of Equations (5) and (6) for different time breaks.

INSERT TABLE 4 HERE

Results suggest that most of the Hawk-Eye effect is observed in 2007 and 2008 when three out of four tournaments adopted the Hawk-Eye technology. In these two years, prudent estimates indicate that the average treatment effect on the treated was about 0.979% (Column 3 of Panel B) and 1.508% (Column 4 of Panel A), and it is statistically significant at a 5% confidence level.

Our results are consistent with stylized facts that can be inferred from the analysis of overturned decisions: the average number of challenges per match was 6.5 at the 2008 US Open, 6.7 at 2009 Wimbledon, and 4.88 and 8.02 at the 2007 and 2010 Australian Open tournaments. Since challenges are more likely to happen on serve than other shots (49.8% vs. 27.3%, for men) and that these are challenged 3-4 times per match (Kovalchik et al., 2017), it can be expected that the maximum number of reversed serves per match should be between 1 and 1.6 (Mather, 2008), as the average rate of correct calls is around 30%-40%. By using our data and multiplying the Hawk-Eye effect by the average number of points in a match (i.e., 221), we get a correction of 2 serves per match, a value compatible with the average treatment effect discussed above.

Notice that, in both specifications, the limited number of observations for matches played with the Hawk-Eye system at the US Open in 2006 reduces both the magnitude of the effect as well as the statistical significance. Similarly, if we restrict the treatment period to 2009 and 2010 and include 2006, 2007, and 2009 in the control group, the Hawk-Eye effect decreases, showing that the break took place before 2009. In the cross-section analysis, we examine the Hawk-Eye effect year-by-year. This will allow us to determine whether the Hawk-Eye effect persists over time or not. Among control variables, in the pair-FE specification, the challenger's ranking exhibits the only additional significant within-estimate; whereas, in the specification without pair FEs, the home-field advantage has a positive and statistically significant effect on the ace ratio.

Table 5 reports the estimates of Equation (7) with and without pair FEs. With respect to Table 4, here, we have added a third group represented by matches played on a clay court (the French Open) as a placebo treatment. For the sake of space, hereafter, we do not report the estimated coefficients of our additional control variables.[19]

INSERT TABLE 5 HERE

This exercise represents an important robustness test for our previous results. Although we are now introducing more unobserved heterogeneity related to the fact that some players systematically prefer to play on specific surfaces, results confirm the main conclusions drawn in Table 4: in 2007 and 2008, when for the first time the Hawk-Eye was fully operative, the within effect of the system is positive (about 1.1-1.2%) and statistically significant. Moreover, if we only consider court FEs, the magnitude of the treatment effect becomes larger and more significant (Panel B of Table 5).

While in the additional robustness section we re-estimate Equations (5) and (7) taking into account tournament-pairs interaction effects, here we test whether our previous results depend on the fact that, even before the implementation of the Hawk-Eye technology, some tennis players experienced a sort of natural Hawk-Eye technology represented by the clay surface. In this case, we cannot exclude *a priori* the possibility that these players already have experience in playing with a verification system and therefore may have changed their service strategy immediately after the introduction of the Hawk-Eye. In this respect, Table 6 reports the estimated coefficients of Equation (8) for different time breaks, with and without pair FEs.

INSERT TABLE 6 HERE

[19] These coefficients are available in the supplementary material available online.

The Hawk-Eye coefficient (i.e., β_1) is large and statistically significant when we consider the periods 2006-2010 and 2007-2010. This means that, after the introduction of the Hawk-Eye, matches involving tennis players with no experience on clay courts ($CE = 0$) exhibited a significant increase in the ace ratio. In contrast, by looking at the coefficient of Hawk Eye$\cdot CE$, it seems that the new technology has initially penalized (in 2006 and 2007) tennis players characterized by a specific experience on the clay surface. This temporary result is consistent with the idea that service skills are particularly useful for players preferring hard and grass surfaces and allows us to rule out the hypothesis of a preexisting service strategy suitable for the Hawk-Eye technology. If we consider that the average CE reported in Table 2 is about 0.26, it is easy to link the results reported in Table 6 with the estimates of the Hawk-Eye effect presented in Table 4.[20] Interestingly, it now emerges a negative impact of challenger's rank on the ace ratio. In other words, weaker opponents reduce the overall ace ratio.

Table 7 provides an alternative method to estimate the average treatment effect on the treated in case of panel data.

INSERT TABLE 7 HERE

This method consists of considering a panel as a sequence of cross-sectional natural experiments (see, e.g., Wooldridge, 2010). In particular, in Table 7, we estimate the double-robust estimator proposed in Wooldridge (2007) for each year separately. Given the small number of treated observations in 2006, we restrict our estimates to tournaments with the same surface of the US Open; otherwise, Wooldridge's algorithm would not converge. In line with Tables 4 and 5 , the sequence of IPWRA estimators shows a significant treatment effect for years 2007, 2008 and 2009. This effect remains around 1.4-1.5% and indicates that before the introduction of the Hawk-Eye technology, line judges systematically called fewer aces. In

[20] By summing the Hawk-Eye coefficient in Column 2 of Table 5, 5.738, with the coefficient of $HawkEye{\cdot}CE$ (-18.481) times 0.26, we get an impact of the Hawk-Eye on the ace ratio of 0.933. This value is close to the estimates reported in Table 4. Notice that, in Table 6, two additional interaction terms dissipate part of the data variability, i.e., $Break \cdot CE$ and $Treated \cdot CE$

contrast, by looking at the 2010 matches, we can notice that the ace ratio in treated courts does not significantly differ from the ace ratio in untreated courts. However, this happens because untreated courts experienced a significant increase in the number of assigned aces. Therefore, we cannot exclude that umpires learned from the use of the Hawk-Eye technology and translated their experience in untreated courts. Yet, the possibility that correction mechanisms generate learning effects and that these new abilities can be transferred to other situations is an interesting implication of the results that we leave for future research.

6.2. *Additional robustness tests*

Since tennis players usually prefer to play on specific surfaces and in specific tournaments, we re-estimate Equations (5) and (7) taking into account pairs-tournament specific effects. Formally, we estimate the following models:

$$Y_{pct} = \alpha_c + \alpha_t + \beta \cdot H_{ct} + \gamma \cdot X_{pct} + \delta_{ps} + e_{pct}, \tag{11}$$

and

$$Y_{pct} = \alpha_c + \alpha_t + \beta_1 \cdot H_{ct} + \beta_2 \cdot C_{ct} + \gamma \cdot X_{pct} + \delta_{ps} + e_{pct}, \tag{12}$$

where s denotes the Grand Slam tournaments. Both specifications restrict the number of non-singleton observations to pairs that played at least two times in the same tournament. For this reason, we consider them as a further robustness check. Panel A of Table 8 provides the coefficients of Equation (11), where we control for possible interactions between pairs of players and tournaments unobserved characteristics. Results are consistent with those presented in Table 3, and the treatment effect is even larger. As a final robustness test, in Panel B of Table 8, we estimate Equation (12), jointly controlling for possible interactions between pairs of players and tournaments unobserved characteristics and placebo court type. Again, results are in line with those presented in Tables 4 and 5, and the magnitude of the treatment effect is larger than the base specification estimates, similarly to results in Table 8.

