Università degli Studi di Padova

dSEA

Thomas Bassetti, Department of Economics and Management, University of Padova
Stefano Bonini, Stevens Institute of Technology
Fausto Pacicco, LIUC Università Carlo Cattaneo
Filippo Pavesi, LIUC Università Carlo Cattaneo and Stevens Institute of Technology

Play it again! A Natural Experiment on Reversibility Bias

August 2019

Marco Fanno Working Papers - 238

Play it again! A Natural Experiment on Reversibility Bias

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

Abstract

Behavioral biases affect a large number of human decisions, many of which have relevant welfare effects. We identify a bias that we denote as "reversibility bias" and explore how the introduction of explicit exposure mechanisms can contribute to attenuate it. To do this, we exploit a unique natural experiment - the introduction of a decision review system represented by player challenges and the associated Hawk-Eye technology in professional tennis. This experiment allows us to identify the bias, by illustrating that if such a bias exists, the challenge rule should reduce the number of calls that postpone the assignment of a point. Our findings may have significant policy implications providing a conceptual framework for the design of institutions to alleviate the welfare costs associated with reversibility bias in different contexts, such as court rulings, human resource management and debt roll-over decisions.

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

[*] 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 reversibility bias, which is characterized by agents deriving greater satisfaction from decisions that are non-definitive and can, therefore, be undone. This bias may play an essential role in several applications in economics and finance. For instance, juries and judges may be biased against acquittals because they preclude chances of further trials. Likewise, human resource managers may avoid promoting employees even if their performance suggests the opposite since this decision is less reversible and can be postponed to the future. A similar behavior at the corporate level may involve a bank's decision to roll over debt to avoid making the definitive decision of forcing the borrower into insolvency, even when expectations on future profitability would suggest otherwise.

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

In this paper, we show that, whenever decision making is characterized by reversibility bias, introducing a review system that allows for the ex-post verification of the correctness of the evaluation, attenuates (or eliminates) the effect of the bias leading to tangible welfare gains. So, for example, a randomized 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.[1]

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.

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

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, reversibility bias 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 choices that lead to a payoff relevant ruling that is permanent and cannot be undone.[2] This bias is consistent with Gilbert and Ebert (2002) that show that individuals expect more satisfaction from reversible decisions than they do from irreversible ones.[3]

Our analysis is also related to the literature that exploits sports markets as an ideal setting to identify behavioral biases (Garicano et al., 2005; Romer, 2006; Bar-Eli et al., 2007; Massey and Thaler, 2013; Pope and Schweitzer, 2011). In this respect, our 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. Our focus is reminiscent of the impact bias identified in the sports literature (Green and Daniels, 2015; Sacheti et al., 2015; Kovalchik et al., 2017), whereby referees or judges avoid making decisions that may impact on the score. However, we depart from this literature in two important ways. First, we devise an accurate quasi-experimental design that allows us to identify a bias that is consistent with all the features of prospect theory (Kahneman and Tversky (1979)) and, to the best of our knowledge, has not been previously documented in the extant literature. Second, we provide support for the conjecture that a review system, that allows any of the interested parties to call on a third party to review the decision, can significantly alleviate the effect of the bias on the final result thus generating potentially significant welfare improvements.

[2] While explaining the ultimate psychological motivations for this bias is outside the scope of our analysis, 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))

[3] It is worth noting that 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.

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 allows us to state that the introduction of a review system will have no effect on the expected share of definitive choices of decision-makers in the absence of a bias and a strictly positive impact on this number if decision-makers are characterized by a reversibility bias.

As a second step, we then test the model's predictions by exploiting a natural experiment 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.[4] 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). The idea is that when calls are close, referees that suffer from reversibility bias 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, the introduction of the decision review system should lead some of these calls to be overturned. 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 reversibility bias, and that decision review systems, combining a challenge rule with a third party review such as that adopted in professional tennis, can attenuate the impact of the bias.

As part of our identification strategy, as a unit of analysis, we focus on what is commonly denoted as "ace" in the tennis lingo, namely a serve that is not returned by the receiver and (unless called out by a judge) attributes a point to the server.

