The University Of Sheffield.

Department Of Economics

Learning in Rent-Seeking Contests with Payoff Risk and Foregone Payoff Information

Aidas Masiliūnas

Sheffield Economic Research Paper Series

SERPS no. 2023002

ISSN 1749-8368

13 Jan 2023

Learning in Rent-Seeking Contests with Payoff Risk and Foregone Payoff Information*

Aidas Masiliūnas†

January 13, 2023

Abstract

We test whether deviations from Nash equilibrium in rent-seeking contests can be explained by the slow convergence of payoff-based learning. We identify and eliminate two noise sources that slow down learning: first, opponents are changing their actions across rounds; second, payoffs are probabilistic, which reduces the correlation between expected and realized payoffs. We find that average choices are not significantly different from the risk-neutral Nash equilibrium predictions only when both noise sources are eliminated by supplying foregone payoff information and removing payoff risk. Payoff-based learning can explain these results better than alternative theories. We propose a hybrid learning model that combines reinforcement and belief learning with risk, social and other preferences, and show that it fits data well, mostly because of reinforcement learning.

Keywords: experiment, contests, reinforcement learning, foregone payoffs, payoff risk, Nash equilibrium

JEL classification: C72 C91 D71 D81

*I would like to thank Friederike Mengel, Heinrich H. Nax, Philipp J. Reiss, Theodore Turocy, Francesco Fallucchi and participants at seminars and conferences in Maastricht (BEElab meeting), Vienna (ESA European meeting 2017), Gothenburg (12th Nordic Conference on Behavioral and Experimental Economics), Marseille (Eco-lunch seminar), Bath (The Micro and Macro Foundations of Conflict), Brno (Young Economists' Meeting), Nice (ASFEE 2018), Vilnius (Baltic Economic Conference), Berlin (ESA World Meeting) and Singapore (Game Theory and Social Dynamics Workshop at SUTD) for comments and suggestions. Financial support from GSBE at Maastricht University, Aix-Marseille School of Economics and Global Asia Institute at the National University of Singapore is gratefully acknowledged.

†Department of Economics, University of Sheffield, 9 Mappin Street, Sheffield S1 4DT, UK. E-mail: aidas.masiliunas@gmail.com

1 Introduction

Behavior in experiments often deviates from the risk-neutral Nash equilibrium.[1] The discrepancy is commonly attributed to the restrictive Nash equilibrium assumptions, such as selfishness, risk neutrality or unlimited cognitive abilities. Consequently, it has been proposed that the discrepancy could be remedied by enriching Nash equilibrium with social, risk and other preferences (DellaVigna, 2009), or decision error (Goeree and Holt, 2001). Instead, we investigate an alternative explanation that deviations from Nash equilibrium arise because behavior is observed in relatively short experiments in an environment not conducive to learning. Convergence is especially problematic for payoff-based learning models, which in the long run should converge to Nash equilibrium (Hopkins, 2002), but might be slow if little is learnt from the realized payoff information due to payoff noise (Thrun and Schwartz, 1993, Hasselt, 2010). Payoff noise can originate either from a stochastic payoff function or from the inter-temporal variability of opponents' choices. We study a rent-seeking contest (Tullock, 1980) in which both noise sources are present. We eliminate these noise sources by removing the probabilistic prize allocation rule and providing information about the ex-post payoffs potentially earned if the player had made a different choice. We test whether these manipulations alter the adaptation process and the explanatory power of the risk-neutral Nash equilibrium.

In decision-making under risk, payoffs are stochastic and may slow down convergence if players learn from their ex-post payoffs. Game theory offers little explanation for how payoff risk affects the explanatory power of Nash equilibrium in repeated games,[2] perhaps because decisions in most games are made only under strategic uncertainty and not risk. Most experimental work on risk has been done using individual choice tasks, both one-shot (e.g., Kahneman and Tversky, 1979) and repeated (often using the "decisions-from-experience" paradigm[3]). In repeated games, it is typically found that payoff risk reduces the payoff maximization rates (Myers and Sadler, 1960, Erev and Barron, 2005). Given these findings, the low explanatory power of Nash equilibrium in rent-seeking contests might not be surprising, as contests are run with a larger strategy space, lower optimization incentives, a smaller number of repetitions and payoffs depend on the choices of other participants.[4]

We study the role of risk in rent-seeking contests by varying the sources of probabilistic payoffs. First, we compare the rent-seeking contest in which prize allocation is probabilistic (SR treatment) to a treatment in which the payoff risk is eliminated by paying the expected

[1] Here and throughout the paper, we refer to the concept of Nash equilibrium under the assumption of selfish expected payoff maximization.

[2] Risk can lead to deviations from the risk-neutral Nash equilibrium predictions if players are risk seeking or risk averse. However, risk plays an additional role in repeated games, where stochastic payoffs might slow down learning. In this paper, we focus on the latter role, which is largely unexplored. Two notable exceptions that studied the effect of risk on convergence are Bereby-Meyer and Roth (2006) and Shafran (2012).

[3] The "decisions-from-experience" paradigm is used mainly in psychology experiments, in which players choose between several options that generate payoffs from an unknown distribution. Typically no information about the nature of the payoff distribution is provided, to test how choices are made only from experience, in contrast to "decisions-from-description" (for an overview, see Erev and Haruvy, 2016).

[4] For example, in one task studied by Grosskopf et al. (2006), participants could select a distribution from which to draw the payoffs ($\mathcal{N}(11,1)$ or $\mathcal{N}(10,3)$, unknown to participants) and the option that generates 9% higher expected payoffs was chosen only about half of the time, even after 200 trials. All contest experiments have a much larger strategy space, none are repeated more than 60 times (Fallucchi et al., 2013) and 7% of the expected payoffs are lost from overbidding the Nash equilibrium level by 100% (the median overbidding rate in contest experiments reviewed by Sheremeta, 2013, is 72%).

value of the lottery (SS treatment, sometimes called a "share contest"). However, convergence to the risk-neutral Nash equilibrium in this treatment could occur for reasons other than payoff-based learning, as removing payoff risk would eliminate the effect of risk preferences, probability weighting or non-monetary utility of winning. To better understand how payoff risk operates, we designed two additional treatments. In the reverse contest (RS treatment), payoff risk is reversed by making contest investment safe, but the amount not invested in the contest risky. In the RR treatment, both types of investment are risky. By comparing behavior in these four treatments, we can understand why risk affects the explanatory power of the risk-neutral Nash equilibrium. If risk matters because stochastic payoffs slow down payoff-based learning, convergence rates should be similarly low in all three treatments in which payoffs are stochastic (SR, RS and RR). In contrast, if risk makes contest investment more attractive because winning provides additional non-monetary utility, players are loss averse or overweight the probability of winning, then we should expect little deviation from the equilibrium predictions in SS and RR, and below-equilibrium contest investment in RS. Finally, if the mechanism that drives behavior does not critically depend on risk (for example, if participants are spiteful or inequality averse), behavioral patterns should be similar in all four treatments.

The second obstacle to convergence is brought by the inter-temporal variability of opponents' choices. Payoff-based learning converges through the iterated elimination of dominated strategies (Beggs, 2005), but the dominance might not be reflected in realized profits if opponents change their behavior across rounds. Therefore, detecting dominance from information about realized payoffs requires many trials. To facilitate learning, we provide foregone payoff information (FPI), informing participants what their ex-post payoffs would have been if they had made a different choice. FPI allows participants to compare the performance of all actions against the same sequence of opponent's actions, improving both the quantity and the quality of information. If there is no payoff risk and participants can learn from FPI, payoff-based learning predicts fast convergence; otherwise, convergence is slow (see Appendix A).

Our experiments support the payoff-based learning predictions. Study 1 finds that when FPI is introduced, convergence to the Nash equilibrium is significantly higher in the treatment without payoff risk, compared to the other three treatments. Study 2 replicates the main result from Study 1 in an environment where learning is more challenging. It also shows that FPI is necessary to observe convergence, as contest investment in treatments without FPI remains significantly above the equilibrium prediction, even in the absence of payoff risk. To explain these results, we estimate a learning model that combines reinforcement and belief learning with probability weighting, non-monetary utility of winning, risk and social preferences. Reinforcement learning can explain high equilibrium rates in the treatment with no payoff risk following the introduction of FPI. Non-monetary utility of winning and probability weighting can explain differences between treatments with payoff risk. Consequently, the full model that combines reinforcement learning and belief learning with preferences fits better than models with only one of these elements.

Our study contributes to the understanding of overbidding in rent-seeking contests, as well as the deviations from theoretical predictions in laboratory experiments more generally. Explanations for why behavior differs from theoretical predictions are often divided into non-standard preferences,[5] incorrect beliefs or non-standard decision making processes

[5] "Non-standard preferences" refer to a utility function that differs from the expected payoff function, for example, because of risk or social preferences (DellaVigna, 2009).

(DellaVigna, 2009). The previous rent-seeking contest literature studied non-standard preferences, non-standard beliefs and decision making from description, but we are the first to focus on decisions from experience and study the obstacles to convergence. We find that information about foregone payoffs significantly increases convergence rates, but only in the absence of payoff risk, as predicted by a payoff-based learning model. We also find that payoff-based learning can organize experimental data better than explanations based on non-standard preferences. To disentangle the potential explanations, we develop a novel method of jointly modeling bounded rationality and non-standard preferences. This approach could be readily adapted to understand the mechanism that drives behavior in other types of games.

2 Literature

Rent-seeking contests in which the success probability is proportional to investment have been widely studied in laboratory experiments (Sheremeta, 2013). Most experiments find that average choices exceed the Nash equilibrium prediction ("over-investment", "over-expenditure", "overspending" or "overbidding") and the distribution of choices is widely spread ("over-spreading", Chowdhury et al., 2014), therefore Nash equilibrium fails to organize experimental data (Sheremeta and Zhang, 2010, Masiliūnas et al., 2014). The gap between behavior and theoretical predictions is often attributed either to the failure to take into account risk preferences (Jindapon and Whaley, 2015), other-regarding preferences (Herrmann and Orzen, 2008), non-monetary utility from winning (Sheremeta, 2010) or non-linear probability weighting (Baharad and Nitzan, 2008). Preference-based explanations are typically assessed by eliciting preferences in a separate task and testing if they are correlated with contest investment (Sheremeta, 2018) or by evaluating whether over-investment disappears when the effect of certain preferences is eliminated (for a detailed overview, see Appendix B). It is typically found that over-investment persists even in contests where risk or social preferences should play no role, such as when there is no payoff risk or when participants are matched with robot opponents (Fallucchi et al., 2013; Cox, 2017). Over-investment also persists when incorrect beliefs about the behavior of other participants are corrected by playing the contest sequentially (Fonseca, 2009) or by informing participants about the action that will be played by a robot opponent (Masiliūnas et al., 2014). An alternative explanation is that participants find it difficult to find the payoff-maximizing action. Some evidence for this hypothesis comes from an increased explanatory power of Nash equilibrium when the game is simplified (Masiliūnas et al., 2014) or framed as a ticket-based lottery (Chowdhury et al., 2020). Our study deviates from the literature by focusing on the information that alters decisions from experience, rather than decisions from the description of the game.

Rent-seeking contests with no payoff risk ("share contests", equivalent to our SS treatment) have been studied by Fallucchi et al. (2013), Chowdhury et al. (2014) and Masiliūnas et al. (2014).[6] All three studies find that the removal of payoff risk does not reduce the over-investment rate, although the magnitude of over-investment and under-investment tends to be reduced. In contrast, the removal of payoff risk reduces over-investment and increases the explanatory power of Nash equilibrium if the optimization premium is high (Chowdhury

[6]Other studies look at share contests with a slightly different design: in Shupp et al. (2013) the strategy space is censored, which seems to be the reason why average investment is below the Nash equilibrium, in Cason et al. (2020) the cost of effort is quadratic and investment is subject to additional noise. Additional noise may also originate from uncertainty about external evaluation (Chung et al., 2020).

et al., 2014), if information about individual choices of other participants is withheld (Fallucchi et al., 2013) or if the opponent is required to play the same action for some time (Masiliūnas et al., 2014).

We are not aware of any literature that manipulates the availability of FPI in a strategic game,[7] but some experiments study other types of information and feedback. One relevant comparison is between partial information treatments, where participants know only own payoffs, and full information treatments, where participants also know the choices and payoffs of others. Full information facilitates convergence in market entry games (Duffy and Hopkins, 2005) and in common value auctions (Armantier, 2004), but has an opposite effect in Cournot oligopoly, where it instead facilitates imitate-the-best dynamics and convergence to the Walrasian equilibrium (Offerman et al., 2002). The effects of feedback have also been studied in first-price auctions, where more overbidding is observed when the loser is informed about the winning bid (Filiz-Ozbay and Ozbay, 2007; Engelbrecht-Wiggans and Katok, 2009).[8] Our FPI manipulation is related to this literature, as we test whether additional information facilitates convergence to the Nash equilibrium. However, participants in all of our treatments have full information about the actions of others. Also related to our study are the experiments that study the format of information, hypothesizing that an improved understanding of the incentive structure should reduce the magnitude of deviations from theoretical predictions. Support for this hypothesis has been found in gift exchange games (Charness et al., 2004) and Cournot oligopoly (Bosch-Domènech and Vriend, 2003). This literature is related to our FPI manipulation, since both study whether information can improve the decisions of boundedly rational participants. However, the approaches differ because we aim to study whether ex-post feedback improves decisions from experience, while the previous papers focused on information that could improve decisions from description.