INSERT TABLE 8 HERE

7. Welfare Implications

In section 2, we established that the introduction of a review system leads to an increase in definitive decisions if and only if the decision maker is characterized by DA. Our natural experiment provides evidence that there is indeed a significant variation in definitive decisions after the introduction of the review system, and this variation is positive, implying that $\triangle d(DA) > 0$. Based on this last result, we now establish under which conditions the introduction of a review system can lead to a welfare improvement by increasing the expected correctness of a decision. This may provide valid policy applications to real-world settings in which the revision technology is likely to be less accurate with respect to the Hawk-Eye system employed in professional sports.

Recalling our theoretical framework, we do not make any specific assumptions on the impact that the review system may have on the decision maker's behavior, beyond stating that it will weakly reduce the salience of definitive decisions. This is formally stated by expressions (2) and (3), that respectively state that the distance between the utility of correct definitive and non-definitive decisions weakly decreases, as does the distance between the disutility of incorrect definitive and non-definitive decisions. Depending on whether the magnitude of this variation is stronger for correct decisions versus incorrect decisions, this may either attenuate the bias, increase the decision maker's definitivity avoidance, or on the contrary, lead to an opposite bias inducing the decision maker to become definitivity loving. Whenever the review system increases the decision maker's bias by either making her more definitivity avoidant or definitivity loving, this negative effect may offset the positive effect that comes from reviewing potentially incorrect decisions, implying that the overall effect on welfare is indeterminate. We purposely avoid to add structure to the model leaving these two potential avenues open, guided by the observation that the impact of the review system on the actual behavior is not observable in our experimental setting.

In order to address this issue, we introduce a parameter B that denotes the percentage increase in the magnitude of the decision maker's bias caused by the introduction of the review system, and is defined by the following expression:

$$B = \frac{|F_1(s^*_H) + F_0(s^*_H) - 1| - |F_1(s^*_{NH}) + F_\omega(s^*_H) - 1|}{|[F_1(s^*_H) + F_0(s^*_H) - 1]|},$$

where $B > 0$ denotes an increase in the magnitude of the bias after the introduction of the review system, and $B < 0$ denotes a decrease in the bias. We therefore establish a general proposition on the impact of the review system on welfare:

Proposition 2. *A review system that leads to an increase in definitive decisions (i.e., $\triangle d(DA) > 0$) is always welfare improving unless it induces the decision maker to become significantly definitivity loving (i.e., for $s^*_H < s_{NB} < s^*_{NH}$ and $B > 0$). In this later case, the introduction of the review system is welfare improving if and only if the following two conditions are jointly satisfied: i) $p - F_1(2p-1) > B$, and; ii) $c < 1 - \frac{B}{p - F_1(2p-1)} = \overline{c}$ (Proof in the Appendix).*

The first part of the proposition comes from the observation that if the review system leads to an increase in the decision maker's definitivity avoidance bias, this increase cannot be too strong if it also leads to an increase in definitive decisions. Moreover, even if the review system induces an opposite bias leading the decision maker to become definitivity loving, as long as the magnitude of this bias is smaller than the initial one, implying that $B < 0$, the review system is always welfare improving.

Thus, the only scenario in which welfare may not increase occurs when the review system increases the magnitude of the bias leading the decision maker to become more definitivity loving than she was definitivity avoidant, implying that $B > 0$ and $s^*_{NH} > s^*_{NB} > s^*_H$. In this case, the necessary (but not sufficient) condition for the system to improve welfare is that $p - F_1(2p - 1) > B$, implying that given the precision of the unbiased signals received by the agent that may potentially contest the decision $(1 - F_1)$, the revision technology (p) is sufficiently precise in relation to the increase in the bias (B). Intuitively, when the bias increases, in order to improve the expected correctness of the decision, the review system must be more accurate. Whenever this necessary condition is satisfied, it must also be that the cost of invoking the challenge, c is smaller than a threshold value $\overline{c}$, where this threshold is increasing in p. This implies that the less precise is the review technology (for low p), the less costly it must be to call upon the review system (c must be lower).

The above proposition suggests that, although reductions in efficiency may occur only in the extreme circumstances in which the introduction of the system induces the decision maker to become more biased in the opposite direction, some caution must be taken in adopting such a system in the attempt to improve efficiency. Thus for example, in the case of financial institutions, if there is a chance that introducing a rule that allows interested parties to invoke an audit committee can lead decision makers to excessively increase definitive decisions (this may, for instance, involve denying to roll over debt even when private information suggests not to, thus engaging in too little risk), our welfare analysis suggests that the review system should be introduced only if certain conditions are satisfied. First of all, the system should be applied only to decisions for which the information available to audit committees involves relatively little noise. So for instance, in periods of high volatility in which systemic risk is muddled with individual business risk, adopting a review system may not be a valid option. Moreover, the less precise is the information available to the audit committee to perform the review, the less costly it must be for the financial institution's shareholders to call on the audit committee for the review.

8. Discussion

8.1. *Definitivity Avoidance and Other Biases*

Psychologists and economists have identified relevant behavioral biases such as status quo bias, omission bias, inaction inertia, and decision deferral, interpreting these biases in terms of reluctance towards making a decision. Here, we briefly discuss the relationship between these biases and our result, arguing that definitivity avoidance may generate situations in which the mentioned biases are observed. For instance, when economic agents have to act in order to make a definitive decision, we may observe an omission or a status quo bias as a consequence of definitivity avoidance. Ritov and Baron (1998) show that people are reluctant to vaccinate a (hypothetical) child, even when the probability of dying as a consequence of the vaccination is lower than the probability of dying from the disease. According to Ritov and Baron, the decision not to vaccinate may be the result of an omission; however, because vaccination is a definitive decision, definitivity avoidance may explain the reluctance to act. Similarly, managers that are subject to omission bias tend not to hedge losses even when the probability of a loss is higher than derivatives losses (Hirshleifer, 2008); nonetheless, corporate hedging may be perceived as a more definitive decision than not hedging, and therefore this bias may be generated by definitivity avoidance. Another suitable example of status quo bias that can be explained by definitivity avoidance is discussed in Westman (1991). According to Westman, children's rights in the United States began to emerge during the twentieth century. However, the protection of children's interests is still an issue: "Children can be placed in foster care, and parental rights can be permanently terminated. The state exercises responsibility for determining custody in divorce cases and for establishing a legal parent-child relationship through adoption. Unfortunately, criteria for making these decisions are not well-defined, so that the general practice is to exercise judicial restraint and perpetuate the status quo rather than resolve issues in a timely and *definitive* manner for a child's benefit. For example, many youngsters spend years in foster care, because no one has assumed the responsibility for making the definitive decisions that are necessary in the legal pursuit of their interests." (p. 47). All these examples clarify that at least part of the decision avoidance behaviors may be related to the nature of the underlying decision instead of the modes associated with decision making. In particular, when decisions differ in terms of definitiveness, economic agents tend to prefer the less definitive choice independently of whether they have to act or not in order to select it. Indeed, our experiment requires umpires to make an action to reveal their bias, and this removes the emphasis on omission or inertia. That is, in many situations, decision-makers do not avoid decisions per se but just those that involve making a final judgment.