This choice is motivated by two distinctive features of "aces" that allow us to clearly

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

identify the existence of a bias. To gather an intuition for this motivation consider that 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 in calling the ball "out" of the service box, this will result in an ace and a point will be assigned to the server. Therefore, the first valuable feature of considering aces is that when a judge chooses to intervene s/he avoids assigning the point which would otherwise determine a permanent effect in the absence of a review system. A second valuable feature of using aces is that a judge's decision on a first serve is not subject to strategic behavior on the part of players when choosing whether to challenge points. It is not unusual in fact, that when confronted with a ball that bounces close to a line, players continue to play and - conditional on the outcome of play - they decide whether to challenge a previous dubious bounce. This would clearly represent a possible confounder in the identification of judges behavior and the existence of a possible bias. Yet, this source of strategic behavior is absent when considering aces, because, by definition, an ace occurs when the receiver is unable to return the ball.

Our results show that following the introduction of a review system, tennis players experienced a significant increase in their ace ratios (i.e., the total number of aces over the total number of served points) even after controlling for a number of covariates and several potential sources of endogeneity. Such a sharp and sudden increase is indicative of an increase in irreversible decisions that is consistent with the empirical implication of our model.

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 discusses the empirical methodology; Section 6 discusses the results and Section 7 concludes.

2. Theoretical Framework

We develop a model in which a decision maker must make a sequence of binary choices, each of which is in favor of one of two agents that have conflicting interests. 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 decision, is less reversible meaning that it involves making a judgment that 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 each agent to challenge the decision-maker's ruling during the course of the interaction between the agents. A challenged decision is then reviewed by an impartial third party and possibly overturned if found to be incorrect. To simplify the exposition and without loss of generality, in the model, we assume that the review technology is perfect (i.e., the correct state can always be verified) and that there is no cost of invoking the review system. Relaxing these two assumptions would not affect the qualitative nature of our results, but would simply reduce the magnitude of the positive effect of the review system on the expected decision.[5]

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 judge and players do not observe the state, but each agent observes an independent signal, $s \in [0, 1]$ that is imperfectly informative on the true state $\omega \in \{0, 1\}$, before making a binary decision $d \in \{0, 1\}$. Here $d = 0$ and $d = 1$ respectively represent the reversible and the irreversible 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

[5] If invoking the review system involves a direct cost or an opportunity cost for the players (as is the case in professional tennis), this reduces the incentives to challenge a decision and produces a milder effect on reversibility. Indeed, irreversible decisions would be turned into possibly reversible, rather than into completely reversible decisions as occurs when exerting the challenge option is costless. It follows that the effect of the review system on the reversibility of decisions from the perspective of the judge is attenuated.

precision. To simplify exposition, we assume that the signals of the judge and the players 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 judge is assumed to have a fair prior, so that $\Pr(\omega = 1) = 1/2$.[6] 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 $b \in \{RB, NB\}$ where RB denotes reversibility bias 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 utility of the judge 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 judge 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 the reversibility bias (RB) may be derived from the following standard properties of prospect theory (Kahneman and Tversky, 1979): 1) utility

[6] 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.

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 is not the complete set of decisions made by the judge over a longer time frame (or the course of her/his career), which is consistent with property 1.[7] 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 assigning a point $(d = 1)$ is a definitive mistake, because it irreversibly assigns a reward to one player. 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 decision is reversible 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 judge's value function in the different states in the presence of reversibility bias and in the absence of a review system:

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

[7] 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 reversibility bias should be more pronounced if we consider only these salient points.

When a review system is introduced, it will make the definitiveness of the decision less relevant, making correctness the salient attribute. More specifically, when an incorrect decision is made, since it becomes reversible the disutility of the choice no longer depends on which was made. On the other hand, the greater public exposure provided by the review system makes a correct decision equally valuable even if one is less definitive than the other. It therefore follows that:

$$0 = v(1,1,b,H) - v(0,0,b,H) = v(0,1,b,H) - v(1,0,b,H). \tag{2}$$

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 judges and the informativeness of signals are such that there always exists a threshold $s^* \in (0,1)$, for which a judge will always set $d = 1(d = 0)$ for $s > s^*$ $(s < s^*)$. This threshold value is defined by the value of s for which the judge 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 $s^*_{b,r}$ as the threshold value of s^* for bias b and regime r. Given relations (1) and (2), it follows that $s^*_{b,H} = s^*_{NB,r} < s^*_{RB,NH}$. Notice that the symmetric signal structure implies that $(1 - F_1(s^*_{NB,r})) = F_0(s^*_{NB,r})$, so that in the absence of a reporting bias the probability of providing a correct evaluation is equal in both states.