FPI has been studied in individual choice tasks, typically using the "decisions-from-experience" paradigm. Rakow et al. (2015) find that the alternative with a higher expected value is chosen more often when FPI is available. Grosskopf et al. (2007) and Fudenberg and Peysakhovich (2016) find that sub-optimal overbidding in the Acquire a Company game and Additive Lemons Problem does not reduce when FPI is provided. Grosskopf et al. (2006) show that FPI increases expected payoff maximization if payoffs from both actions are positively correlated, but reduces it in other cases, a result attributed to over-exploration and increased focus on large positive payoffs generated by the more risky option ("big eyes effect"). Otto and Love (2010) find that in a dynamic game FPI reduces the choice of an option that maximizes long-run payoffs, in favor of an option that provides a high immediate payoff. This result can be explained by reinforcement learning, because FPI increases the attractiveness of the option with high immediate payoffs and the strategy space remains under-explored. Yechiam and Busemeyer (2006) investigate a task in which the more risky action usually generates a higher payoff, but has a lower expected payoff because of a small probability of a large loss. The risky option is chosen more often when FPI is available, and the treatment difference is larger when the probability to receive the large negative reward is small (1/200 compared to 1/20). These results can be well explained by reinforcement learning. Overall, this literature shows

[7] To the best of our knowledge, the only study that manipulated FPI in a strategic game is Grosskopf et al. (2007), who do so in the Acquire a Company task. The game is strategic only in control treatments, in which the role of the seller is assumed by human participants, but their decisions are trivial.

[8] These studies are different from our design not only because of a different game, but also because they do not study repeated strategic games: Filiz-Ozbay and Ozbay (2007) use a one-shot design while participants in Engelbrecht-Wiggans and Katok (2009) compete against robots.

that the effect of FPI depends on the task and can be explained by reinforcement learning. We contribute to this literature by testing the reinforcement learning predictions about the effect of FPI on Nash equilibrium play rates in games with varying sources of payoff risk.

Also related to our study are continuous-time experiments in which participants receive the average instantaneous payoff against the choices of all other participants ("mean-matching"). Behavior in these experiments closely approximates Nash equilibrium predictions in hawk-dove game (Oprea et al., 2011), rock-paper-scissors game (Cason et al., 2014), coordinated attacker-defender games (Stephenson, 2019), Hotelling games (Kephart and Friedman, 2015) and all-pay auctions (Stephenson and Brown, 2021). The latter study is particularly relevant for our paper as it fails to find the commonly observed overbidding in all-pay auctions. Our study is different as it uses a discrete period design and provides additional insights; for example, we separate the effect of risk from other factors (in continuous-time games, it is confounded with the stability of opponent's actions) and we compare the role of risk in decisions from description and in decisions from experience. Since we use a factorial design, we can identify whether the improved convergence is driven by better information or the lack of payoff risk, while the previous studies differed in multiple design elements (either the standard discrete time framework, or continuous-time experiments with mean-matching and enhanced feedback).

3 Experimental design

In the regular rent-seeking contest, the part of endowment invested in the contest is risky, while the remaining part is safe and directly converted into earnings. We designed three additional contest variations in which players also chose how to divide the endowment, but the riskiness of contest and non-contest investment was manipulated using a 2x2 between-subject design. Treatment differences are displayed in Figure 1.

- **Treatment SR** (*safe* non-contest investment, *risky* contest investment) is equivalent to the regular rent-seeking contest.
- **Treatment SS** (*safe* non-contest investment, *safe* contest investment) is equivalent to the share contest (Fallucchi et al., 2013). The probabilistic prize allocation rule, used in SR, is replaced by the expected value of the lottery, removing payoff risk.
- **Treatment RS** (*risky* non-contest investment, *safe* contest investment, or "reverse") makes income from non-contest investment risky and income from contest investment safe, reversing the payoff risk compared to the SR treatment. In RS, contest investment generates deterministic payoffs, just as in SS, but non-contest investment determines the probability to win a prize. The prize value is the same as in SR, but the odds to win are proportional to non-contest investment.
- **Treatment RR** (*risky* non-contest investment, *risky* contest investment) combines two lotteries with identical prize values: in one the probability to win is determined by contest investment, in the other by non-contest investment.

The payoffs in each treatment depend on the choices and on the lottery outcomes. Each player $i \in \{1, 2\}$ divides the endowment E between contest investment ($c_i \in C$) and non-contest investment ($E - c_i$). In the experiment, participants were informed about E and

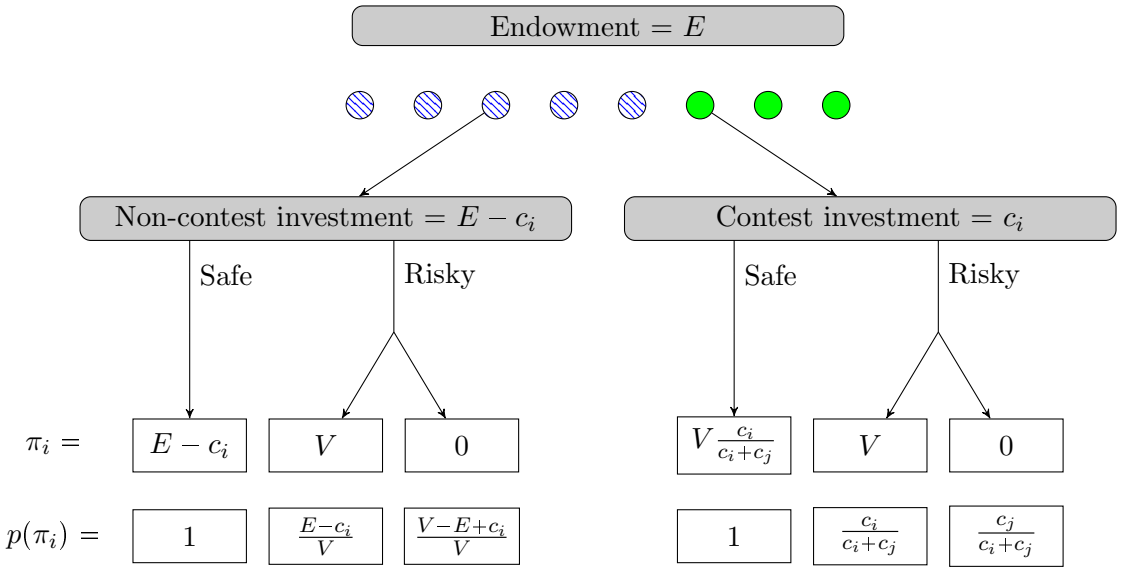


Figure 1: Sources of payoff risk in four contest treatments.

chose what part of it to invest in contest (framed as the purchase of "tokens" in SS and RS, and "tickets" in SR and RR). We will therefore refer to c_i as the choice variable. Depending on the treatment, investment determines either the share of the prize or the probability to receive the prize. Denote the value of the contest and non-contest prize by V, the probability/share of the contest prize by p_i^c and the probability/share of the non-contest prize by p_i^{nc} such that:

$$p_i^c(c_i, c_j) = \begin{cases} \frac{c_i}{c_i+c_j} & \text{if } c_i + c_j > 0 \\ 0.5 & \text{otherwise} \end{cases} \tag{1}$$

$$p_i^{nc}(c_i) = \frac{E - c_i}{V} \tag{2}$$

Payoffs also depend on the realization of up to two lotteries, depending on the treatment. We model the lottery draw that determines the allocation of the contest prize using a random variable $r_i^c \sim \mathcal{U}(0,1)$.[9] In SR and RR, player i receives the contest prize if $r_i^c \leqslant p_i^c(c_i, c_j)$. In SS and RS, the lottery outcome is irrelevant and i receives the expected value $p_i^c(c_i, c_j)V$. We model the lottery draw that determines the non-contest prize using another random variable r_i^{nc}, independently drawn for each participant from $\mathcal{U}(0,1)$. In RS and RR, i receives the non-contest prize if $r_i^{nc} \leqslant p_i^{nc}(c_i, c_j)$. In SS and SR, i receives $p_i^{nc}(c_i, c_j)V$. Overall, realized payoffs $\pi_i(c_i, c_j, r_i^c, r_i^{nc})$ are a function of the chosen contest investment level and the realizations of up to two random variables, as specified in Table 1.

Besides the between-subject manipulation of payoff risk, we also manipulated foregone payoff information (FPI) within-subject. When FPI was not provided, players only learnt their realized payoff $\pi_i(c_i(t), c_j(t), r_i^c(t), r_i^{nc}(t))$, where the arguments denote the choices and realized values in round t. When FPI was provided, players also learnt what they would have

[9]Exactly one player in each pair receives the contest prize, therefore one independent draw is performed for each pair of participants. Assume that $r_1^c \sim \mathcal{U}(0,1)$, and $r_2^c = 1 - r_1^c$. Then r_2^c is also a variable with a standard uniform distribution.

Table 1: Realized payoffs in each treatment. Prize equals V, contest investment of player i is c_i and $r_i^c, r_i^{nc} \sim \mathcal{U}(0,1)$. Indicator function I is equal to 1 if the condition is satisfied, and 0 otherwise.

		Contest investment	
		Safe	Risky
Non-contest investment	Safe	SS $\pi_i = p_i^{nc}V + p_i^cV$	SR $\pi_i = p_i^{nc}V + I_{r_i^c \leqslant p_i^c}V$
	Risky	RS $\pi_i = I_{r_i^{nc} \leqslant p_i^{nc}}V + p_i^cV$	RR $\pi_i = I_{r_i^{nc} \leqslant p_i^{nc}}V + I_{r_i^c \leqslant p_i^c}V$

earned had they made a different choice, i.e., they were informed about $\pi_i(c_i, c_j(t), r_i^c(t), r_i^{nc}(t))$, $\forall c_i \in C$. All treatments started with two non-incentivized rounds, used to familiarize participants with the software, followed by 10 incentivized rounds in which FPI was not available, 20 rounds with FPI and 10 more rounds without FPI. The first 10 rounds provide a baseline comparison of treatments, with no additional information. The second block had 20 rounds, allowing us to study the long-run effects of FPI. If the payoff-based learning hypothesis is correct, a treatment difference should appear when FPI is introduced. The third block tests whether behavior observed in an environment with FPI persists when FPI is removed. If players consider the entire history of outcomes (as in reinforcement learning), behavior in the third block should be similar to the second block, and treatment differences would persist. If, on the other hand, behavior is affected only by immediate feedback (as in regret minimization, or directional learning), behavior and treatment order in the third block would resemble the first block.

At the end of the experiment, we collected demographic data and asked what action participants would recommend to a friend who would hypothetically take part in this experiment (adapted from Grosskopf et al., 2007). Responses to this question provide further insight into what players learn from the game, without the potential confounds of reputation building or exploration-exploitation tradeoff.

FPI was manipulated within-subject to understand the learning process of the participants who initially deviate from the equilibrium predictions. Evidence that the addition of information can change behavior would further support the hypothesis that the initial behavior is driven not by preferences (which should not change over rounds), but by the complexity of the environment. However, the within-subject design does not allow us to determine whether the change in behavior is driven by the introduction of FPI, or if it would have occurred without it. We address this question in Study 2, which compares treatments with FPI to a baseline in which FPI is never available.

Information was presented using a graphical interface that looked similar in all the treatments. The computer screens seen by the participants are reproduced in Figure H.1-H.6 (Appendix H).[10] The decision screen (Figure H.1) was identical in each round and each treatment, but the presentation of feedback varied. In each treatment, potential income from each source (contest and non-contest investment) was depicted as a box. In SS, the filling of each box represented sure earnings; income in the non-contest box was linearly increasing in non-contest investment and contest box was divided proportionally to contest investment. In

[10] A video illustrating the graphical interface can be viewed at `https://masiliunas.github.io/files/learning_in_contests_video.mp4`

SR, the screen looked similar, but the share of the contest box represented the probability to win. After players had observed their probability to win, a lottery was performed by placing a marker at a random location of the contest box. If the marker stopped in the player's area ($r_i^c \leqslant p_i^c$), the prize was won and the entire contest box was filled. In RS, squares in the non-contest box represented the probability to win, and the lottery was performed by placing a marker at a random location of the non-contest box. If the marker stopped in the player's area ($r_i^{nc} \leqslant p_i^{nc}$), the prize was won and the non-contest box was filled. In RR, two lotteries were performed, one for contest and one for non-contest investment. Two markers stopped at random locations of each box, and a player received zero, one or two prizes.

In addition to the payoff information, players received feedback about the action of the other participant, the received share of the contest prize in percentage and in points (in SS and RS) or the probability to win the contest prize (in SR and RR). When FPI was available, players were additionally informed about the probability to win and lottery outcomes for the unchosen actions, conditional on the realizations of r_i^c and r_i^{nc}.

4 Theoretical predictions

Behavior in contest experiments is systematically different from the predictions of the risk-neutral Nash equilibrium (Sheremeta, 2013), which could be attributed to the restrictive assumptions about the rationality or the utility function of the players. This section first shows the Nash equilibrium predictions under standard assumptions and then explores how the predictions change under bounded rationality or under alternative utility specifications.

4.1 Nash equilibrium

From Table 1 and definitions in (1) and (2), it can be verified that all treatments have identical expected payoffs, equal to:

$$E[\pi_i(c_i, c_j)] = \begin{cases} E - c_i + V\frac{c_i}{c_i+c_j} & \text{if } c_i + c_j > 0 \\ E + \frac{V}{2} & \text{otherwise} \end{cases} \tag{3}$$

A unique stage-game Nash equilibrium in all four treatments is $(c_i^*, c_j^*) = (V/4, V/4)$. In Study 1, we set $E = V = 8$, therefore the Nash equilibrium is $(2, 2)$. In Study 2, $E = V = 80$, therefore the Nash equilibrium is $(20, 20)$. We use the term "risk-neutral Nash equilibrium" (RNNE) for the Nash equilibrium calculated under the assumption of standard preferences, to avoid confusion with the Nash equilibria calculated with non-standard preferences. RNNE predictions are invariant to changes in payoff risk (because expected payoffs are held constant) and FPI (because the payoff function is known in all treatments), therefore RNNE predicts no difference between treatments and between rounds with and without FPI.