8.2. Strategic Behavior vs. Behavioral Bias

Our model assumes that points are independent and identically distributed. Empirically, in Section 5, we have shown that this assumption holds reasonably well under some conditions that are satisfied by our empirical design. However, strategic behavior by players may hinder the validity of this assumption. Assuming that tennis players can modulate the speed of their serves, where more rapid serves are also riskier, Gillman (1985) shows that a serving strategy based on a risky serve first, and eventually, a safer second serve is optimal. Now, suppose that players change their service strategy according to the importance of the point. For instance, Anbarci et al. (2018) show that when behind in score, servers become loss averse and take more risk than when they are ahead, increasing the speed of the service. By knowing this fact, an umpire might formulate a prior probability that the ball is in and use the signal to infer the posterior through Bayesian updating.[21] Similarly, strategic players could anticipate the umpire's belief reducing the speed of the serve when the Hawk-Eye technology is absent, and they cannot challenge the umpire's call. Vice versa, with the availability of the Hawk-Eye, rational servers should take a risk in difficult times more easily. This change in their

[21] Green and Daniels (2018) find that, in baseball matches, umpires adopt similar reasoning to identify the strike zone. Indeed, with respect to the service box, the strike zone cannot be directly observed by umpires, and therefore they use the current number of strikes and balls to formulate a prior on the probability of the pitch being in or out and update this prior with the signal.

service strategy could lead to an increase in the number of aces as well as in the number of errors in the first serve. In this case, a deviation of the call from the signal would not be a sign of psychological bias but the result of rational behavior.[22] Therefore, to test whether the introduction of the Hawk-Eye affected players' serving strategy in this direction, we repeated our analysis for the ratio of valid first serves. However, we did not find significant changes in the fraction of first serves in when we use pair fixed effects and a positive effect when we use court fixed effects. This result allows us to reject the hypothesis of a change in players' strategy after the introduction of the Hawk-Eye.

8.3. Learning by exposure

Another relevant question concerns the introduction of the review system. In particular, we might want to understand whether the review system simply neutralizes the effect of the bias that continues to characterize behavior, or if the exposure to such a system may actually make decision-makers more aware of their inefficient behavior, therefore inducing them to attempt to overcome the bias and improve the quality of their decisions. Following Tetlock and Gardner (2015), we may say that inducing experts to focus on the correctness of their decisions might improve upon their ability to make better decisions. Although this is beyond the scope of our analysis, the empirical evidence in Tables 7 and 8 suggests that this may be the case. Notice indeed, that as shown in the last columns of Tables 7 and 8 the treatment effect disappears, but the control group (those not exposed to the review system) displays a number of aces that does not significantly differ from those of the treated group when the technology was first introduced (i.e., 2006-07). This suggests that once decision-makers have been exposed to a review system, this makes them aware of their bias and induces them to correct their behavior even when the review system is absent.

[22] This prediction is consistent with Klaassen and Magnus (2001) that fined that, at important points, it is more difficult for servers to win the point, although they do not specifically focus on the probability of scoring an ace.

9. Conclusion

In this paper, we propose that decision-makers may be subject to a definitivity avoidance bias, which leads valuable information to be disregarded, therefore producing sub-optimal decisions. More specifically, this bias leads agents that are called on to make decisions that have material consequences for other parties to refrain from following their (imperfect) private information when this involves making definitive decisions. By exploiting the introduction of a review system that allows tennis players to challenge the decision of the officials by invoking the use an impartial monitoring technology that can overturn incorrect calls, we are able to identify the existence of the bias and establish that such a review system may lead to its attenuation.

This natural experiment suggests that in all those contexts in which definitivity is salient, welfare gains may arise from introducing a review system that allows an agent that is affected by the consequence of the judgment to call for a revision of the decision by a neutral third party. Nonetheless, our analysis shows that a review system may not always be welfare improving if it induces agents to overreact, becoming definitivity lovers in very uncertain decisions. An interesting avenue for future research is, therefore, to explore how the introduction of review systems affects the behavior of decision makers in different contexts in order to draw more accurate indications for the design of welfare-improving decision review systems.

Moreover, identifying neutral third party reviewers in the real world that have the same desirable features of the Hawk-Eye system, namely competence and neutrality, may not be a simple task. In this respect, the growing use of artificial intelligence may provide a valid tool for designing non-human review systems. For instance, the use of intelligent algorithms to evaluate the *ex-post* correctness of decisions based on objective parameters combined with vastly available datasets may serve the purpose of producing real-world equivalents of the Hawk-Eye system in professional tennis.[23]

[23] Along these lines, Chen (2019) suggests how to create decision support systems for judges that combine

References

Anbarci, N., Arin, K. P., Kuhlenkasper, T., and Zenker, C. (2018). Revisiting loss aversion: Evidence from professional tennis. *Journal of Economic Behavior and Organization*, 153(1):1–18.

Anderson, C. J. (2003). The psychology of doing nothing: forms of decision avoidance result from reason and emotion. *Psychological bulletin*, 129:139–167.

Angrist, J. D. and Pischke, J.-S. (2008). *Mostly harmless econometrics: An empiricist's companion*. Princeton university press.

Antoniou, C. and Mavis, C. (2019). Do beliefs reflect information reliability? evidence from odds of tennis matches. *Working Paper, Available at SSRN 2757037*.

Arellano, M. (2003). *Panel data econometrics*. Oxford university press.

Bar-Eli, M., Azar, O.H., a. R. I., Keidar-Levine, Y., and Schein, G. (2007). Action bias among elite soccer goalkeepers: The case of penalty kicks. *Journal of Economic Psychology*, 28(3):606–621.

Bernanke, B. S. (1989). Is there too much corporate debt? *Business Review*, (Sep):3–13.

Cameron, A. C., Gelbach, J. B., and Miller, D. L. (2008). Bootstrap-based improvements for inference with clustered errors. *The Review of Economics and Statistics*, 90(3):414–427.