In terms of player behavior, based on the signal observed, a given player 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, every decision is challenged when one player disagrees with the judge, meaning that the judge's decision does not correspond to the state the player believes is more likely to be true. Moreover, for strategic reasons, the decision may be challenged with positive probability even when a player's private information does not contradict that of the judge. Whenever a judgment is challenged, given the assumption that the review technology is perfect, the decision is overturned if it does not match the true state.

We denote the judge's decision conditional on the bias, the regime and the signal observed with $d(b, r, s) \in [0, 1]$. The expected decision (i.e., before observing the signal) conditional on the bias and the regime can therefore we written in the following way:

$$\begin{aligned} E[d(b,r)] &= \Pr(\omega = 1)\left[1 \Pr(s > s^* \mid \omega = 1, b, r) + 0 \Pr(s < s^* \mid \omega = 1, b, r)\right] + \\ \Pr(\omega &= 0)\left[1(\Pr(s > s^* \mid \omega = 0, b, r) + 0 \Pr(s < s^* \mid \omega = 0, b, r)\right]. \end{aligned}$$

From this expression we define $\Delta d(b) \equiv E[d(b, H)] - E[d(b, NH)]$ that represents the difference between the expected decision before and after the introduction of the review system when the bias is b, which is equal to[8]:

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

Once again considering that $s^*_{b,H} = s^*_{NB,r} < s^*_{RB,NH}$, it is straightforward to observe that $\Delta d(RB) > 0$ and $\Delta d(NB) = 0$, which leads to the following conclusion:

$$E[d(NB, NH)] = E[d(b, H)] > E[d(RB, NH)].$$

This result allows us to clearly state our empirical prediction:

Prediction: If decision makers exhibit a reversibility bias (RB) the introduction of the review system leads to an increase in the share of definitive decisions, while there is no effect on this share in the absence of a bias (NB) .

[8] 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.[9] 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.[10]

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

[9] 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.

[10] 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 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.[11] In this respect, players may engage in strategic behavior when choosing whether to challenge points. It is not unusual, in fact, that when confronted with a ball that bounces close to a line, players continue to play and - conditional on the outcome of play - they decide whether to challenge a previous dubious bounce. This would clearly represent a possible confounder in the identification of possible biases in judges behavior. 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. 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

[11] 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.

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 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, as mentioned above, 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[12]. This dataset contains detailed statistics and results on most of the professional tennis matches 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 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

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

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.[13] 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

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

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

We present descriptive statistics in Table 2.

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.

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).

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:

$$Y_{pct} = \alpha_c + \alpha_t + \beta \cdot H_{ct} + \gamma \cdot X_{pct} + \delta_p + e_{pct}, \tag{3}$$

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.

One can argue that, after the introduction of the Hawk-Eye, players might 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. First, we estimate a triple difference model in which clay courts constitute a placebo control group. Our specification can be modified as follows:

$$Y_{pct} = \alpha_c + \alpha_t + \beta_1 \cdot H_{ct} + \beta_2 \cdot C_{ct} + \gamma \cdot X_{pct} + \delta_p + e_{pct}, \tag{4}$$

where C_{ct} indicates clay courts after the treatment period. In the presence of a reversibility bias, we expect the estimate of β_2 to be statistically insignificant. 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. This implies that, 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 (3) as follows:

$$\begin{aligned} Y_{pct} = \alpha_c + \alpha_t + \beta_1 \cdot H_{ct} + \beta_2 \cdot Break \cdot CE_p + \beta_3 \cdot Treated \\ \cdot CE_p + \beta_4 \cdot H_{ct} \cdot CE_p + \gamma \cdot X_{pct} + \delta_p + e_{pct}, \end{aligned} \tag{5}$$

where CE_p denotes players' 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 Haw 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). Therefore, 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:

$$p(x_i) = Pr(H = 1 \Big| X_i) = \frac{1}{1 + e^{-\delta \cdot X_i}}, \tag{6}$$

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

$$\tau_{ATT} = \sum_{i=1}^{N} \cdot H_i \cdot Y_i - \sum_{i=1}^{N} \frac{p(x_i)}{1 - p(x_i)} \cdot (1 - H_i) \cdot Y_i. \tag{7}$$

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, whereas the outcome model would suffer from the omission of pair-specific effects, 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 (3) and (4) taking into account pairs-tournament specific effects.