Contest investment above RNNE ("over-investment") is strictly dominated by the RNNE investment. With the discrete strategy space, only the RNNE action profile survives the iterated elimination of dominated strategies.

4.2 Payoff-based learning

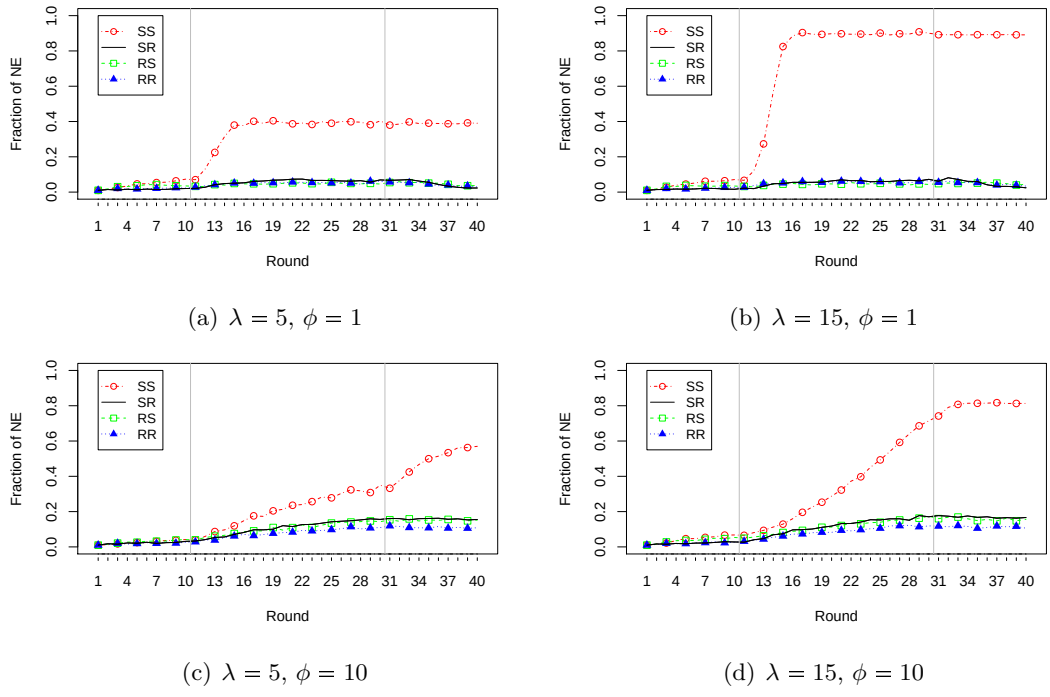
Calculating RNNE using only the game's description is cognitively demanding: participants would need to compute the expected payoffs for each action profile and either eliminate actions

that cannot be supported by any belief hierarchy or find a best-response to each possible action of the opponent. An alternative justification for the Nash equilibrium arises from it being the long-run outcome of learning dynamics, such as of belief learning (Lehrer and Kalai, 1993) or replicator dynamics (Weibull, 1997). Belief learning is less cognitively demanding than iterated elimination of dominated strategies, as it requires expected payoffs to be calculated not for all action profiles but only once for each action. Still, best-responding is difficult and in experiments the theoretical best-response is rarely chosen even when opponent's action is known (Fonseca, 2009, Masiliūnas et al., 2014). The most likely path to RNNE under bounded rationality is through payoff-based learning. The main challenge lies not with the cognitive abilities of the participants, but with the quantity and quality of observed feedback.

"Payoff-based" or "completely uncoupled" learning depends only on player's own past payoffs (Foster and Young, 2006). It has been shown that in some games some models in this class converge to RNNE (aspiration learning, or win-stay, lose-shift, Cho and Matsui, 2005, regret testing, Foster and Young, 2006, trial and error, Young, 2009, probe and adjust, Huttegger, 2013, reinforcement learning, Beggs, 2005). We use the most popular model in this class, reinforcement learning (Erev and Roth, 1998), which has been successfully used to explain deviations from expected payoff-maximization in individual choice tasks with FPI (Grosskopf et al., 2006, Otto and Love, 2010, Yechiam and Busemeyer, 2006, summarized in Section 2). Theoretically, reinforcement learning converges in contests through the iterated elimination of dominated strategies (Beggs, 2005), even when the payoffs are stochastic (Bravo and Mertikopoulos, 2017). In practice, learning can be slow and incomplete because of the payoff risk and high behavioral variability.

We changed two aspects of the game that should increase the convergence speed of payoff-based learning. The manipulation of payoff risk makes SS the only treatment in which payoffs are deterministic. However, even in SS the correlation between realized and expected payoffs is low because of the variability of the actions chosen by the opponent. The FPI manipulation solves the problem by enabling a direct comparison of all actions against the same distribution of opponent's actions. Convergence by iterated elimination of dominated strategies should be fast in SS with FPI because dominated strategies always generate low payoffs.

Formally, we obtain the predictions about the treatment difference by simulating the path of play for a reinforcement learning model (Roth and Erev 1995, Erev and Roth 1998). Denote the action space by $C \equiv \{c^1, c^2, \ldots, c^K\}$, where action c^k is a contest investment level indexed by k. Simulations are performed assuming that $C = \{0, 1, \ldots, 8\}$, as used in Study 1. Each action c^k has an associated attraction in round t, denoted $A_k(t)$. Initial attractions $A_k(0)$ are set to 0, for all k. The foregone or realized payoff from choosing action c^k in round $t \in \{1, 2, \ldots, 40\}$ is $\pi_k(t)$. Attractions are reinforced using only realized payoffs in rounds 1-10 and 31-40, and both realized and foregone payoffs in rounds 11-30. The reinforcement function is equal to the difference between the payoff and the reference point. We assume that the reference point is equal to the initial endowment (E), which is also the maximin payoff and the average payoff if both players choose all actions with equal probabilities. If the action's payoff is unobserved, the attraction remains unchanged. If it is observed (because it was played or because foregone payoffs were provided), the attraction in round t is a weighted average of the attraction in round $(t-1)$ and the received reinforcement:



(a) $\lambda = 5$, $\phi = 1$

(b) $\lambda = 15$, $\phi = 1$

(c) $\lambda = 5$, $\phi = 10$

(d) $\lambda = 15$, $\phi = 10$

Figure 2: Fraction of pairs choosing RNNE action profiles in reinforcement learning simulations. FPI is observed in rounds 11-30.

$$A_k(t) = \begin{cases} \frac{\phi A_k(t-1)+\pi_k(t)}{\phi+1} & \text{if } k \text{ is chosen or } t \in [11,\ldots,30], \\ A_k(t-1) & \text{otherwise.} \end{cases}$$

Attractions are averaged rather than cumulated, to keep the rate of learning constant across rounds. Attractions are mapped into choice probabilities using a logistic choice rule:

$$P_k(t) = \frac{e^{\lambda A_k(t-1)}}{\sum_j e^{\lambda A_j(t-1)}}$$

Reinforcement learning is governed by two parameters: ϕ measures the weight placed on the previous attractions and λ controls the sensitivity to differences between attractions. Simulations were run with four combinations of parameter values: the weight placed on previous history was either low ($\phi = 1$) or high ($\phi = 10$), and the sensitivity to differences between attractions was either low ($\lambda = 5$) or high ($\lambda = 15$). Agents were randomly rematched for 40 rounds within a 6-person group, and each simulation was repeated 1000 times. Figure 2 shows the average rate of RNNE play in matched pairs. The simulated difference between treatments is very small in the first ten rounds, but a gap between SS and the other three treatments appears when FPI is introduced and it remains even when the information is removed. A higher weight placed on recent payoffs (lower ϕ) speeds up convergence in SS but slows it down in the other three treatments. It happens because dominated strategies in SS always generate low payoffs, therefore a high weight placed on recently obtained payoffs

(low ϕ) and high sensitivity to differences between attractions (high λ) reduce the subsequent probability to choose dominated strategies and speeds up convergence through the iterated elimination of dominated strategies. In the other three treatments, convergence is slow when recent payoffs receive a high weight (low ϕ) because payoffs from a small number of rounds are noisy due to the probabilistic prize allocation. Payoff noise prevents the elimination of dominated strategies even when λ is high because dominated strategies can generate high average earnings in the short run.

We investigate the robustness of these results by considering a larger set of parameter values. Figure G.1 in Appendix G plots the simulated fraction of RNNE pairs in round 10 and round 30. In panel (a), λ is set to an intermediate value of 5, and ϕ is set to values between 0.1 and 20. In panel (b), ϕ is set to an intermediate value of 5, and λ is set to values between 1 and 20. One hundred simulations were run for each combination of parameters. Treatments are no different in round 10, but RNNE is more commonly chosen in SS than in the other three treatments in round 30, for all the considered parameter combinations.

FPI affects the path of choices because we assume that learning is driven by both realized and foregone payoffs. However, the origins of reinforcement learning lie in behaviorist psychology, which assumes that people respond only to realized payoffs. We interpret reinforcement learning more broadly, as a statistical method to estimate expected payoffs through the aggregation of observed payoff information.[11] This assumption is not uncommon in the literature; for example, experience-weighted attraction model (Camerer and Ho, 1999) includes a parameter measuring sensitivity to FPI ("law of simulated effect"), and sensitivity to FPI is used in reinforcement learning models to explain choices in experiments with full feedback (Yechiam and Rakow, 2012, Otto and Love, 2010).

4.3 Preferences and probability weighting

As an alternative to bounded rationality, we consider enriching the utility function with risk or other-regarding preferences, non-monetary utility of winning and non-linear probability weighting. We discuss the literature and show the predictions of these theories in Appendix B. If deviations from RNNE are driven by risk preferences, we should observe RNNE play in SS, which has no risk, and deviations from RNNE in all other treatments. Instead, if participants have other-regarding preferences, deviations from RNNE should be observed in all treatments because contest investment always reduces opponent's expected earnings. Non-monetary utility of winning and S-shaped probability weighting predict higher investment into the option with a probabilistic outcome, and similar behavior in SS and RR. Impulse balance equilibrium predicts RNNE in SS but not in the other three treatments because RNNE maximizes ex-post payoffs only in SS. It is also possible that several mechanisms operate at once; for example, participants might be both risk-averse and receive non-monetary utility from winning. We allow such combinations in Section 7, which compares the fit of reinforcement learning to a belief learning model with multiple elements in the utility function.

The only models that predict higher RNNE rates when FPI is provided are reinforcement learning (because FPI increases the quantity and quality of feedback) and impulse balance

[11] Grosskopf et al. (2006) use the term "fictitious play" to refer to a model that is sensitive to FPI, while "reinforcement learning" refers to a model that does not depend on foregone payoffs. Their formulation of the "fictitious play" model is almost identical to the one presented in this section (only with an added separation between exploration and exploitation, and no reference point). We do not use the term "fictitious play" to avoid confusion with models based on explicit updating of beliefs about opponent's type.

equilibrium (because FPI makes it easier to find the ex-post rational action and increases the intensity of impulses). Other models implicitly assume that the common knowledge of the payoff function is sufficient to calculate payoffs, therefore the provision of FPI does not affect the predictions.

5 Study 1

5.1 Design

Study 1 compared behavior in the four treatments explained in Section 3. In all treatments, participants made decisions without FPI in the first 10 rounds, followed by 20 rounds with FPI and 10 more rounds without FPI. In the rounds with FPI, participants could uncover the FPI of any action with a mouse click. Information revelation was costless and it was recorded to know what information players observed (Figure D.4 in Appendix D shows the dynamics of information revelation).

In each round, players were randomly rematched within a 6-person matching group. The prize and the endowment were set to 8 points and players could invest integer amounts between 0 and 8 points.[12] Four rounds were randomly chosen for payment, one from each block of 10 rounds. Players received instructions for each of the three parts only at the start of that part, so that decisions in the first part would not be influenced by information about the subsequent introduction of FPI.

A total of 144 participants took part in Study 1, 36 in each of the four treatments. The average duration of the experiment was 80 minutes and the average payment was €16.50. Experiments were programmed using z-Tree (Fischbacher, 2007) and run at the BEElab laboratory of Maastricht University. Subjects were recruited using ORSEE (Greiner, 2015).