Carey, M., Kashyap, A. K., Rajan, R., and Stulz, R. M. (2012). Market institutions, financial market risks, and the financial crisis. *Journal of Financial Economics*, 104(3):421 – 424.

Chen, D. L. (2019). Machine learning and the rule of law. *Computational Analysis of Law*, 27(1):15–42.

Cohen-Zada, D., Krumer, A., and Shapir, O. M. (2018). Testing the effect of serve order in tennis tiebreak. *Journal of Economic Behavior & Organization*, 146:106–115.

Dodd-Frank (2010). Wall street reform and consumer protection act. Pub. L. No. 111-203, §929-Z, 124 Stat. 1376, 1871 (2010) (codified at 15 U.S.C. §78o) [Bluebook R. 12.4].

Ely, J., Gauriot, R., and Page, L. (2017). Do agents maximise? risk taking on first and second serves in tennis. *Journal of Economic Psychology*, 63:135–142.

Garicano, L., Palacios-Huerta, I., and Prendergast, C. (2005). Favoritism under social pressure. *Review of Economics and Statistics*, 87:208–216.

George, S. L. (1973). Optimal strategy in tennis: A simple probabilistic model. *Journal of the Royal Statistical Society*, 22(1):97–104.

Gilbert, D. T. and Ebert, J. E. J. (2002). Decisions and revisions: The affective forecasting of changeable outcomes. *Journal of Personality and Social Psychology*, 82:503–514.

large datasets with artificial intelligence in order to correct for behavioral biases.

Gillman, L. (1985). Missing more serves may win more points. *Mathematics Magazine.*

Green, E. and Daniels, D. P. (2018). Bayesian instinct. *Available at SSRN 2916929.*

Hirshleifer, D. (2008). Psychological bias as a driver of financial regulation. *European Financial Management*, 14(5):856–874.

Hsiao, C., Ching, S. H., and Ki, W. S. (2012). A panel data approach for program evaluation: measuring the benefits of political and economic integration of hong kong with mainland china. *Journal of Applied Econometrics*, 27:705–740.

Ibragimov, R. and Müller, U. K. (2010). t-statistic based correlation and heterogeneity robust inference. *Journal of Business & Economic Statistics*, 28:453–468.

Kahneman, D. and Tversky, A. (1979). Prospect Theory: An Analysis of Decision under Risk. *Econometrica*, 47(2):263–291.

Kang, J. D. Y. and Schafer, J. L. (2007). Inverse probability weighted estimation for general missing data problems. *Statistical Science*, 22:523–539.

Klaassen, F. and Magnus, J. (2001). Are points in tennis independent and identically distributed? evidence from a dynamic binary panel data model. *Journal of the American Statistical Association*, 96:500–509.

Klaassen, F. and Magnus, J. (2009). The efficiency of top agents: An analysis through service strategy in tennis. *Journal of Econometrics*, 148(1):72–85.

Kovalchik, S. A., Sackmann, J., and Reid, M. (2017). Player, official or machine?: uses of the challenge system in professional tennis. *International Journal of Performance Analysis in Sport*, 17(6):961–969.

Leipold, A. D. (2005). Why are federal judges so acquittal prone? *Washington University Law Journal*, 83(151).

Massey, C. and Thaler, R. (2013). The loser's curse: Overconfidence vs. market efficiency in the national football league draft. *Management Science*, 59(7):1479–1495.

Mather, G. (2008). Perceptual uncertainty and line-call challenges in professional tennis. *Proceedings of the Royal Society of London B: Biological Sciences*, 275:1645–1651.

Pope, D. and Schweitzer, M. (2011). Is tiger woods loss averse? persistent bias in the face of experience, competition, and high stakes. *American Economic Review*, 101:129–157.

Ritov, I. and Baron, J. (1998). Status quo and omission biases. *Journal of Risk and Uncertainty*, pages 49–62.

Robins, J. M., Sued, M., Lei-Gomez, Q., and Rotnitzky, A. (2007). Comment: Performance of double-robust estimators when "inverse probability" weights are highly variable. *Statistical Science*, 22:544–559.

Rodenberg, R. M., Sackmann, J., and Groer, C. (2016). Tennis integrity: a sports law analytics review. *The International Sports Law Journal*, 16(1-2):67–81.

Romer, D. (2006). Do firms maximize? evidence from professional football. *Journal of Political Economy*, 114:340–365.

Sacheti, A., Gregory-Smith, I., and Paton, D. (2015). Home bias in officiating: Evidence from international cricket. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 3:741–755.

Samuelson, W. and Zeeckhauser, R. (1998). Status quo bias in decision making. *Journal of Risk and Uncertainty*, 1:7–59.

Słoczyński, T. and Wooldridge, J. M. (2018). Inverse probability weighted estimation for general missing data problems. *Econometric Theory*, 34:112–133.

Tetlock, P. and Gardner, D. (2015). *Superforecasting: The art and science of prediction*. New York: Crown.

Tsiros, M. and Mittal, V. (2000). Regret: A model of its antecedents and consequences in consumer decision making. *Journal of Consumer Research*, 26:401–417.

Tykocinski, O. E., Pittman, T. S., and Tuttle, E. S. (1995). Inaction inertia: Foregoing future benefits as a result of an initial failure to act. *Journal of Personality and Social Psychology*, pages 793–803.

Walker, M. and Wooders, J. (2001). Minimax play at wimbledon. *American Economic Review*, 91(5):1521–1538.

Westman, J. C. (1991). *Who speaks for the children? The handbook of individual and class child advocacy.* Professional Resource Exchange, Inc.

Whitney, D., Wurnitsch, N., Hontiveros, B., and Louie, E. (2008). Perceptual mislocalization of bouncing balls by professional tennis referees. *Current Biology*, 18:R947–R949.

Wooldridge, J. M. (2007). Inverse probability weighted estimation for general missing data problems. *Journal of Econometrics*, 141:1281–1301.

Wooldridge, J. M. (2010). *Econometric Analysis of Cross Section and Panel Data*. Cambridge: MIT press, ii edition.

Zeelenberg, M., Beattie, J., Van der Pligt, J., and De Vries, N. (1996). Consequences of regret aversion: Effects of expected feedback on risky decision making. *Organizational Behavior and Human Decision Processes*, 65:148–158.

Table 1: **Matches in treated courts before and after the treatment**

	2002	2003	2004	2005	2006	2007	2008	2009	2010	Total
Baseline sample (only matches on grass and hard surface)										
Without Hawk Eye	31	56	63	79	70	42	33	41	30	445
With Hawk Eye	0	0	0	0	16	48	48	45	37	194
Total	31	56	63	79	86	90	81	86	67	639
Enlarged sample (with matches on clay surface)										
Without Hawk Eye	38	66	69	99	85	55	47	55	38	552
With Hawk Eye	0	0	0	0	17	52	55	51	43	218
Clay surface	18	16	20	29	33	32	26	36	30	240
Total	56	82	89	128	135	139	128	142	111	1,010

Notes: This table shows the number of matches played in treated and untreated courts before and after the introduction of the Hawk-Eye technology.