6. Results

6.1. Main results

Table 4 shows the estimates of Equation (3) 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 years, the average treatment effect on the treated was about 1.35-1.5%, 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 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. The challenger's ranking exhibits the only additional significant within-estimate besides the main independent variable.

Table 5 reports the estimates of Equation (4) where we add a third group represented by matches played on a clay court (the French Open) as a placebo treatment.

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.

While in the additional robustness section we re-estimate Equations (3) and (4) taking into account tournament-players 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 (5) for different time breaks.

INSERT TABLE 6 HERE

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.[14] 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.

[14] 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 $Treated \cdot CE$

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 6, 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 3 and 4, 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 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 (3) and (4) taking into account pairs-tournament specific effects. Formally, we estimate the following models:

$$Y_{pct} = \alpha_c + \alpha_t + \beta \cdot H_{ct} + \gamma \cdot X_{pct} + \delta_{ps} + e_{pct}, \tag{8}$$

and

$$Y_{pct} = \alpha_c + \alpha_t + \beta_1 \cdot H_{ct} + \beta_2 \cdot Cl_{ct} + \gamma \cdot X_{pct} + \delta_{ps} + e_{pct}, \tag{9}$$

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. Table 8 provides the coefficients of Equation (8), where we control for possible interactions between pairs of players and tournaments unobserved characteristics.

INSERT TABLE 8 HERE

Results are consistent with those presented in Table 3 and the treatment effect is even larger.

As a final robustness test, in Table 9 we estimate Equation (9), jointly controlling for possible interactions between pairs of players and tournaments unobserved characteristics and placebo court type

INSERT TABLE 9 HERE

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.

7. Discussion and conclusions

Our analysis suggests that decision makers may be subject to reversibility 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 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 reversibility is salient, welfare improvements can 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.

While our analysis allows us to state that the introduction of a review system is welfare improving, two main issues remain to be explored. The first involves identifying neutral third party reviewers in the real world that have the same desirable features of the Hawk-Eye system, namely competence and neutrality. The second issue involves understanding whether the introduction of the review system simply neutralizes the effect of the bias that continues to characterize behavior, or whether 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.

Regarding the first issue, it is often the case that identifying a competent and unbiased third party may be particularly challenging. 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.[15] Considering the second issue, as suggested by Tetlock and Gardner (2015) inducing experts to focus on the correctness of their decisions, may actually improve upon their ability to make better decisions. Although this is beyond the scope of our analysis, Tables 7 and 9 suggest that our empirical evidence is not inconsistent with this being the case. Notice indeed, that as shown in the last columns of Tables 7 and 9 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.,

[15] Along these lines, Chen (2019) suggests how to create decision support systems for judges that combine large datasets with artificial intelligence in order to correct for behavioral biases.

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.

References

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.

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].

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

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.

Green, E. and Daniels, D. P. (2015). Impact aversion in arbitrator decisions. *Available at SSRN 2391558*.

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.

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.

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.

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.

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

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

	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

Table 2

Descriptive Statistics

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

	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

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

This table presents results of a set of the DD model specified in Equation 1. The dependent variable is the aces to points ratio measured for pair p, in court c at time t, α_c and α_t are courts and time dummies absorbing for 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, δ_p are pairs fixed effects, and e_{pct} is the error term. We take into account pairs time series dependence using clustered-robust standard errors. Significance levels: *10%, **5%, ***1%

	2006-2010	2007-2010	2008-2010	2009-2010	2010
	(1)	(2)	(3)	(4)	(5)
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)
Pair FE	Yes	Yes	Yes	Yes	Yes
Court FE	Yes	Yes	Yes	Yes	Yes
Time FE	Yes	Yes	Yes	Yes	Yes
N	639	639	639	639	639
adj. R2	0.540	0.541	0.543	0.541	0.537
within R2	0.158	0.160	0.163	0.161	0.153