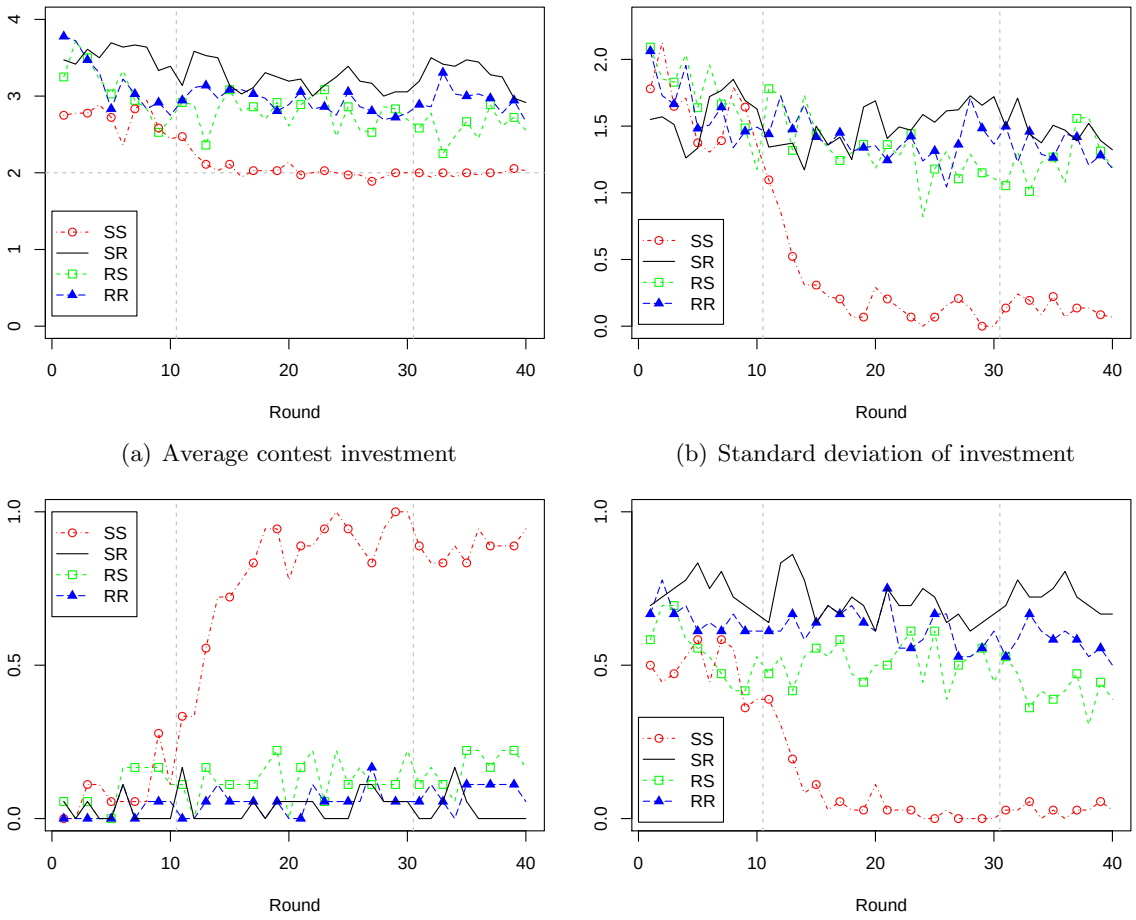
5.2 Results

We find that over-investment and over-spreading remain in the treatments with payoff risk, but disappear in SS when FPI is introduced. Panels (a) and (b) of Figure 3 show that the average contest investment and the standard deviation of investment[13] are high in the first ten rounds, but sharply decrease in SS following the introduction of FPI in round 10. In the other three treatments, high dispersion and investment persist throughout the game. We test if the difference between treatments is significant by calculating the average and standard deviation of contest investment in the last 10 rounds for each matching group. Table 2 shows the average outcomes in each treatment, as well as the two-sided p-values of Mann-Whitney U tests that compare SS to the other three treatments[14]. We find that both the average investment and the standard deviation of investment are significantly lower in SS than in each of the other three treatments (p-values at most 0.016), but there are no significant differences between SR, RS and RR (the most significant pairwise difference is in average contest investment between RS and SR, $p = 0.078$). We find the same pattern if we use data from rounds 11-30, instead of 31-40 (see Table D.2 in Appendix D). In the first ten rounds, the average investment is

[12] In the experiment, all earnings were denominated in "points" with an exchange rate of 1 point = €0.45.

[13] Standard deviation is calculated separately for each matching group, using the decisions of six participants in each group, and then averaged. See Figure D.1 in Appendix D for the evolution of the entire choice distribution.

[14] All the reported p-values are two-sided, unless indicated otherwise.



(a) Average contest investment

(b) Standard deviation of investment

(c) Fraction of pairs in RNNE

(d) Fraction of dominated strategies

Figure 3: Dynamics in each treatment. FPI was available from round 11 to round 30.

lower in SS than in SR ($p = 0.01$), but there are no significant differences between the other treatments (Table D.1 in Appendix D).

A treatment difference in average investment and dispersion does not indicate whether choices converge to the RNNE prediction. We therefore use three measures of convergence and compare them across treatments. First, we test whether the average investment is significantly different from the RNNE prediction. In the first ten rounds, contest investment is significantly different from RNNE in all treatments (one sample two-sided t-test p-values are at most 0.0111). In the last ten rounds, investment is not different from RNNE in SS (two-sided p-value = 0.8125) but different from RNNE in the other three treatments (all two-sided p-values below 0.02). Second, we compare the treatments in terms of the fraction of pairs in each matching group who choose the RNNE action profile. Panel (c) in Figure 3 indicates that convergence occurs only in SS and statistical tests confirm this observation (Table 2). RNNE rates are low in all treatments at the start of the game (the only significant difference is found between RS and SR, $p = 0.03$). In rounds 10-30, RNNE rates reach 100% in SS but remain at most 20% in SR, RS and RR. In the last block, the difference between SS and the other three treatments is highly significant (p-values below 0.004). Third, we look at

Table 2: Aggregate outcomes in rounds 31-40. Outcomes are average contest investment, standard deviation of contest investment, frequency with which both matched participants choose the RNNE action, average absolute deviation from the RNNE prediction and the frequency of dominated strategies. All measures are aggregated on the treatment level. Statistical significance is evaluated using two-sided p-values of a Mann-Whitney U test, when data is averaged on the matching group level.

	SS	SR	p-value	RS	p (vs SS)	RR	p (vs SS)
Contest investment	1.99	3.28	0.0039	2.60	0.0039	2.95	0.0156
SD	0.28	1.54	0.0039	1.36	0.0039	1.46	0.0039
RNNE investment	88%	3%	0.0036	17%	0.0033	8%	0.0038
RNNE deviation	0.07	1.61	0.0039	0.93	0.0039	1.24	0.0039
Dominated	3%	72%	0.0038	42%	0.0038	57%	0.0038

average RNNE deviation, defined as the absolute value of the difference between the chosen contest investment and the RNNE prediction. In the last ten rounds, deviation from RNNE is significantly lower in SS than in any of other three treatments (MWU $p = 0.0039$; see Table 2).

Deviations from RNNE could be rationalized by non-equilibrium beliefs, but dominated strategies should not be chosen even under much weaker assumptions on rationality. It is known that dominated strategies are commonly chosen in contests (Masiliūnas et al., 2014), and we test whether their frequency decreases when players receive good feedback on mistakes. Panel (d) in Figure 3 shows that there is no significant difference in terms of dominated strategies in the first ten rounds (the only significant difference is between SS and SR, MWU two-sided $p = 0.0103$). The gap between treatments appears after round 10, when the fraction of dominated strategies shrinks to 0% in SS but remains above 50% in the other treatments. Table 2 shows that the difference between SS and the other three treatments in the last ten rounds is highly significant ($p = 0.0038$), although the difference between RS and SR becomes significant as well ($p = 0.0247$).

Within each treatment, we can identify what has been learnt by comparing choices in the first ten rounds to the last ten rounds. Both blocks are comparable because they have the same number of rounds and are played without FPI. Contest investment significantly decreases in SS (paired t-test two-sided $p = 0.0069$), but not in the other treatments ($p = 0.074$ in RS, $p = 0.3976$ in SR and $p = 0.2612$ in RR). Figure D.2 in Appendix D compares investment distributions in the first ten and the last ten rounds. The fraction of RNNE choices increases from 18% to 32% in RR (t-test two-sided p-value = 0.0145), from 25% to 42% in RS (p = 0.0341), from 29% to 94% in SS (p = 0.0004), but remains at 17% in SR (p = 0.9249). An increase in the explanatory power of RNNE is therefore highest in SS, lower in RR and RS, and not observed in SR. Further evidence about what participants learn in the experiment comes from answers about the action that they would recommended to a friend in a hypothetical future experiment. RNNE action is recommended by 81% of participants in SS, 44% in RS, 36% in RR and only 8% in SR (see Figure D.3 in Appendix D for more details). The proportion is significantly higher in SS than the other treatments (test of proportions two-sided p-value is at least 0.0016) and it is lower in SR compared to the other treatments (p-value at least 0.0046). Both measures indicate that learning is strongest in SS and weakest in SR.

We find that the average contest investment in SS is significantly below the other three

treatments, and not different from the RNNE prediction. In the other three treatments, contest investment exceeds the RNNE prediction even at the end of the game. These findings are closest to the predictions of reinforcement learning. Social preferences never predict lower investment in SS than in the other three treatments. Risk-seeking preferences correctly predict the difference between SS and the other three treatments, but the effect size predicted by CRRA utility function is much smaller than the difference observed in experiments. The effect of risk preferences should also appear already in the first block. Non-monetary utility of winning, probability weighting and impulse balance equilibrium predict that if investment is lower in SS than in SR, then in RS it should be even lower. Experiments fail to find such pattern, instead showing that the presence of payoff risk is more important than its source. However, the comparison of theories based solely on the predicted rank of treatments excludes the possibility that multiple factors might interact. Section 7 will test whether the results can be explained by a richer model that combines learning and a utility function with a combination of preferences.

6 Study 2

6.1 Design

The purpose of Study 2 is twofold. First, we test whether FPI is necessary for the convergence in the SS treatment, found in Study 1. For this purpose, we compare the SS treatment to a baseline SS treatment in which FPI was never provided (SSB treatment). The difference between these treatments shows whether convergence might occur even without FPI. We also compare the SR treatment to a baseline without FPI (SRB treatment) to measure the effect of FPI in the probabilistic rent-seeking contest.

Second, we test whether the results of Study 1 reproduce when learning is more difficult. High convergence rates in the SS treatment might be a result of a rather coarse strategy space used in Study 1. In Study 2, we allow the participants to invest any integer amounts between 0 and 80, instead of 0-8 as in Study 1. A finer strategy space makes convergence more challenging; for example, reinforcement learning converges via the iterated elimination of dominated strategies, which requires three steps of iterated elimination when strategy space is 0-8, but six steps when it is 0-80. A finer strategy space should slow down the convergence and increase the amount of information that participants need to process, which might reduce the effectiveness of FPI.

Other parameters of the game were accordingly adjusted to suit the finer strategy space: the endowment and the reward were set to 80 points and the exchange rate was reduced by a factor of ten. To keep the same visual representation of the options as in Study 1, we used a two-stage decision-making procedure: first, participants chose their contest investment from the set $\{0, 10, \ldots, 80\}$ and then they were shown the ten options that were closest to the initial choice (for example, participants who chose 30 would subsequently see the details of set $\{26, 27, \ldots, 35\}$). The additional complexity of two-stage decisions could have delayed the experiment, which would be especially problematic for an online study, therefore we did not ask the participants to uncover FPI; instead, all FPI was visible in the rounds when it was available. Since the purpose of Study 2 was to replicate Study 1 and to explore the mechanism, information about the data acquisition process was no longer necessary.

We ran Study 2 online, instead of in the lab, which required additional adjustments to

the design.[15] First, we imposed a 45 second time limit for making each decision, and another 45 second limit for viewing feedback.[16] The time limit was needed so that the experiment could continue even if one participant became temporarily unavailable. Second, we reduced the size of the matching groups from 6 participants to either 4 or 6 participants.[17] The smaller group size was chosen to minimize the data loss if someone disconnected and the group could not continue (which never happened). We also added a 4 SGD show-up fee, in accordance with the IRB regulations, and changed the exchange rate to keep the incentives similar to Study 1, taking into account the added show-up fee, exchange rate between euros and Singapore dollars and the typical compensation for internet experiments in the National University of Singapore.

The online environment is challenging for experiments on learning, as experimenters have less control than in the laboratory and participants might be less attentive. For example, there is some evidence that participants are less likely to follow the instructions (Dickinson and McEvoy, 2021) and perform worse in an allocation task and in the cognitive reflection test (Li et al., 2021). Overall, we expect that learning in Study 2 would be more difficult due to the online environment, larger strategy space and the additional time limits. It is interesting to understand whether FPI and removal of risk still facilitate convergence when learning is more difficult.

Study 1 found little convergence in SR, RS and RR treatments; therefore, Study 2 compared only SS (where we found high convergence rates) and SR (which has been widely studied in the previous literature) to their baseline versions without FPI. Overall, we had four treatments:

- **SR treatment** was identical to the SR treatment from Study 1.
- **SS treatment** was identical to the SS treatment from Study 1.
- **SRB treatment** ("SR baseline") was identical to SR, but FPI was not available in any of the rounds. The experiment was still divided into three blocks of 10, 20 and 10 rounds, but at the start of the second and third block participants were informed that the game will be the same as in the first block.
- **SSB treatment** ("SS baseline") was identical to SS, but FPI was not available in any of the rounds.

All participants were students at the National University of Singapore, recruited from the NUS Centre for Behavioural Economics subject pool (CBElab) using ORSEE (Greiner, 2015).[18] Experiments were programmed using z-Tree (Fischbacher, 2007) and conducted over the internet using z-Tree unleashed (Duch et al., 2020). We conducted two sessions for each

[15] At the time of the data collection (summer 2021), conducting in-person experiments in the laboratory was not possible due to the Covid-19 safe management measures imposed by the Singapore government (Phase 2, Heightened Alert).

[16] If no decision was made, the previous round decision was used; in the first round, the investment would be drawn from a uniform distribution. In the experiment, no decision was not made on average 1% of the time. We include non-active decisions in the analysis, but results do not change if they are excluded.

[17] In each session, matching groups were formed dynamically: all but one group had 4 participants and the last group had either 4 or 6 participants. This was done to allow any even number of participants to take part, minimizing the number of participants who signed up but could not participate.

[18] The experiments were approved by NUS Institutional Review Board (NUS-IRB-2021-425).

treatment and 36 participants took part in SR and SRB, 38 in SS and 32 in SSB. In total, 142 participants took part in Study 2. The average duration of the experiment was 70 minutes and the average earnings were 21.3 SGD (at the time of the experiment, the exchange rate was 1 SGD = 0.75 USD).