Table 2: **Descriptive Statistics**

	Mean	SD	Min	25%	50%	75%	Max
Ace ratio	7.529	3.870	0	4.663	7.041	9.934	33.333
Hawk Eye	0.362	0.481	0	0	0	1	1
Clay	0.237	0.426	0	0	0	0	1
Grass	0.236	0.425	0	0	0	0	1
Hard	0.527	0.500	0	0	1	1	1
Break	0.549	0.498	0	0	1	1	1
Favorite Rank	16.738	19.128	1	3	10	24	134
Challenger Rank	65.250	70.326	2	26	52.5	87	1141
Favorite Age	25.195	3.032	18.626	22.976	25.050	27.146	36.534
Challenger Age	25.507	3.743	16.572	22.773	25.166	28.246	36.726
Home player	0.143	0.350	0	0	0	0	1
Minutes	148.257	48.834	6	112	142	182	393
Clay Experience (CE)	0.260	0.133	0	0.195	0.255	0.310	1
OBS.	1,010	1,010	1,010	1,010	1,010	1,010	1,010

Notes: This table presents descriptive statistics for our selected variables. Variables are defined as follows: Ace Ratio is a variable that captures the total number of aces over the total number of served points; Hawk-Eye is a dummy variable that takes a value of 1 if the match is played with Hawk-Eye technology in place; Clay, Grass and Hard are dummy variables taking the value of 1 for the type of court the match has been played on; Favorite and Challenger Rank(Age) capture the ranking(age) of the highest and lowest seeded (oldest and youngest) player in the match, respectively; Home player indicates whether one of the two players comes from the country organizing the tournament; Minutes is the length of the match.

Table 3: **Pairwise Correlation Coefficients**

	#	1	2	3	4	5	6	7	8
Ace ratio	1	1							
Hawk Eye	2	0.204***	1						
Favorite Rank	3	-0.031	-0.295***	1					
Challenger Rank	4	-0.031	-0.150***	0.311***	1				
Favorite Age	5	0.05	-0.073**	0.126***	0.048	1			
Challenger Age	6	0.076**	0.055*	0.041	0.094***	0.162***	1		
Home player	7	0.051	0.061*	-0.013	0.036	-0.017	-0.014	1	
Minutes	8	-0.057*	0.05	-0.002	-0.128***	-0.014	0.022	0.027	1
Clay Experience	9	-0.180***	0.029	-0.020	-0.014	0.012	0.004	-0.061*	0.046

Table 4: **Hawk Eye effect on Ace Ratio**

	2006-2010	2007-2010	2008-2010	2009-2010	2010
			Panel A. With pair FEs (N=639)		
Hawk Eye	1.131*	1.350**	1.508**	1.378*	-0.692
	(0.676)	(0.658)	(0.642)	(0.701)	(0.952)
Favorite Rank	0.005	0.004	0.002	0.004	0.003
	(0.015)	(0.014)	(0.014)	(0.014)	(0.014)
Challenger Rank	-0.005**	-0.005**	-0.005*	-0.005**	-0.005**
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
Favorite Age	0.252*	0.199	0.199	-0.031	0.066
	(0.151)	(0.156)	(0.150)	(0.103)	(0.101)
Challenger Age	0.277**	0.220	0.208	-0.009	0.083
	(0.137)	(0.144)	(0.138)	(0.101)	(0.100)
Home player	-0.355	-0.326	-0.333	-0.336	-0.334
	(0.568)	(0.573)	(0.573)	(0.569)	(0.569)
Minutes	0.001	0.001	0.001	0.001	0.000
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
adj. R2	0.540	0.541	0.543	0.541	0.537
within R2	0.158	0.160	0.163	0.161	0.153
Pair FE	Yes	Yes	Yes	Yes	Yes
Court and Time FE	Yes	Yes	Yes	Yes	Yes
			Panel B. Without pair FEs (N=639)		
Hawk Eye	0.781*	0.979**	1.638**	1.637*	0.189
	(0.471)	(0.426)	(0.634)	(0.921)	(0.634)
Favorite Rank	-0.004	-0.003	-0.001	-0.004	-0.009
	(0.010)	(0.010)	(0.010)	(0.009)	(0.008)
Challenger Rank	-0.002	-0.002	-0.002	-0.002	-0.003
	(0.005)	(0.005)	(0.005)	(0.005)	(0.004)
Favorite Age	0.021	0.024	0.030	0.020	0.011
	(0.053)	(0.053)	(0.057)	(0.054)	(0.046)
Challenger Age	0.020	0.021	0.025	0.022	0.016
	(0.029)	(0.029)	(0.028)	(0.029)	(0.028)
Home player	1.514***	1.512***	1.466***	1.523***	1.642***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Minutes	-0.006	-0.006	-0.006	-0.006	-0.006
	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
adj. R2	0.074	0.076	0.083	0.079	0.069
within R2	0.122	0.124	0.131	0.127	0.118
Court and Time FE	Yes	Yes	Yes	Yes	Yes

Notes: This table reports the estimates of Equations (5) and (6) for different treatment periods. The dependent variable is the aces to points ratio measured for pair p, in court c at time t. All regressions include courts and time dummies absorbing for the direct effects of the treatment group and period. Hawk Eye is a dummy indicating whether the Hawk-Eye technology was available in court c at time t. In Panel A, we also consider a vector of pair FEs. Standard errors are clustered at the pair level in Panel A and based on 1,000 wild-bootstrap replications in Panel B. Significance levels: *10%, **5%, ***1%.