Table 5

Hawk Eye effect on Ace Ratio: placebo tests

This table presents results of a set of the triple DDD model specified in Equation 2 introducing a placebo category expresented by Clay courts matches. The dependent variable is the aces to points ratio measured for pair p, in court c at time t, α_c and α_t are courts and time dummies absorbing for 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, δ_p are pairs fixed effects, and e_{pct} is the error term. We take into account pairs time series dependence using clustered-robust standard errors. Significance levels: *10%, **5%, ***1%

	2006-2010	2007-2010	2008-2010	2009-2010	2010
	(1)	(2)	(3)	(4)	(5)
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)	-
Favorite Rank	0.005	0.005	0.004	0.004	0.005
	(0.010)	(0.010)	(0.010)	(0.010)	(0.010)
Challenger Rank	-0.004**	-0.004**	-0.004**	-0.004**	-0.004**
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
Favorite Age	0.065	0.022	0.023	0.099	0.173*
	(0.207)	(0.208)	(0.206)	(0.094)	(0.095)
Challenger Age	0.100	0.056	0.050	0.133	0.201**
	(0.201)	(0.202)	(0.200)	(0.094)	(0.093)
Home player	-0.293	-0.279	-0.288	-0.290	-0.289
	(0.376)	(0.378)	(0.382)	(0.379)	(0.375)
Minutes	-0.002	-0.002	-0.002	-0.002	-0.002
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
Pair FE	Yes	Yes	Yes	Yes	Yes
Court FE	Yes	Yes	Yes	Yes	Yes
Time FE	Yes	Yes	Yes	Yes	Yes
N	1010	1010	1010	1010	1010
adj. R2	0.576	0.577	0.577	0.575	0.576
within R2	0.155	0.157	0.159	0.153	0.155

Table 6

Hawk Eye effect on Ace Ratio: player experience on placebo court

This table presents results of a set of the DD model specified in Equation 5. 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, α_c and α_t are courts and time dummies absorbing for 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, δ_p are pairs fixed effects, and e_{pct} is the error term. We take into account pairs time series dependence using clustered-robust standard errors. Significance levels: *10%, **5%, ***1%

	2006-2010	2007-2010	2008-2010	2009-2010	2010
	(1)	(2)	(3)	(4)	(5)
Hawk Eye	5.363**	5.738***	3.513**	3.406**	0.362
	(2.361)	(2.035)	(1.645)	(1.719)	(2.110)
$Break \cdot CE$	-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)
$HawkEye \cdot CE$	-18.683*	-18.481**	-7.441	-7.719	-3.995
	(9.506)	(7.719)	(6.067)	(6.297)	(6.930)
Favorite Rank	0.005	0.008	0.003	0.005	0.003
	(0.015)	(0.015)	(0.014)	(0.015)	(0.014)
Challenger Rank	-0.005**	-0.005**	-0.005**	-0.006**	-0.005**
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
Challenger Age	0.040	0.045	0.012	0.034	0.025
	(0.092)	(0.090)	(0.092)	(0.093)	(0.091)
Home player	-0.504	-0.355	-0.323	-0.272	-0.355
	(0.565)	(0.571)	(0.573)	(0.575)	(0.576)
Minutes	0.001	0.001	0.001	0.001	0.001
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
Pair FE	Yes	Yes	Yes	Yes	Yes
Court FE	Yes	Yes	Yes	Yes	Yes
Time FE	Yes	Yes	Yes	Yes	Yes
N	639	639	639	639	639
adj. R2	0.546	0.553	0.541	0.541	0.533
within R2	0.041	0.058	0.032	0.032	0.016

Table 7

Sequential IPWRA estimates

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%

	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

Table 8

Pair-tournament effects

This table reports the estimates of Equation (8), considering different treatment periods. Now, pair-tournament FEs absorb the effect of the home player variable. Robust-clustered standard errors are in parentheses. Significance levels: *10%, **5%, ***1%