6.2 Results

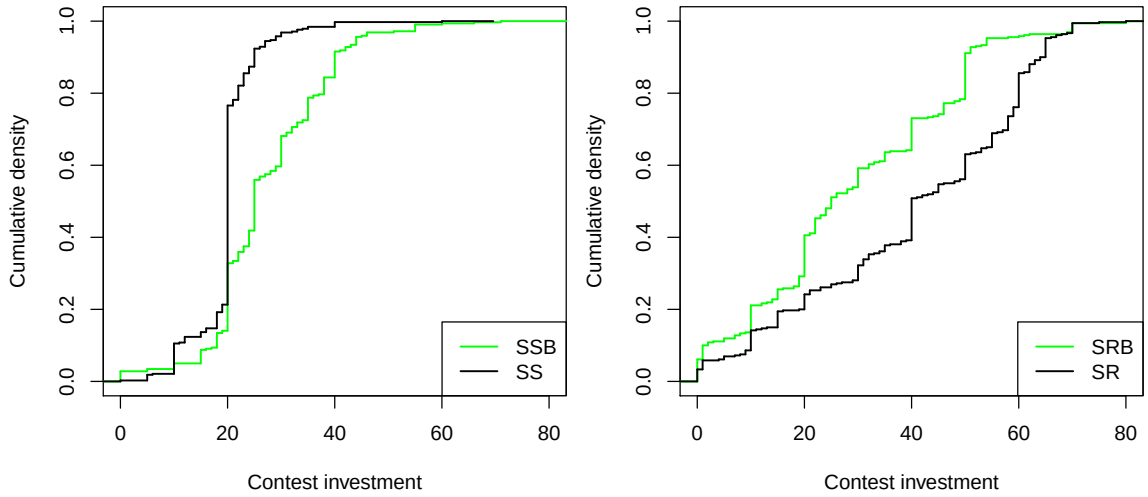


Figure 4: Cumulative distribution of contest investment in rounds 31-40. RNNE predicts an investment of 20.

Figure 4 shows the cumulative distribution of contest investment in the last ten rounds of each treatment. In these rounds, FPI was not available, but participants in SS and SR had previously experienced playing the game with FPI. In SS, the RNNE contest investment is chosen 55% of the time, while investing more than RNNE is as likely (23%) as under-investing (21%). In SSB, where participants had no previous experience with FPI, only 19% of choices are exactly at RNNE and 67% exceed RNNE. Overall, we find that FPI is necessary for convergence in SS. In contrast, experience with FPI does not facilitate convergence in SR, where RNNE is chosen very rarely (4% in SR, 11% in SRB) and players usually over-invest (76% in SR, 59% in SRB). The average contest investment in the last ten rounds significantly exceeds RNNE in SR (t-test two-sided $p = 0.0019$) and SSB ($p = 0.0104$), the difference is marginal in SRB ($p = 0.0611$) but investment is not significantly different from RNNE in SS ($p = 0.8557$). This evidence suggests that convergence to RNNE occurs only when participants have access to FPI in a treatment with no payoff risk, just as found in Study 1. We will further test this result by comparing the measures of convergence across treatments.

First, we compare several measures of convergence in SS and SR using data from the last ten rounds, just as we did in Study 1. Table 3 shows that at the end of the experiment, SS has significantly lower average investment (MWU two-sided $p = 0.0029$), lower standard deviation ($p = 0.0005$), higher frequency of Nash equilibrium play ($p = 0.0009$), lower average absolute deviation from the equilibrium prediction ($p = 0.0005$) and lower frequency of dominated strategies ($p = 0.0014$), compared to SR. Study 2 therefore replicates the finding that convergence rates are higher in SS than in SR.

To test whether FPI is necessary for convergence, we compare each treatment (SS or SR)

Table 3: Aggregate outcomes in rounds 31-40 in Study 2. Outcomes are (1) average contest investment, (2) standard deviation of contest investment, (3) frequency with which both matched participants chose the RNNE action, (4) frequency of dominated strategies and (5) average absolute deviation from the RNNE prediction. All measures are aggregated on the treatment level. Statistical significance is evaluated using two-sided p-values of a Mann-Whitney U test, when data is averaged on the matching group level.

	SR	SS	p-value	SRB	p (vs SR)	SSB	p (vs SS)
Contest investment	40.1	19.9	0.0029	28.9	0.2076	27.4	0.0050
SD	20.6	5.7	0.0005	18.5	0.0742	10.9	0.0036
RNNE investment	0%	37%	0.0009	4%	0.1441	3%	0.0134
RNNE deviation	24.8	3.0	0.0005	16.0	0.1415	9.6	0.0026
Dominated	76%	23%	0.0014	59%	0.4285	67%	0.0049

to their baseline version in which FPI was never available (SSB or SRB). Table 3 indicates that convergence is stronger in SS than in SSB: after the experience with FPI, investment is lower ($p = 0.005$) and has a lower standard deviation ($p = 0.0036$), RNNE rates are higher ($p = 0.0134$), choices are on average closer to RNNE ($p = 0.0026$) and dominated strategies are played less often ($p = 0.0049$). In contrast, FPI does not decrease contest investment or improve the explanatory power of RNNE in the probabilistic rent-seeking contest, as none of the differences between SR and SRB are significant (see Table 3). We also compare each treatment to their baseline version in the first block, before FPI was introduced, to test whether there are any differences in the randomization of participants to treatments (see Table E.2 in Appendix E). Out of all measures, a significant difference is found only in the standard deviation of investment, which is lower in SS than in SSB (15.0 vs 17.8, MWU two-sided $p = 0.039$). We conclude that FPI facilitates convergence in SS, but not in the SR treatment. In fact, all convergence measures are worse in SR than in SRB, suggesting that FPI can even increase the deviation from RNNE when payoffs are noisy, although the differences are not statistically significant.

We quantify how much participants learn by comparing their choices in the first ten rounds to the last ten rounds. First, we calculate the change in average contest investment. We find a significant decrease in SS (paired t-test two-sided $p = 0.0047$) and in SRB ($p = 0.0402$), a marginally significant decrease in SSB ($p = 0.0962$) and no change in SR ($p = 0.4297$). Results are similar for the change in average deviations from the equilibrium prediction, which significantly decrease in SS ($p = 0.0006$), and, to a lesser extent, also in SRB ($p = 0.0213$) and SSB ($p = 0.0359$), but not in SR ($p = 0.2782$). In addition, we test what participants learn from the entire experiment by comparing how often they would recommend the RNNE action to a friend. RNNE action is recommended by 63% of participants in SS, compared to 22% in SSB, 8% in SRB and 3% in SR (see Figure E.2 in Appendix E). The frequency of RNNE recommendations is significantly higher in SS than in SR (test of proportions two-sided $p < 0.0001$) or SSB ($p = 0.0005$), but there is no difference when comparing SSB to SRB ($p = 0.1155$) or SR to SRB ($p = 0.3035$). We conclude that learning is strongest in SS, as participants who experienced the environment with FPI subsequently decrease their contest investment, behave more in line with the RNNE predictions and more often recommend the RNNE action. Some learning occurs even without FPI (in SSB and SRB treatments), but it is much weaker than in SS.

As a robustness check, we replicate the analysis by excluding the decisions that were not actively chosen (due to no decision being made within the time limit) and find the same results. The results are also unchanged if instead of using the third block, we use the second block, in which FPI was available in SS and SR but not in SSB or SRB (see Table E.1 in Appendix E). For completeness, we also compare SSB to SRB, which have been studied in the previous literature. We find no significant difference in average contest investment, standard deviation of investment, average deviation from RNNE, average frequency of RNNE or dominated strategies, whether we use the decisions from all the rounds or only the last ten rounds (see Table E.3 in Appendix E).

Overall, Study 2 replicates the main finding from Study 1: convergence occurs only when there is no payoff risk and participants observe FPI. Support for this result is found even when learning is more difficult due to a finer strategy space, additional time limits and an online environment. As expected, these changes in design reduce the rates of RNNE play: in Study 1, nearly all pairs in the SS treatment converge to RNNE, but only a half do so in the same treatment of Study 2. Some convergence occurs even in the treatments without FPI, although it is much lower than in SS and choices exceed RNNE even after 40 rounds of play.

7 Learning and non-standard preferences

Studies 1 and 2 found that RNNE is played only in the treatment with no payoff risk and with FPI, just as predicted in reinforcement learning simulations. Next, we use the data from Study 1 to test whether reinforcement learning can explain the entire adjustment process in each treatment. We are also interested in comparing reinforcement learning to a model with non-standard preferences and probability weighting. The comparison would not be fair if we modeled preferences using a static solution concept, therefore we use a belief learning model extended with non-standard preferences and probability weighting. We then combine the two models using an extension of EWA. This estimation strategy identifies how much the addition of non-standard preferences improves the fit of a belief learning model, whether it fits better than reinforcement learning, and how much each of these models contributes to the fit of the most general EWA model.

Many previous papers compared the fit of reinforcement and belief learning. It is usually found that either learning model outperforms the point predictions of a static Nash equilibrium and the relative fit depends on the features of the game; for example, reinforcement learning tends to do better in longer games (Feltovich, 2000) and when participants are informed only about own payoffs (Blume et al., 2002). Additional support for payoff-based learning comes from similar adaptation patterns in games with full information and a low information environment, where participants are informed only about their own payoffs (see Nax et al., 2016, for public goods games). We are also interested in understanding the importance of FPI, which has been studied in a few previous studies. For example, it was found that EWA fits the data from centipede games better if FPI is included in the model (Ho et al., 2008) and a model that assigns similar weights to foregone and realized payoffs performs better than the alternatives in individual choice tasks (Yechiam and Rakow, 2012).

7.1 Adaptation rules

7.1.1 Reinforcement learning

Denote the action space by $C \equiv \{c^1, c^2, \ldots, c^K\}$, and the action of player i in round $t \in \{1, 2, \ldots, 40\}$ by $c_i(t) \in S$. The reinforcement learning model follows Roth and Erev (1995) and Erev and Roth (1998) in assuming that the attraction of action c^k, for $k \in \{1, 2, \ldots, K\}$, is a weighted sum of the payoff flow $\pi_k(t)$ that was generated or would have been generated by playing c^k:

$$A_k(t) = \phi A_k(t-1) + \delta_k(t)\pi_k(t) \tag{4}$$

Parameter $\phi \in [0, 1]$ determines the rate at which old payoff information is discounted. If $\phi = 1$, all past payoffs receive the same weight. If $\phi = 0$, only the most recent payoff is taken into account. Variable $\delta_k(t)$ does not appear in reinforcement learning models (e.g., Erev and Roth, 1998) and is added to allow learning from FPI. It attains one of the following values:

$$\delta_k(t) = \begin{cases} 1 & \text{if } c^k \text{ was chosen in round } t, \\ \delta_o & \text{if } c^k \text{ was not chosen, but its FPI was observed,} \\ \delta_a & \text{if } c^k \text{ was not chosen, its FPI was available but not observed,} \\ \delta_u & \text{if } c^k \text{ was not chosen and FPI was unavailable.} \end{cases} \tag{5}$$

Chosen actions are reinforced using realized payoffs and a weight of $\delta_k(t) = 1$. Unchosen actions are reinforced using foregone payoffs and a weight $\delta_k(t) \in [0, 1]$.[19] We allow the weight to depend on the type of FPI (observed, available but not observed, unavailable). If δ_o, δ_a or $\delta_u = 0$, that class of FPI is ignored. If δ_o, δ_a or $\delta_u = 1$, that class of FPI receives the same weight as the realized payoff.

7.1.2 Belief learning with non-standard preferences

The second learning model assumes learning from observed opponent's actions. Denote the strategy chosen by i's opponent by $c_{-i}(t)$. Using an indicator function $I(c^k, c_{-i}(t))$, which equals 1 if the opponent's chosen action in round t is c^k, and 0 otherwise, define the weighted frequency of strategy k played by i's opponent up to time t by $N^k(t) = \phi N^k(t-1) + I(c^k, c_{-i}(t))$. If $\phi = 1$, $N^k(t)$ reduces to the total number of times that action c^k was played by the opponent. Beliefs are normalized to attain values between 0 and 1 using an experience weight $N(t) = \phi N(t-1) + 1$. The belief about the probability that i's opponent will play c^k in round t is then calculated by:

$$b_k(t) = \frac{N^k(t)}{N(t)} = \frac{\phi^t N^k(0) + \sum_{u=1}^{t} \phi^{u-1} I(c^k, c_{-i}(t+1-u))}{\phi^t N(0) + \sum_{u=1}^{t} \phi^{u-1}} \tag{6}$$

Initial belief is calculated as $b_k(0) = \frac{N^k(0)}{N(0)}$, where $N^k(0)$ represents experience prior to the experiment, and $N(0)$ is its weight.

[19] A similar approach was used to model reinforcement learning with FPI by Yechiam and Rakow (2012) and Yechiam and Busemeyer (2006).

Attractions used to calculate choice probabilities are calculated as a belief-weighted average of expected utility:[20]

$$A_k(t) = \sum_{j=1}^{K} E[u_i(c^k, c^j)]b_j(t) \tag{7}$$

If expected utility equals expected payoff, the belief learning model reduces to weighted fictitious play by Cheung and Friedman (1997). The extension to expected utility permits non-standard preferences and learning to be modeled in one unified framework.

7.1.3 Generalized Experience-Weighted Attraction learning

Next, we introduce a generalized experience-weighted attraction (EWA) learning model (Camerer and Ho, 1999), which combines belief and reinforcement learning. In EWA, belief learning is implemented through sensitivity to FPI. Instead of forming explicit beliefs from the weighted path of observed choices and calculating expected payoffs conditional on these beliefs, players are directly calculating the weighted average of foregone payoffs in all previous rounds, which depend on the path of opponent's choices. The equivalence between belief and reinforcement learning exists only if the foregone payoffs of all actions receive the same weight, and if there is no payoff risk. If payoffs are risky, expected foregone payoffs must be used for EWA to be equivalent to weighted fictitious play. Generalized EWA (adapted from Shafran, 2012) uses both the expected payoffs, needed for belief learning, and the realized and foregone payoffs, needed for reinforcement learning.