Table 5: **Hawk Eye effect on Ace Ratio: placebo tests**

	2006-2010	2007-2010	2008-2010	2009-2010	2010
Panel A. With pair FEs (N=1010)					
Hawk Eye	0.842	1.066*	1.203**	0.607	-1.159
	(0.572)	(0.556)	(0.528)	(0.610)	(0.835)
Clay courts	-1.355	-0.823	-0.807	0.701	-
	(2.553)	(2.573)	(2.541)	(1.514)	-
adj. R2	0.576	0.577	0.577	0.575	0.576
within R2	0.155	0.157	0.159	0.153	0.155
Additional controls	Yes	Yes	Yes	Yes	Yes
Pair FE	Yes	Yes	Yes	Yes	Yes
Court and Time FE	Yes	Yes	Yes	Yes	Yes
Panel B. Without pair FEs (N=1010)					
Hawk Eye	1.076**	1.215**	1.854***	1.502*	0.079
	(0.529)	(0.493)	(0.000)	(0.857)	(1.014)
Clay courts	-3.411	-3.411	-3.395	-0.715	-2.034
	(3.052)	(3.052)	(3.063)	(1.173)	(2.633)
adj. R2	0.176	0.177	0.183	0.175	0.168
within R2	0.210	0.211	0.217	0.209	0.203
Additional controls	Yes	Yes	Yes	Yes	Yes
Court and Time FE	Yes	Yes	Yes	Yes	Yes

Notes: This table presents a set of DDD estimates for Equation (7), where Clay courts matches constitute a placebo category. The dependent variable is the aces to points ratio measured for pair p, in court c at time t. All regressions include a set of time-court varying pair's characteristics potentially affecting the ace ratio (i.e., players ages, ranking, and home-field advantage) as well as courts and time dummies absorbing for the direct effects of the treatment group and the Hawk-Eye introduction period. Hawk Eye is a dummy indicating whether the Hawk-Eye technology was available in court c at time t. In Panel A, we also consider a vector of pair FEs. Standard errors are clustered at the pair level in Panel A and based on 1,000 wild-bootstrap replications in Panel B. Significance levels: *10%, **5%, ***1%.

Table 6: **Hawk Eye effect on Ace Ratio: player experience on placebo court**

	2006-2010	2007-2010	2008-2010	2009-2010	2010
Panel A. With pair FEs (N=639)					
Hawk Eye	5.363**	5.738***	3.513**	3.406**	0.362
	(2.361)	(2.035)	(1.645)	(1.719)	(2.110)
$HawkEye \cdot CE$	-18.683*	-18.481**	-7.441	-7.719	-3.995
	(9.506)	(7.719)	(6.067)	(6.297)	(6.930)
$Break \cdot CE$	-1.476	-0.725	4.171	0.920	0.183
	(6.477)	(6.238)	(3.736)	(3.660)	(4.036)
$Treated \cdot CE$	13.319	11.036	2.363	2.357	-0.750
	(9.594)	(7.692)	(6.112)	(5.877)	(4.664)
adj. R2	0.546	0.553	0.541	0.541	0.533
within R2	0.041	0.058	0.032	0.032	0.016
Additional controls	Yes	Yes	Yes	Yes	Yes
Pair FE	Yes	Yes	Yes	Yes	Yes
Court and Time FE	Yes	Yes	Yes	Yes	Yes
Panel B. Without pair FEs (N=639)					
Hawk Eye	2.015*	2.445*	3.129**	2.952*	2.630
	(1.142)	(1.298)	(1.270)	(1.683)	(2.789)
$HawkEye \cdot CE$	-5.782*	-6.260**	-4.589*	-4.827	-8.630
	(3.447)	(3.072)	(2.702)	(3.134)	(6.164)
$Break \cdot CE$	-0.576	-0.379	1.297	-0.592	1.903
	(3.362)	(3.968)	(1.787)	(2.572)	(1.961)
$Treated \cdot CE$	1.179	0.765	-1.330	-0.555	0.100
	(2.765)	(2.418)	(2.665)	(3.555)	(0.737)
adj. R2	0.080	0.084	0.086	0.081	0.069
within R2	0.132	0.136	0.138	0.133	0.121
Additional controls	Yes	Yes	Yes	Yes	Yes
Court and Time FE	Yes	Yes	Yes	Yes	Yes

Notes: This table presents results of a set of the DD model specified in Equation (8). Here, we interacted the Hawk-Eye effect with the players' experience on clay (CE). The dependent variable is the aces to points ratio measured for pair p, in court c at time t. All regressions include a set of time-court varying pair's characteristics potentially affecting the ace ratio (i.e., players ages, ranking, and home-field advantage) as well as courts and time dummies absorbing for the direct effects of the treatment group and the Hawk-Eye introduction period. Hawk Eye is a dummy indicating whether the Hawk-Eye technology was available in court c at time t. In Panel A, we also consider a vector of pair FEs. Standard errors are clustered at the pair level in Panel A and based on 1,000 wild-bootstrap replications in Panel B. Significance levels: *10%, **5%, ***1%.

Table 7: **Sequential IPWRA estimates**

	2006	2007	2008	2009	2010
ATT	0.812	1.437***	1.437**	1.538**	0.511
	(1.076)	(0.480)	(0.621)	(0.775)	(0.740)
Control group mean	7.646***	7.348***	7.494***	8.159***	9.066***
	(0.880)	(0.365)	(0.505)	(0.605)	(0.597)
N	252	502	508	504	505

Notes: This table provides results for an alternative method to estimate the average treatment effect considering a panel as a sequence of cross-sectional natural experiments. We estimate the double-robust estimator proposed in Wooldridge (2007) for each year separately. Robust-clustered standard errors are in parentheses. Significance levels: *10%, **5%, ***1%.

Table 8: **Pair-tournament effects**

	2006-2010	2007-2010	2008-2010	2009-2010	2010
Panel A. With pair-tournament FEs (N=208)					
Hawk Eye	1.025	3.249**	3.492**	3.107*	-1.730
	(1.583)	(1.415)	(1.733)	(1.810)	(1.815)
adj R2	0.503	0.522	0.530	0.524	0.506
within R2	0.385	0.408	0.418	0.410	0.388
Additional controls	Yes	Yes	Yes	Yes	Yes
Pair-Tournment FE	Yes	Yes	Yes	Yes	Yes
Court and Time FE	Yes	Yes	Yes	Yes	Yes
Panel B. With pair-tournament FEs and placebo court type (N=293)					
Hawk Eye	0.490	2.740*	3.415**	3.106*	-1.741
	(1.624)	(1.423)	(1.658)	(1.658)	(1.816)
Clay courts	-11.390	-7.394	-5.503		5.688*
	(6.965)	(6.996)	(6.935)		(3.263)
adj R2	0.573	0.581	0.589	0.587	0.576
within R2	0.355	0.368	0.379	0.376	0.359
Additional controls	Yes	Yes	Yes	Yes	Yes
Pair-Tournment FE	Yes	Yes	Yes	Yes	Yes
Court and Time FE	Yes	Yes	Yes	Yes	Yes

Notes: This table reports the estimates of Equations (11) and (12), considering different treatment periods. Now, pair-tournament FEs absorb the effect of the home player variable and control for possible interactions between pairs of players and tournaments unobserved characteristics. Panel B also considers clay courts as a placebo control group. Robust-clustered standard errors are in parentheses. Significance levels: *10%, **5%, ***1%