	2006-2010	2007-2010	2008-2010	2009-2010	2010
	(1)	(2)	(3)	(4)	(5)
Hawk Eye	1.025	3.249**	3.492**	3.107*	-1.730
	(1.583)	(1.415)	(1.733)	(1.810)	(1.815)
Favorite Rank	0.009	0.013	0.012	0.022	-0.002
	(0.026)	(0.024)	(0.026)	(0.033)	(0.031)
Challenger Rank	0.007	0.008	0.007	0.003	0.008
	(0.011)	(0.011)	(0.011)	(0.011)	(0.011)
Favorite Age	0.680	0.365	0.405	-0.307*	-0.059
	(0.587)	(0.594)	(0.556)	(0.178)	(0.195)
Challenger Age	0.576	0.315	0.261	-0.486**	-0.223
	(0.497)	(0.499)	(0.488)	(0.194)	(0.197)
Minutes	-0.006	-0.005	-0.007	-0.006	-0.006
	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)
Pair-Tournment FE	Yes	Yes	Yes	Yes	Yes
Time FE	Yes	Yes	Yes	Yes	Yes
N	208	208	208	208	208
adj R2	0.503	0.522	0.530	0.524	0.506
within R2	0.385	0.408	0.418	0.410	0.388

Table 9

Pair-tournament and placebo court regressions

This table reports the estimates of Equation (8), jointly controlling for possible interactions between pairs of players and tournaments unobserved characteristics and placebo court type. Pair-tournament FEs absorb the effect of the home player variable. Robust-clustered standard errors are in parentheses. Significance levels: *10%, **5%, ***1%

	2006-2010	2007-2010	2008-2010	2009-2010	2010
	(1)	(2)	(3)	(4)	(5)
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)
Favorite Rank	0.004	0.006	0.007	0.011	0.000
	(0.020)	(0.020)	(0.020)	(0.022)	(0.021)
Challenger Rank	-0.002	-0.002	-0.002	-0.002	-0.001
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
Favorite Age	1.100**	0.835*	0.744	1.019**	0.041
	(0.467)	(0.466)	(0.451)	(0.419)	(0.181)
Challenger Age	0.747*	0.511	0.377	0.635	-0.337**
	(0.432)	(0.433)	(0.431)	(0.396)	(0.170)
Minutes	-0.006	-0.006	-0.007	-0.006	-0.006
	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
Pair-Tournment FE	Yes	Yes	Yes	Yes	Yes
Time FE	Yes	Yes	Yes	Yes	Yes
N	293	293	293	293	293
adj R2	0.573	0.581	0.589	0.587	0.576
within R2	0.355	0.368	0.379	0.376	0.359