The updating rules are governed by the experience weight $N(t)$ and attractions $A_k(t)$:

$$N(t) = \phi(1-\kappa)N(t-1) + 1 \tag{8}$$

$$A_k(t) = \frac{\phi N(t-1)A_k(t-1) + \gamma E[u(c^k, c_{-i}(t))] + (1-\gamma)\delta_k(t)\pi_k(t)}{\phi N(t-1)(1-\kappa) + 1} \tag{9}$$

This version of generalized EWA extends Shafran (2012) by allowing the weight of foregone payoffs to depend on the observability ($\delta_k(t)$ is calculated using equation 5) and by permitting expected utility. Generalized EWA reduces to standard EWA if there is no payoff risk (as in SS), $\gamma = 0$ and $\delta_o = \delta_a = \delta_u$. It reduces to cumulative reinforcement learning (equation 4) if $N(0) = 1$, $\gamma = 0$ and $\kappa = 1$. It reduces to averaged reinforcement learning if $N(0) = \frac{1}{1-\phi}$, $\gamma = 0$ and $\kappa = 0$. It reduces to weighted fictitious play (equation 6) if $\gamma = 1$, $\kappa = 0$ (see Camerer and Ho, 1999, and Shafran, 2012, for a proof).

We set $N(0) = 1$, as is common in the literature (see Camerer, 2003). Initial attractions $A_k(0)$ were chosen to maximize the likelihood of first round observations (Ho et al., 2008).[21]

Since the attractions depend on expected utility instead of expected payoffs, parameters such as λ are no longer comparable across specifications. For example, risk aversion would reduce all expected utilities, therefore models with risk aversion would need to "compensate"

[20] Expected utility of player i from playing strategy c^k and opponent playing c^j is denoted by $E[u_i(c^k, c^j)]$.

[21] If f^k is the frequency of strategy c_i^k in round 1, $A_k(0) = \log(f^k)/\lambda$ if $f^k > 0$ and $A_k(0) = 0$ otherwise. It is straightforward to verify that these initial attractions generate the appropriate initial choice frequencies, but they are not unique, as the exponential choice rule is invariant to adding a constant to all attractions. Alternatively, we could have estimated the initial attractions as separate parameters, but additional 8 parameters would make model fitting very challenging.

using a larger λ value to obtain the same noise level. To preserve comparability, we use mean normalization, subtracting the mean of all expected utilities and dividing by the range between the highest and lowest expected utility.[22] Consequently, normalized expected utilities have the range of one and the mean of zero in each treatment. We normalized the realized payoffs using the same approach, to keep the estimated values of λ comparable in reinforcement and belief learning models. However, we performed normalization using the expected payoffs (as in the SS treatment), so as not to change the estimated values of the other parameters, compared to the case without normalization.

The combination of expected utilities and realized payoffs used in the generalized EWA allows us to identify the relative importance of reinforcement and belief learning using only the γ parameter, while δ measures purely the difference between realized and foregone payoffs, manipulated in the experiment. In the standard EWA, δ could show either the difference between belief and reinforcement learning, or the sensitivity to FPI.

In all models, the probability to choose action k is calculated using a logistic choice rule:

$$P_k(t) = \frac{e^{\lambda A_k(t-1)}}{\sum_{j=1}^{K} e^{\lambda A_j(t-1)}}$$

where parameter $\lambda \in [0, \infty)$ measures sensitivity to differences between attractions.

We estimate the parameters that maximize the following log-likelihood function:

$$LL = \sum_{i=1}^{96} \sum_{t=1}^{40} \log(P_{c_i(t)}(t))$$

A potential problem for generalized EWA is overfitting, as models with many parameters may fit well but make inaccurate predictions. For this reason, we estimate parameters using 2/3 of the observations from Study 1 (96 participants), and use the remaining 1/3 (48 participants) to evaluate the out-of-sample goodness of fit.[23] Since a higher number of parameters increases log-likelihood, we measure the in-sample fit using the Akaike information criterion (AIC) and the Bayesian information criterion (BIC), which penalize models for the number of parameters.[24]

Data for all estimations is pooled from all four treatments and all rounds. We do so to test if one set of parameter values can explain the differences between all treatments and information conditions. If parameters were allowed to vary by treatment, we would likely observe overfitting, as models would predict less noise and lower strength of non-standard preferences in SS compared to the other three treatments.

Estimations were performed using quasi-Newton and derivative-free optimization routines with various starting values and the following constraints: $\lambda > 0$, $\phi, \kappa, \delta, \gamma \in [0, 1]$, $s \in [-5, 5], r \in [-2, 1), \omega \in [0, 16], \beta \in [0, 3]$. Standard errors were calculated from the variance-covariance matrix, estimated from the numerical approximation of the Hessian matrix.

[22] Define the normalized expected utility by $NEU_i(c^k, c^j) = \frac{E[u_i(c^k,c^j)] - \sum_{c^k \in C, c^j \in C} E[u_i(c^k,c^j)]/|C|^2}{\max_{c^k \in C, c^j \in C} E[u_i(c^k,c^j)] - \min_{c^k \in C, c^j \in C} E[u_i(c^k,c^j)]}$, for all $c^k \in C$ and $c^j \in C$.

[23] Study 1 has 6 independent matching groups in each treatment, thus we use the first 4 groups for estimation (i.e., the four groups that were run in earlier sessions) and the last 2 for out-of-sample fit.

[24] AIC $= -2\log(\mathcal{L}) + 2k$, BIC $= -2\log(\mathcal{L}) + k\log(N)$, where k is the model degrees of freedom and N is the number of observations.

7.2 Estimation results

Since generalized EWA nests belief and reinforcement learning, we estimate all models using equation (9) and obtain belief or reinforcement learning by appropriately constraining the parameter values.

7.2.1 Reinforcement learning

Reinforcement learning is obtained by setting $\gamma = 0$. We allow κ to vary between 0 and 1, to allow for both averaging and cumulation of attractions, although in all models the estimated value of κ is equal to 1 (see Table 4), reducing the updating rule to equation (4).

We estimate and compare the fit of five reinforcement learning models that differ in assumptions about how FPI is used to update attractions:

1. FPI is ignored, so only realized payoffs receive a positive weight. This assumption is made in classical reinforcement learning models, e.g., Erev and Roth (1998).

$$\delta_a = \delta_o = \delta_u = 0$$

2. All FPI receives the same weight. This assumption is made in studies that do not provide explicit FPI. In the treatment with no payoff risk, this model is equivalent to the standard EWA model, and δ measures sensitivity to foregone expected payoffs.

$$\delta_a = \delta_o = \delta_u \in [0, 1]$$

3. Available FPI receives a different weight than unavailable FPI. This is the approach taken by studies that explicitly provide FPI (Yechiam and Rakow, 2012, Yechiam and Busemeyer, 2006).

$$\delta_o = \delta_a \in [0, 1], \delta_u \in [0, 1]$$

4. Only observed FPI receives a positive weight. This specification is unique to our study because we have information about the FPI that was observed.

$$\delta_o \in [0, 1], \delta_a = \delta_u = 0$$

5. Each type of FPI receives a different weight. This is the most general model, allowing attractions to be affected differently by each type of foregone payoffs.

$$\delta_o \in [0, 1], \delta_a \in [0, 1], \delta_u \in [0, 1]$$

Table 4 lists the estimated parameter values and the goodness of fit for all five reinforcement learning models. The standard model (1) that ignores FPI has a much worse fit than all other models. If all three types of FPI are treated the same way (model 2), the estimated weight placed on foregone payoffs is 0.14, much smaller than the weight of 1 that realized payoffs receive. Allowing a different weight for available but unseen FPI (model 3) does not improve the fit, as the estimated value of δ_a remains unchanged. If it is assumed that players

Table 4: Estimated parameter values and goodness of fit for reinforcement learning models in which (1) FPI is ignored, (2) all FPI is perceived the same, (3) available FPI receives a different weight than unavailable FPI, (4) only observed FPI receives a positive weight, (5) each type of FPI receives a different weight. Standard errors are in parentheses. Values in square brackets indicate the exogenously set parameter values.

	(1)	(2)	(3)	(4)	(5)
λ	1.63	2.04	2.04	1.81	2.06
	(0.17)	(0.26)	(0.26)	(0.20)	(0.26)
ϕ	0.96	0.94	0.94	0.95	0.94
	(0.0038)	(0.0042)	(0.0042)	(0.0041)	(0.0042)
κ	1	1	1	1	1
	(0.078)	(0.11)	(0.11)	(0.09)	(0.11)
δ_u	[0]	0.14	0.14	[0]	0.14
		(0.011)	(0.020)		(0.020)
δ_a	[0]	[δ_u]	0.14	[0]	0.11
			(0.016)		(0.019)
δ_o	[0]	[δ_u]	[δ_a]	0.30	0.24
				(0.037)	(0.036)
LL (in)	-5568.37	-5481.71	-5481.71	-5534.49	-5476.81
AIC	11142.73	10971.41	10973.41	11076.97	10965.61
BIC	11161.49	10996.43	11004.68	11101.99	11003.13
LL (out)	-2623.07	-2559.84	-2559.81	-2581.60	-2552.30

react only to observed FPI, while unseen FPI is ignored (model 4), the fit is worse compared to the two previous models. Allowing different weights for each type of FPI (model 5) has the best fit, both in-sample and out-of-sample, even when accounting for the additional number of parameters (although model (2) does better in terms of BIC, which penalizes additional parameters more than AIC). In model (5), observed FPI receives twice higher weight than unobserved FPI, although it is still much lower than the weight of realized payoffs (normalized to 1). The estimated value of δ_o is similar to that in model (4). We conclude that information about foregone payoffs affects the adaptation process and including it improves the fit. Process data on information revelation further improves the fit, as model (5) that includes such data fits better than model (3).

7.2.2 Belief learning with non-standard preferences

We estimate the belief learning model with risk and social preferences, non-monetary utility of winning and non-linear probability weighting (see Appendix B for more details). Additionally, non-monetary utility from winning is allowed to depend on the type of contest. Lottery outcome $l \in \{1, 2, \ldots, L\}$ is therefore defined as $(\pi_i^l, \pi_j^l, C_i^l, NC_i^l)$, where $C_i^l = 1$ if i receives the contest prize and $NC_i = 1$ if i receives the non-contest prize (and 0 otherwise). The payoffs of i and j are denoted by π_i and π_j. Utility at each outcome is calculated as:

$$u_i(\pi_i, \pi_j, C_i, NC_i) = \frac{\pi_i^{1-r}}{1-r} + \omega_c C_i + \omega_{nc} NC_i + s\pi_j \tag{10}$$

where s is the social preference parameter, r is the constant relative risk aversion (CRRA)

Table 5: Estimated parameter values and goodness of fit for belief learning models with (6) standard preferences, (7) social preferences, (8) risk preferences, (9) non-monetary utility of winning, (10) probability weighting and (11) all preferences combined. Standard errors are in parentheses. Values in square brackets indicate the exogenously set parameter values.

	(6) NO	(7) SOC	(8) RISK	(9) NMU	(10) PW	(11) ALL
λ	0.66	10.56	1.14	2.23	0.73	1.46
	(0.11)	(0.38)	(0.22)	(0.47)	(0.13)	(0.21)
ϕ	0.97	1	0.93	0.91	0.97	0.94
	(0.0076)	(0.033)	(0.011)	(0.011)	(0.008)	(0.008)
κ	1	0.03	1	0.84	1	1
	(0.086)	(0.01)	(0.12)	(0.15)	(0.090)	(0.11)
s	[0]	-0.27	[0]	[0]	[0]	-0.78
		(0.01)				(0.25)
r	[0]	[0]	-0.79	[0]	[0]	-1.18
			(0.067)			(0.04)
ω_c	[0]	[0]	[0]	3.95	[0]	16
				(0.18)		(1.19)
ω_{nc}	[0]	[0]	[0]	0	[0]	0
				(0.13)		(0.72)
β	[1]	[1]	[1]	[1]	1.31	1.70
					(0.03)	(0.06)
LL(in)	-6889.65	-6731.22	6803.27	-6569.05	-6818.30	-6518.71
AIC	13785.3	13470.45	13561.19	13148.09	13644.61	13053.41
BIC	13804.05	13495.46	13586.21	13179.36	13669.62	13103.44
LL (out)	-3306.83	-3321.89	-3311.41	-3349.78	-3310.42	-3360.67

coefficient, ω_c is non-monetary utility from winning the contest prize and ω_{nc} is non-monetary utility from winning the non-contest prize.

Objective probabilities of lottery outcomes are transformed to subjective weights using the cumulative prospect theory weighting function (Tversky and Kahneman, 1992):

$$w(p) = \frac{p^\beta}{(p^\beta + (1-p)^\beta)^{1/\beta}} \tag{11}$$

If the probability of lottery outcome l is p_l, expected utility is calculated as:

$$E[u_i] = \sum_{l=1}^{L} w(p_l) u_i(\pi_i^l, \pi_j^l, C_i^l, NC_i^l) \tag{12}$$

Equation (12) is used to calculate expected utilities in equation (9). Parameters of the utility function are estimated jointly with the parameters of the learning model. Since we use EWA formulation, the belief learning model is comparable to reinforcement learning, having identical initial attractions and the same flexibility in aggregating past payoff information. The sole difference is the use of expected utility instead of realized payoffs.

Table 5 shows the estimated parameter values for belief learning. The baseline model (6) with standard preferences has a poor fit, much worse than the baseline reinforcement learning

Table 6: Estimated parameter values and goodness of fit for generalized EWA models that combine belief and reinforcement learning. Standard errors in parentheses. Values in square brackets indicate the exogenously set parameter values.