Fig. 1. Definitivity Avoidance and Prospect Theory

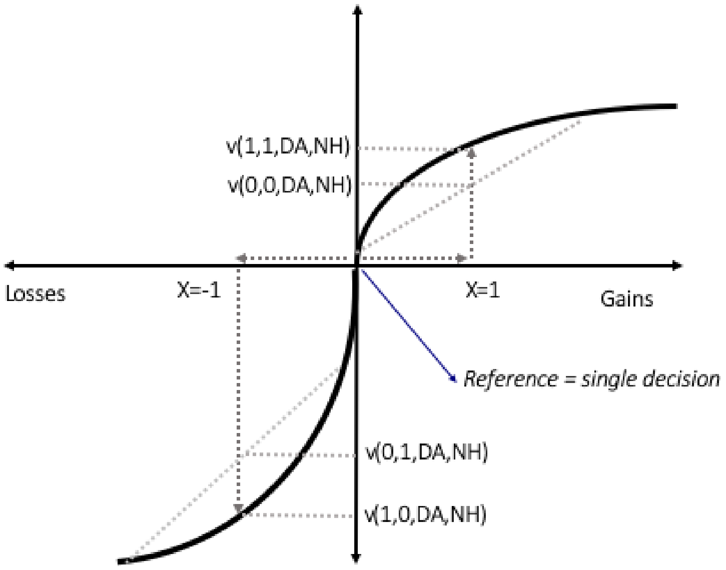
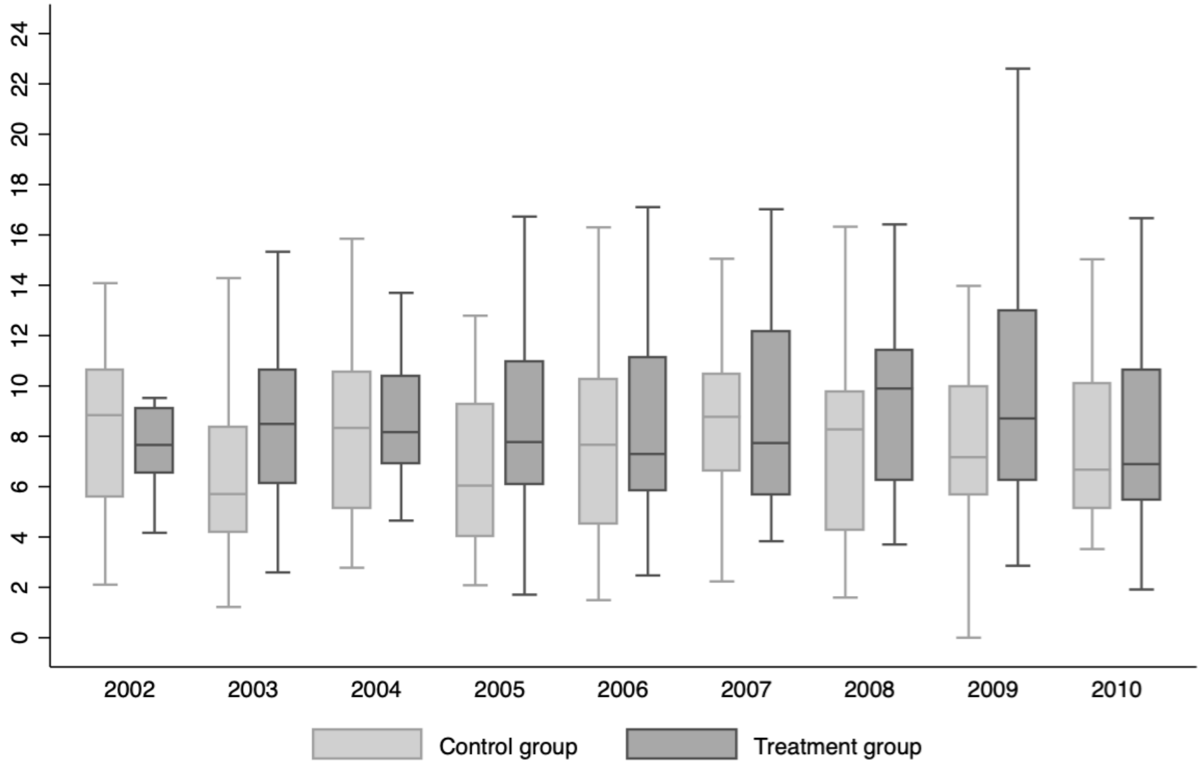


Fig. 2. Distribution of aces per match in control and treatment group over time
This figure reports the box-plot distribution of ace ratios over the sample period. Ace ratio is measured as the total number of aces over the total number of served points and is reported in percentage points on the vertical axis



Appendix A. Proofs

A.1. Proof of Prediction 1

We begin by deriving the expression for $\Delta d(b)$. The expressions for the expected decision with and without the review system are the following

$$E[d(b,H)] = 1/2\left[(1-F_1(s^*_H))[(1-F_1(s^*_{NB}))+F_1(s^*_{NB})(1-c)p]\right]$$

$$+F_1(s^*_H)[((1-F_1(s^*_{NB}))(1-c)p)]+$$

$$+(1-F_0(s^*_H))[(1-F_0(s^*_{NB}))+F_0(s^*_{NB})(1-c)(1-p)]$$

$$+F_0(s^*_H)[(1-F_0(s^*_{NB}))(1-c)(1-p)]$$

$$E[d(b,NH)] = 1/2\left[(1-F_1(s^*_{NH}))+(1-F_0(s^*_{NH}))\right].$$

So, simplifying $\Delta d(b)$ can be rewritten in the following way:

$$\Delta d(b) = 1/2[\sum_{\omega}(F_\omega(s^*_{NH})-F_\omega(s^*_H))]+$$

$$+\left[(F_0(s^*_H)+F_1(s^*_H)-1)(1-c)(F_1(s^*_{NB})(1-p)+(F_0(s^*_{NB})p)\right]. \tag{13}$$

Now using Proposition 1 (i.e., $s^*_{NH} \geq s^*_{NB}$ where $s^*_{NH} > s^*_{NB}$ implies that there is DA and $s^*_{NH} = s^*_{NB}$ implies that there is no bias) and the symmetry of the signal structure $(1-F_1(s^*_{NB})) = F_0(s^*_{NB})$, notice that if a bias exists there are 3 possible cases that may arise in which $\Delta d(b) > 0$ ($\Delta d(b) < 0$):

case 1) $s^*_H = s^*_{NB}$ so if there is a bias it is completely corrected by the review system, which implies that the 2nd term of (13) is equal to 0, and for $\Delta d(b) > 0$ ($\Delta d(b) < 0$) it must be that $s_H^* < s_{NH}^*$ ($s_H^* > s_{NH}^*$) which implies that there must be DA.

case 2) $s_H^* > s_{NB}^*$ which implies that there is DA, and the 2nd term of (13) is greater than 0, so $\Delta d(b) > 0$ ($\Delta d(b) < 0$) is consistent with $s_H^* \gtrless s_{NH}^*$ ($s_{NH}^* < s_H^*$). In other words the review system may lead to an uncertain variation (decrease) in the propensity to make definitive decisions with respect to the pre-review system level.

case 3) $s_H^* < s_{NB}^*$ which implies that there is DA and the 2nd term of (13) is less than 0, so for $\Delta d(b) > 0$ ($\Delta d(b) < 0$) it must be that $s_H^* < s_{NB}^* < s_{NH}^*$ ($s_H^* \lesseqgtr s_{NH}^*$) implying that the review system induces an increase (uncertain variation) in definitive decisions.