Fig. 1. Reversibility Bias and Prospect Theory

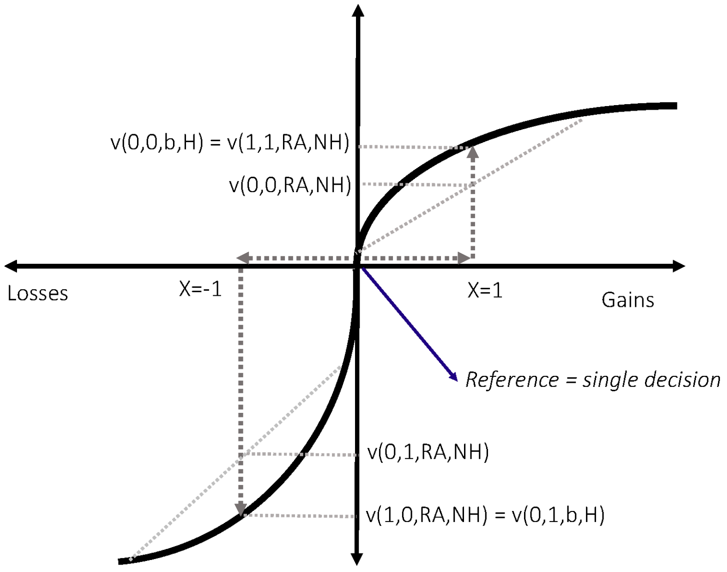
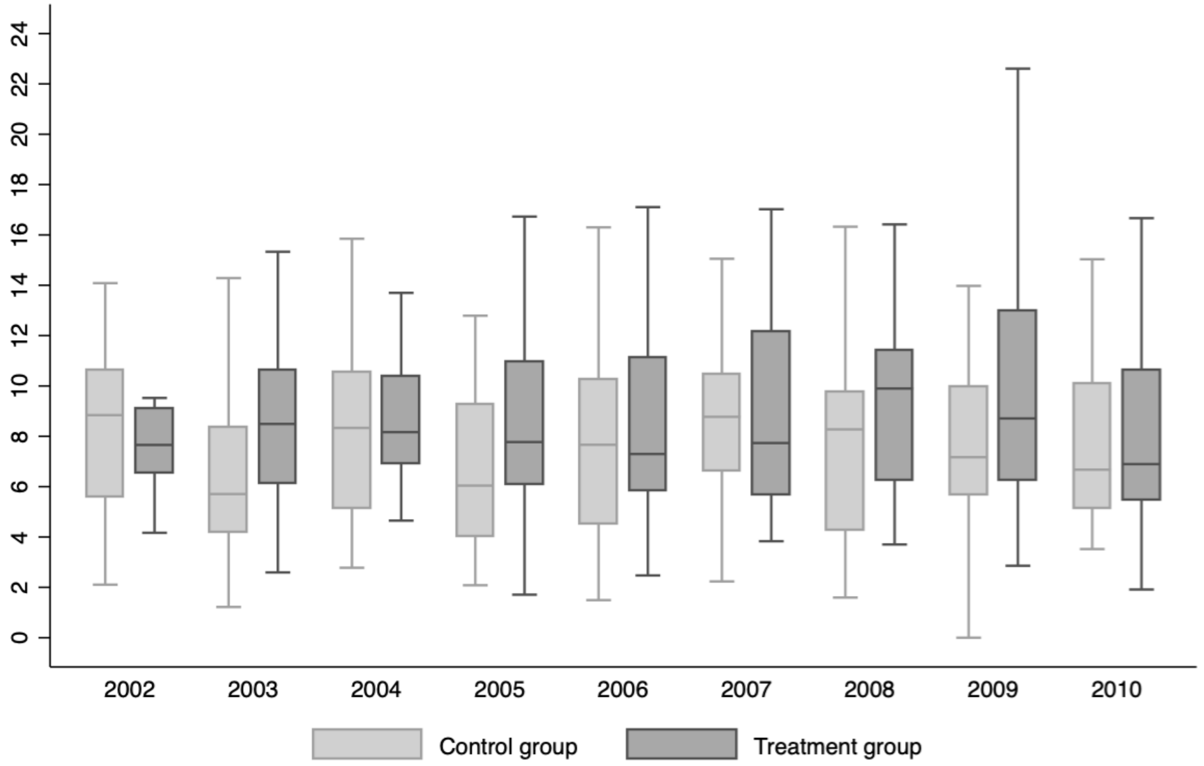

v(0,0,b,H) = v(1,1,RA,NH)
v(0,0,RA,NH)
Losses
X=-1
X=1
Gains
Reference = single decision
v(0,1,RA,NH)
v(1,0,RA,NH) = v(0,1,b,H)

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. Derivation of $\Delta d(b)$

Let us denote with $\varepsilon \in [0,1)$, the probability that a judge's decision is challenged when an agent's private information does not contradict that of the judge. As mentioned previously, this may occur with positive probability for strategic reasons. Given that the bias is specific to those that are in a position to judge, we naturally assume that the agents (or players) are not subject to reversibility bias.

Therefore, the expressions for the expected decision with and without the review system are the following:

$$E[d(b,H)] = 1/2[(1 - F_1(s^*_{b,H})) + F_1(s^*_{b,H})(1 - F_1(s^*_{NB,r}) + F_1(s^*_{NB,r})\varepsilon)$$
$$+ (1 - F_0(s^*_{b,H}))(1 - F_0(s^*_{NB,r})(1-\varepsilon))].$$

and

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

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

$$\Delta d(b) = 1/2[\sum_{\omega}((F_\omega(s^*_{b,NH}) - F_\omega(s^*_{b,H})) + F_1(s^*_{b,H})(1 - F_1(s^*_{NB,r})(1-\varepsilon))$$
$$- (1 - F_0(s^*_{b,H}))(F_0(s^*_{NB,r})(1-\varepsilon))].$$

By the symmetry of the signal structure $(1 - F_1(s^*_{NB,r}) = F_0(s^*_{NB,r})$, and by expressions (1) and (2) $s^*_{b,H} = s^*_{NB,r}$. Thus, it follows that the second and third term in the square brackets cancel out, which completes the proof.

Appendix B. Additional descriptive statistics

Table A1 reports the number of matches played in each treated court over time.

Table B1

Matches played in each treated court over time

This table reports the number of matches played in each treated court over the observational period

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