	(12)	(13)	(14)	(15)
λ	2.63	2.64	3.28	3.46
	(0.37)	(0.38)	(0.36)	(0.47)
ϕ	0.92	0.92	0.91	0.91
	(0.0051)	(0.0052)	(0.0051)	(0.0053)
κ	1	1	1	0.97
	(0.13)	(0.14)	(0.12)	(0.15)
δ_u	[0]	0	[0]	0
		(0.027)		(0.030)
δ_a	[0]	0	[0]	0
		(0.028)		(0.031)
δ_o	[0]	0.075	[0]	0.074
		(0.046)		(0.051)
s	[0]	[0]	-0.04	-0.036
			(0.023)	(0.024)
r	[0]	[0]	0.65	0.66
			(0.056)	(0.062)
ω_c	[0]	[0]	1.65	1.86
			(0.15)	(0.16)
ω_{nc}	[0]	[0]	0	0
			(0.28)	(0.30)
β	[1]	[1]	0.98	0.95
			(0.028)	(0.0076)
γ	0.25	0.25	0.43	0.43
	(0.014)	(0.020)	(0.017)	(0.034)
LL(in)	-5370.27	-5368.92	-5219.50	-5217.60
AIC	10748.55	10751.83	10457	10459.19
BIC	10773.56	10795.6	10513.27	10534.23
LL (out)	-2514.26	-2508.27	-2565.75	-2563.91

model. The next four models consider each of the four expected utility modifications. Each one fits better than the baseline model, even when accounting for the number of parameters. However, the out-of-sample fit is reduced in all of the models, a result driven by very small differences between SR, RS and RR in the out-of-sample groups. Of the four theories, non-monetary utility of winning fits the best on all measures, except for the out-of-sample fit, where it fits the worst. The estimated parameter values indicate that the data is best explained if participants were anti-social, risk-seeking, received non-monetary utility from winning the contest prize, but no non-monetary utility from winning the non-contest prize, or probability weighting was S-shaped. When all non-standard preferences are estimated jointly (model 11), the estimated values of preference parameters remain similar, although further away from the standard preferences. The model that combines all four components of the utility function fits better than other models, although still worse than the baseline reinforcement learning model. However, the out-of-sample fit of the combined model is low and indicates potential overfitting.

7.2.3 Generalized EWA

Next, we estimate the parameters for EWA model that combines reinforcement and belief learning. To maintain a meaningful interpretation of the γ parameter, we rescale the utilities from belief learning models, keeping average utilities equal to average payoffs. Table 6 shows the estimated parameter values. Model (12) is identical to generalized EWA (Shafran, 2012) and includes neither FPI nor the non-standard utility parameters. This model fits data better than any of the belief or reinforcement models, and the low estimated value of γ suggests that it is mainly driven by reinforcement rather than belief learning. The addition of foregone payoff parameters in model (13) improves the fit, but not by much, as the estimated δ parameters are all close to zero, except for the observed foregone payoffs. The additional utility function parameters in model (14) improve the in-sample fit further, although the out-of-sample fit is reduced. The estimated utility function parameters are of similar magnitude to the beliefs-only model parameters reported in Table 5. Model (15) is the most general and the estimated parameter values are similar to model (14). As in model (13), only the observed foregone payoff parameter δ_o receives a positive weight, which drives the improved log-likelihood compared to model (14), although the improvement is not sufficient to increase the fit in terms of AIC or BIC due to the three additional parameter values.

7.3 Goodness of fit

To understand why some models fit better than others, we compare the average contest investment and RNNE play rates observed in experiments (panels (a) and (c) in Figure 3) to the predictions made by the best-fitting model in each class (the most interesting predictions are displayed in Figure 5). The baseline reinforcement learning model (1) accurately predicts the difference of RNNE frequency between SS and the other three treatments, as well as the treatment order at the end of the experiment. Estimated parameter values of model (4) are close to model (1), therefore the two share a similar in-sample fit. Models (2), (3) and (5) make similar predictions because they add similar weights to all FPI, although learning in (5) is slightly faster due to the positive weight placed on unobserved FPI and a higher value of λ. All reinforcement models replicate the general features of the data, except for the under-predicted speed of convergence in SS.[25]

The baseline belief learning model (6) fails to predict persistent over-investment and a difference between treatments, because expected payoffs are identical by design, and the distribution of beliefs differs only in SS. The addition of social preferences in model (7), risk preferences in model (8) or probability weighting in model (10) improves the fit by predicting slightly higher contest investment, but with no observable treatment difference. Model (9) assumes non-monetary utility of winning only for the contest prize, while the non-monetary utility from winning the non-contest prize is estimated to be zero. The model therefore predicts higher average contest investment and slightly lower RNNE frequency in SR and RR than in RS and SS. Model (11) combines all non-standard preferences and probability weighting: non-monetary utility increases the predicted contest investment in RR and SR,

[25]Fast learning in the SS treatment when FPI is introduced might be of different type than the slow learning in the other treatments. Therefore, dynamics at the start of the second block in the SS treatment might be better explained by models of epiphany learning (Chen and Krajbich, 2017), in which agents that recognize a dominant strategy play it for the rest of the game. Such type of epiphany learning might occur in other dominance solvable games, such as the ascending clock auctions (Li, 2017), Nim games (Dufwenberg et al., 2010; McKinney and Van Huyck, 2013) or two-player beauty contests (Chen and Krajbich, 2017).

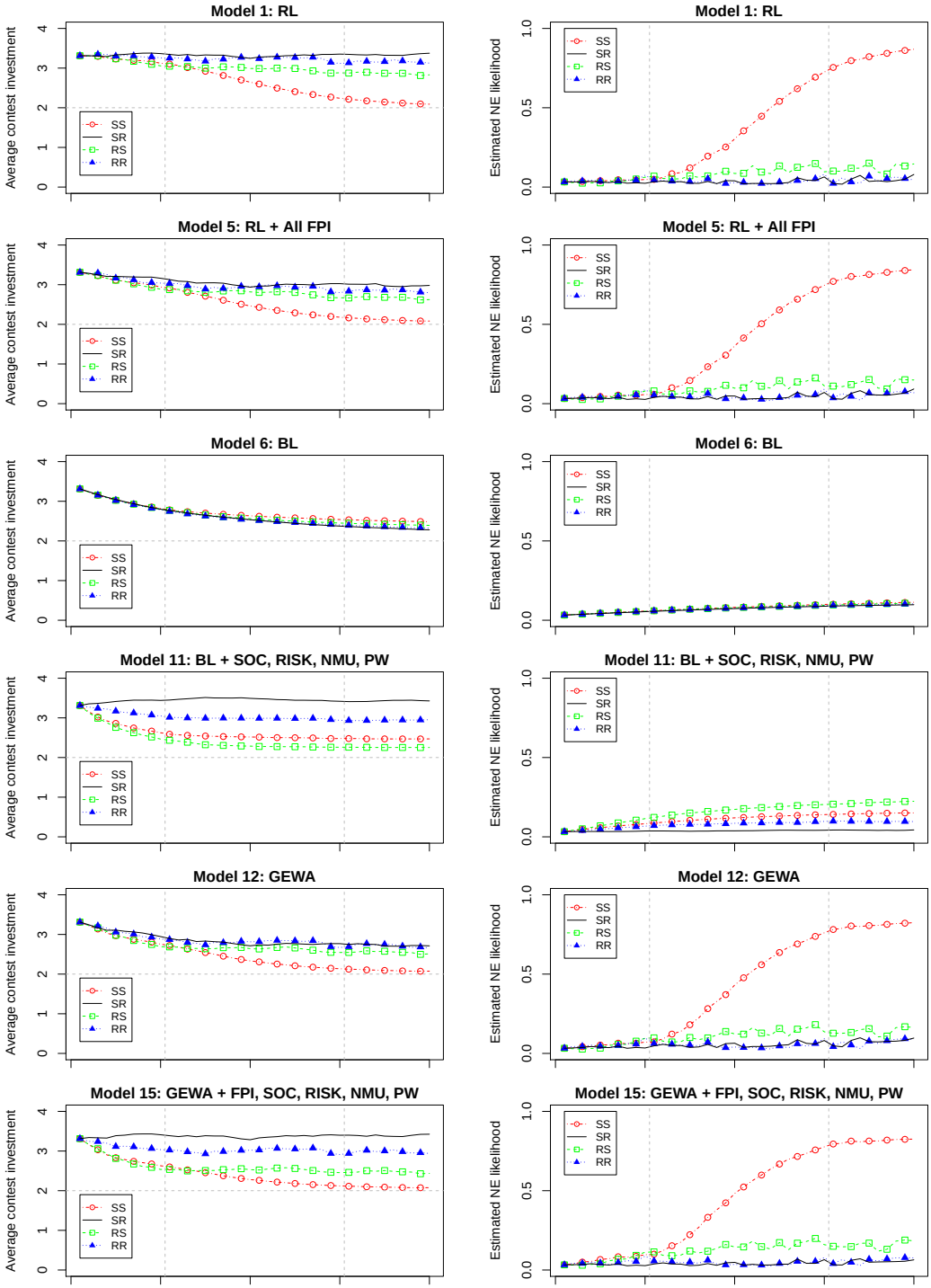


Figure 5: Average contest investment (left) and RNNE frequency (right), estimated using one-step-ahead predictions.

while probability weighting additionally increases the prediction in SR and decreases in RS. The full belief learning model correctly explains the treatment ranking in terms of average choices between SR, RS and RR, but fails to explain lower investment and higher RNNE rates in SS.

The generalized EWA model (12) that combines belief and reinforcement learning explains lower contest investment and higher RNNE rates in SS, compared to the other three treatments, but fails to predict a difference between SR, RS and RR. The difference between these three treatments is not predicted even with the added sensitivity to FPI (model 13), but it is predicted with the addition of non-standard preferences in models (14) and (15). Thus the main contribution of reinforcement learning lies in explaining lower contest investment and higher frequency of RNNE in SS (compare the belief model 11 to the full EWA model 15), while the main contribution of non-standard preferences and probability weighting is in explaining differences between SR, RR and RS (compare the reinforcement learning model 5 to the full EWA model 15). But while both models improve the fit, the contribution of reinforcement learning is much larger: it improves the total log-likelihood by 1301 points (difference between model 15 and 11), while the improvement from belief learning with non-standard preferences is only 259 points (difference between model 15 and 5).

Results from the estimations should be interpreted with caution. It is known that when learning models are misspecified, EWA fails to accurately identify populations from simulated data 50% of the time (Salmon, 2001). The problem is even more severe if the population is heterogeneous, as is often true in contest experiments; in that case, the estimated parameters could be very inaccurate (Wilcox, 2006). In particular, there tends to be a large downward bias in the estimation of the δ parameter, which in our model is interpreted as sensitivity to FPI. It is thus likely that participants are more sensitive to FPI than we found. The model that combines EWA with non-standard preferences is completely new, thus it is unknown to what extent and in which direction the estimated preference parameters could be biased. One benefit of our estimation strategy is that all models are estimated using the data from all four treatments (in contrast to using one game, as is typically done in the literature), and models differ in how the variation in risk is predicted to affect behavior. Thus even if the estimated parameter values were biased, the ranking of models in terms of the goodness of fit should remain similar. Ultimately, future research should empirically study the potential biases of the estimated parameter values, perhaps by testing how well the parameter values can be retrieved from simulated data.

8 Concluding remarks

We test whether deviations from Nash equilibrium in rent-seeking contests occur because learning is subject to sampling error caused by the low informational value of realized payoffs. Payoff-based learning simulations show that participants would take a very long time to discover the payoff-maximizing action using only information about realized payoffs. Experiments show that when the sampling error is reduced by removing payoff risk and providing foregone payoff information, risk-neutral Nash equilibrium play rates are significantly increased. These results can be explained by payoff-based learning, but not preference-based theories. The learning hypothesis is further supported by the findings that participants continue behaving as predicted by the risk-neutral Nash equilibrium even when foregone payoff information is removed, and most participants would recommend such a strategy to their

friends.

Our results complement rather than substitute the preference-based explanations proposed in the rent-seeking contest literature. The finding that participants respond to foregone payoff information as predicted by payoff-based learning does not mean that participants also use payoff-based learning when foregone payoffs are not observed. In fact, it is likely that when learning is difficult, participants do not even attempt to learn and rely on heuristics instead. For example, they might focus on the chances to win or the payoff of other participants, which are easier to evaluate than the expected payoffs. Sheremeta (2018) shows that impulsiveness can explain behavior in contests better than other personal characteristics. From the dual processes model perspective (Sloman, 1996), it is important to know why so many players use the impulsive type I rather than the deliberative type II system. We show that the answer may lie in the contest environment, which makes it difficult to utilize the type II system. We find that theories based on non-standard preferences and probability weighting can explain differences between treatments in which learning is difficult. When learning is made easier, the effect of non-standard preferences disappears, and players converge to the risk-neutral Nash equilibrium.

Variability in the quality of feedback can organize our data, but it can also organize data from other studies that manipulated payoff risk in rent-seeking contests. Chowdhury et al. (2014) find high Nash equilibrium rates only in the treatment with no payoff risk and quadratic investment costs. Quadratic costs increase the penalty of non-equilibrium choices; as a result, dominated actions typically provide low payoffs, and payoff-based learning would predict faster convergence. Masiliūnas et al. (2014) find that choices are close to theoretical predictions only when payoff risk is removed and opponents play a fixed action over time, manipulations that would dramatically increase the speed of payoff-based learning. Fallucchi et al. (2013) find increased Nash equilibrium rates only in the treatment with no payoff risk and no feedback on individual choices. Feedback about the choice and payoff of each opponent may nudge players to imitate the best, crowding out the forms of learning that would converge to the Nash equilibrium. We also find evidence for imitation at the start of the game, but its effect decreases in all treatments when foregone payoff information is introduced.