This completes the proof of our empirical prediction, namely that a variation in definitive decisions due to the introduction of the review system may occur only if the decision maker is characterized by DA before the review system is introduced.

A.2. Proof of Proposition 2

We define $\Delta L(r) =\mid E(d(r)) - E(d(NB)) \mid$, which represents the expected informational loss of each regime r. Therefore, the condition for the review system to be welfare improving is the following:

$$\Delta L(H) \leq \Delta L(NH) \tag{14}$$

We again consider 3 cases:

case 1) $s_H^* = s_{NB}^*$ in this case $\Delta L(H) = 0$ and therefore welfare maximizing for all values of c and p.

case 2) $s_H^* > s_{NB}^*$. If $s_{NH}^* > s_H^*$ the review system reduces the bias and is therefore always welfare improving. If instead $s_H^* > s_{NH}^* > s_{NB}^*$ the review system induces a greater bias in the decision maker's behavior (more definitivity avoidance) and we must find if there exist values of c and p such that the review system can be welfare improving.

case 3) $s_H^* < s_{NB}^*$ in this case as implied by the proof of the empirical prediction, it must be that $s_{NH}^* > s_{NB}^* > s_H^*$ meaning that regime H inverts the bias and makes the decision maker definitivity loving. In this case, again we must verify for which values of c and p the system guarantees an increase in welfare.

First notice that

$$\Delta L(H) = \left|1/2[1 - F_1(s_H^*) - F_0(s_H^*)](1 - (1-c)(p - F_1(s_{NB}^*)(2p-1))\right|,$$

and

$$\Delta L(NH) = \left|1/2[1 - F_1(s_{NH}^*) - F_0(s_{NH}^*)]\right|.$$

Now defining

$$(1-B) = \left|\frac{[F_1(s_{NH}^*) + F_0(s_{NH}^*) - 1]}{[F_1(s_H^*) + F_0(s_H^*) - 1]}\right| > 0,$$

where B denotes percentage increase in bias of the decision maker after the introduction of the review system, and

$$F_1(s_{NB}^*) \equiv F_1 < 1/2,$$

where the last inequality follows from the assumption that unbiased signals are informative. Condition (14) can therefore be written as:

$$1 - (1-c)(p - F_1(2p-1)) < (1-B) \tag{15}$$

We first consider Case 2 ($s_H^* > s_{NH}^* > s_{NB}^*$)

Recalling the expression for $\Delta d(b)$ and using the finding that $\Delta d(b) > 0$, it follows that:

$$(1-c)(p - F_1(2p-1)) > \frac{|[\sum_\omega (F_\omega(s_{NH}^*) - F_\omega(s_H^*))]|}{[F_1(s_H^*) + F_0(s_H^*) - 1]}$$

Since both sides of the above expressions are positive and the right hand side is less than one, it can be rewritten as follows:

$$1 - (1-c)(p - F_1(2p-1)) < 1 - \frac{|[\sum_\omega (F_\omega(s_{NH}^*) - F_\omega(s_H^*))]|}{[F_1(s_H^*) + F_0(s_H^*) - 1]}.$$

Now it is straightforward to show that

$$1-\frac{|[\sum_{\omega}(F_{\omega}(s^*_{NH})-F_{\omega}(s^*_{H}))]|}{[F_1(s^*_H)+F_0(s^*_H)-1]}=(1-B),$$

which allows to state that (15) is always satisfied, implying that the introduction of the review system is always welfare improving

We now consider case 3 $(s^*_{NH}>s^*_{NB}>s^*_{H})$

In this case, $(1-B)>0$, but it can also be greater than 1. In this latter case, $B<0$ (i.e., the introduction of the review system reduces the bias) implies that condition (15) is always satisfied. In the other case, in which $(1-B)\in(0,1)$ it is once again necessary to analyze when (15) is satisfied as for case 2.

In the case in which $(1-B)\in(0,1)$ however, using the finding that $\Delta d(b)>0$ does not allow us to rule out cases in which the review system may lead to a decrease in welfare. To see this, notice that using the expression for $\Delta d(b)$ and the finding that $\Delta d(b)>0$ implies that:

$$1-(1-c)(p-F_1(2p-1))>1-\frac{[\sum_{\omega}(F_{\omega}(s^*_{NH})-F_{\omega}(s^*_{H}))]}{[F_1(s^*_H)+F_0(s^*_H)-1]},$$

Moreover, it straightforward to show that since $[\sum_{\omega}(F_{\omega}(s^*_{NH})-F_{\omega}(s^*_{H}))]>[F_1(s^*_H)+F_0(s^*_H)-1]>0$

$$1-\frac{[\sum_{\omega}(F_{\omega}(s^*_{NH})-F_{\omega}(s^*_{H}))]}{[F_1(s^*_H)+F_0(s^*_H)-1]}<0<(1-B).$$

which does not allow us to state whether or not (15) is satisfied.

Therefore, in order to pin down when the review system may be welfare improving, we can rewrite (15) in the following way:

$$c<1-\frac{B}{p-F_1(2p-1)}=\overline{c}.$$

In order for there to exist a c that satisfies this condition it must be that $\overline{c}>0$ which requires that the following necessary (but not sufficient) condition be satisfied

$$p - F_1(2p - 1) > B. \tag{16}$$

In other words, the bias ratio induced by the review system must sufficiently small in relation to the precision of the review technology in order for the necessary condition to be satisfied. Notice also that $\bar{c}$ is increasing in p implying that for higher values of p, the review system can be welfare improving even for higher values of the cost of invoking a challenge.

Appendix B. Additional descriptive statistics

Table B1 reports the number of matches played in each treated court over time.

Table B1: **Matches played in each treated court over time**

Tournament	Court	2002	2003	2004	2005	2006	2007	2008	2009	2010	Total
US Open	Arthur Ashe	4	11	9	16	10	16	11	17	12	106
US Open	Louis Armstrong	2	5	3	5	7	5	5	5	6	43
Wimbledon	Centre Court	1	9	2	9	10	6	12	10	6	65
Wimbledon	Court 1	0	1	7	6	7	10	8	6	5	50
Australian Op.	Rod Laver	2	2	9	16	12	15	19	13	14	102
Total		9	28	30	52	46	52	55	51	43	366