Important conclusions can also be drawn from the treatment differences that were not observed. We find that behavior in the regular rent-seeking contest is not affected by foregone payoff information. Throughout the experiment, participants had access to 200 payoff realizations, but the distribution of choices at the end of the game was nearly identical to the initial distribution. This finding suggests that deviations from theoretical predictions in probabilistic rent-seeking contests are unlikely to disappear when the number of repetitions is increased. We also found no difference between the treatments in the first ten rounds, when foregone payoff information was unavailable. Therefore, over-investment persists and is robust to changes in payoff risk in the absence of foregone payoff information. This result is important because, in practice, contests differ in the degree and nature of payoff risk. Originally, the rent-seeking contest was used to study competition for monopoly rents (Tullock, 1980). In such winner-take-all markets, the assumption of probabilistic prize allocation is reasonable because only one product can become the industry standard (e.g., Blu-ray or HD DVD). In other markets, profits are divided between competitors; for example, advertising and technical improvements may increase the market share, but not guarantee the entire market. Similarly, resources not spent on competition are rarely kept as cash but instead invested in other risky projects. Risk in these projects originates not from the uncertainty about the behavior of the competitors but from the potential external technological or legal challenges. Our paper

studies such variations in payoff risk and finds similar behavioral patterns in all treatments (if foregone payoffs are not displayed). This result gives confidence that the numerous results obtained using the rent-seeking contest success function can be extended to a more general framework.

Our results might explain why convergence to the Nash equilibrium is often observed in continuous-time experiments (Brown and Stephenson, 2020). First, in these experiments, participants usually receive the average payoff from being matched with all other participants ("mean-matching"), which reduces the payoff risk and the variability of opponent's actions over time, since the average action of many opponents is more stable than the action of a single opponent. Second, participants in these experiments also usually receive salient feedback about their own payoffs or the choices and payoffs of all other participants (Cason et al., 2014, Stephenson and Brown, 2021). As a result, these studies create an environment with little payoff risk, stable choices of the opponents and clear feedback on how to improve own earnings, in a similar way as our treatment with information about foregone payoffs and no payoff risk. Our results suggest that both mean-matching and clear feedback might be necessary to observe convergence in continuous-time games, and these results could be explained by payoff-based learning. In future research, it would be interesting to study continuous-time rent-seeking contests with mean-matching and clear feedback. Given the findings from the previous literature and from our study, we would expect significantly less over-investment in this environment, compared to the discrete-time design.

Our approach of reducing payoff noise and improving feedback might be applied to first-price auctions, where overbidding is commonly observed. In first-price auctions, paying the expected value of the gamble would not be possible because payoffs are not stochastic. Still, the complexity that arises from the strategic uncertainty about the opponent's type and bid could be reduced by matching participants with multiple other bidders and paying the average earnings across all pairings. Better still, participants could be informed about their average earnings but paid based on one randomly selected pairing to identify the pure effect of the additional information. Kirchkamp et al. (2010) used a similar design in first-price and second-price auctions, where the private values were drawn from a uniform distribution and decisions were made using a strategy method. In the control treatment, payoffs were determined by a single draw. In other treatments, fifty draws are made, and participants are informed about the earnings from all the fifty auctions. Their earnings were decided by all fifty auctions or one randomly selected auction. Therefore, this design disentangles the effect of extensive feedback from risk reduction. Authors find that risk reduction significantly reduces overbidding, but extensive feedback does not. The insignificant effect of feedback is consistent with our results because participants in Kirchkamp et al. (2010) were not informed about their foregone payoffs. We are unaware of any auction experiments that supply foregone payoff information, and the intervention could be designed the same way as in our study. An alternative is to inform players about the choices and earnings of other players, although such information should facilitate imitate-the-best rather than best-response dynamics. Nevertheless, Stephenson and Brown (2020) found that average bids were closer to the equilibrium prediction when such information was provided in all-pay auctions. Similarly, Ockenfels and Selten (2005) found that when winners were informed about the loser's bid in a two-person first-price auction, average bids were lower and closer to the equilibrium prediction, compared to the treatment with no additional feedback. Engelbrecht-Wiggans and Katok (2008) manipulated information about loser's regret (additional money earned by outbidding the highest bidder) and winner's regret (money saved by reducing the bid to the second-highest

bidder) in first-price auctions with one human participant and two computerized opponents. In addition, each decision was matched with ten pairs of computerized opponents and regret was calculated across all ten auctions. As predicted, the information about loser's regret increased bids, and winner's regret reduced them, compared to the no-feedback treatment. Finally, the closest study to ours in the auction setting is Stephenson and Brown (2021), who study all-pay auctions in continuous time. The experiment used mean-matching and participants were in some treatments informed about the instantaneous payoffs from all actions. The study found no overbidding in terms of average bids, consistent with our expectations. It would be interesting to replicate these results in a discrete-time first-price auctions and manipulate both mean-matching and payoff information to identify if they are necessary for convergence.

Standard solution concepts assume that players are fully informed about the incentive structure. However, in complex games, boundedly rational participants may learn little from the game description and instead rely on information acquired from experience. In such games, models that assume no prior information about the game structure may have higher explanatory power than those assuming full information. We show how insights from individual choice tasks in which participants have little information about the payoff structure can be used to improve the explanatory power of solution concepts in strategic situations. We hope that these results will make a step towards developing a successful positive model of human behavior.

Our study contributes to the question of how the informational value of received payoff information influences the explanatory power of solution concepts in repeated games. Bereby-Meyer and Roth (2006) show that noisy payoffs slow down learning in prisoner's dilemma, and Shafran (2012) shows that stochastic payoffs slow down convergence in coordination games. We show that reduced noise dramatically improves Nash equilibrium play in rent-seeking contests. It is still to be tested if these findings can be extended to other games and if heterogeneity in the informational value of feedback can explain why convergence to equilibrium is observed in some games (e.g., competitive guessing game, Weber, 2003, race game, Gneezy et al., 2010) but not in others.

References

Armantier, O. (2004). Does observation influence learning? *Games and Economic Behavior*, 46(2):221–239.

Baharad, E. and Nitzan, S. (2008). Contest efforts in light of behavioural considerations. *The Economic Journal*, 118(533):2047–2059.

Beggs, A. W. (2005). On the convergence of reinforcement learning. *Journal of Economic Theory*, 122(1):1–36.

Bereby-Meyer, Y. and Roth, A. E. (2006). The speed of learning in noisy games: Partial reinforcement and the sustainability of cooperation. *American Economic Review*, 96(4):1029–1042.

Blume, A., DeJong, D. V., Neumann, G. R., and Savin, N. (2002). Learning and communication in sender-receiver games: an econometric investigation. *Journal of Applied Econometrics*, 17(3):225–247.

Bosch-Domènech, A. and Vriend, N. J. (2003). Imitation of successful behaviour in cournot markets. *The Economic Journal*, 113(487):495–524.

Bravo, M. and Mertikopoulos, P. (2017). On the robustness of learning in games with stochastically perturbed payoff observations. *Games and Economic Behavior*, 103:41–66.

Brookins, P. and Ryvkin, D. (2014). An experimental study of bidding in contests of incomplete information. *Experimental Economics*, 17(2):245–261.

Brown, A. L. and Stephenson, D. G. (2020). Games with continuous-time experimental protocols. In *Handbook of Experimental Game Theory*. Edward Elgar Publishing.

Camerer, C. and Ho, T.-H. (1999). Experience-weighted attraction learning in normal form games. *Econometrica*, 67(4):827–874.

Camerer, C. F. (2003). *Behavioral game theory*. Russell Sage Foundation New York.

Cason, T. N., Friedman, D., and Hopkins, E. (2014). Cycles and instability in a rock–paper–scissors population game: a continuous time experiment. *Review of Economic Studies*, 81(1):112–136.

Cason, T. N., Masters, W. A., and Sheremeta, R. M. (2020). Winner-take-all and proportional-prize contests: theory and experimental results. *Journal of Economic Behavior & Organization*, 175:314–327.

Charness, G., Frechette, G. R., and Kagel, J. H. (2004). How robust is laboratory gift exchange? *Experimental Economics*, 7(2):189–205.

Chen, W. J. and Krajbich, I. (2017). Computational modeling of epiphany learning. *Proceedings of the National Academy of Sciences*, 114(18):4637–4642.

Cheung, Y.-W. and Friedman, D. (1997). Individual learning in normal form games: Some laboratory results. *Games and Economic Behavior*, 19(1):46–76.

Cho, I.-K. and Matsui, A. (2005). Learning aspiration in repeated games. *Journal of Economic Theory*, 124(2):171–201.

Chowdhury, S. M., Mukherjee, A., and Turocy, T. L. (2020). That's the ticket: explicit lottery randomisation and learning in Tullock contests. *Theory and Decision*, 88:405–429.

Chowdhury, S. M., Sheremeta, R. M., and Turocy, T. L. (2014). Overbidding and overspreading in rent-seeking experiments: Cost structure and prize allocation rules. *Games and Economic Behavior*, 87:224–238.

Chung, K., Kim, K., and Lim, N. (2020). Social structures and reputation in expert review systems. *Management Science*, 66(7):3249–3276.

Cornes, R. and Hartley, R. (2003). Risk aversion, heterogeneity and contests. *Public Choice*, 117(1):1–25.

Cornes, R., Hartley, R., et al. (2003). Loss aversion and the Tullock paradox. Working paper, University of Nottingham Discussion Paper No. 03/17.

Cox, C. A. (2017). Rent-seeking and competitive preferences. *Journal of Economic Psychology*, 63:102–116.

Cox, J. C., Smith, V. L., and Walker, J. M. (1988). Theory and individual behavior of first-price auctions. *Journal of Risk and Uncertainty*, 1(1):61–99.

DellaVigna, S. (2009). Psychology and economics: Evidence from the field. *Journal of Economic literature*, 47(2):315–72.

Dickinson, D. L. and McEvoy, D. M. (2021). Further from the truth: The impact of moving from in-person to online settings on dishonest behavior. *Journal of Behavioral and Experimental Economics*, 90:101649.

Duch, M. L., Grossmann, M. R., and Lauer, T. (2020). z-Tree unleashed: A novel client-integrating architecture for conducting z-Tree experiments over the internet. *Journal of Behavioral and Experimental Finance*, 28:100400.

Duffy, J. and Hopkins, E. (2005). Learning, information, and sorting in market entry games: theory and evidence. *Games and Economic Behavior*, 51(1):31–62.

Dufwenberg, M., Sundaram, R., and Butler, D. J. (2010). Epiphany in the game of 21. *Journal of Economic Behavior & Organization*, 75(2):132–143.

Engelbrecht-Wiggans, R. and Katok, E. (2008). Regret and feedback information in first-price sealed-bid auctions. *Management Science*, 54(4):808–819.

Engelbrecht-Wiggans, R. and Katok, E. (2009). A direct test of risk aversion and regret in first price sealed-bid auctions. *Decision Analysis*, 6(2):75–86.

Erev, I. and Barron, G. (2005). On adaptation, maximization, and reinforcement learning among cognitive strategies. *Psychological review*, 112(4):912.

Erev, I. and Haruvy, E. (2016). Learning and the economics of small decisions. In Kagel, J. H. and Roth, A. E., editors, *The Handbook of Experimental Economics, Volume 2*, chapter 10, pages 638–716. Princeton university press, Princeton, NJ.

Erev, I. and Roth, A. E. (1998). Predicting how people play games: Reinforcement learning in experimental games with unique, mixed strategy equilibria. *American Economic Review*, 88(4):848–881.

Ert, E. and Erev, I. (2007). Replicated alternatives and the role of confusion, chasing, and regret in decisions from experience. *Journal of Behavioral Decision Making*, 20(3):305–322.

Fallucchi, F., Renner, E., and Sefton, M. (2013). Information feedback and contest structure in rent-seeking games. *European Economic Review*, 64:223–240.

Fehr, E. and Schmidt, K. M. (1999). A theory of fairness, competition, and cooperation. *The Quarterly Journal of Economics*, 114(3):817–868.

Feltovich, N. (2000). Reinforcement-based vs. belief-based learning models in experimental asymmetric-information games. *Econometrica*, 68(3):605–641.

Filiz-Ozbay, E. and Ozbay, E. Y. (2007). Auctions with anticipated regret: Theory and experiment. *American Economic Review*, 97(4):1407–1418.

Fischbacher, U. (2007). z-Tree: Zurich toolbox for ready-made economic experiments. *Experimental economics*, 10(2):171–178.

Fonseca, M. A. (2009). An experimental investigation of asymmetric contests. *International Journal of Industrial Organization*, 27(5):582–591.

Foster, D. P. and Young, H. P. (2006). Regret testing: Learning to play Nash equilibrium without knowing you have an opponent. *Theoretical Economics*, 1(3):341–367.

Fudenberg, D. and Peysakhovich, A. (2016). Recency, records, and recaps: Learning and nonequilibrium behavior in a simple decision problem. *ACM Transactions on Economics and Computation (TEAC)*, 4(4):1–18.

Gneezy, U., Rustichini, A., and Vostroknutov, A. (2010). Experience and insight in the race game. *Journal of Economic Behavior & Organization*, 75(2):144–155.

Goeree, J. K. and Holt, C. A. (2001). Ten little treasures of game theory and ten intuitive contradictions. *American Economic Review*, 91(5):1402–1422.

Goeree, J. K., Holt, C. A., and Palfrey, T. R. (2002). Quantal response equilibrium and overbidding in private-value auctions. *Journal of Economic Theory*, 104(1):247–272.

Gonzalez, R. and Wu, G. (1999). On the shape of the probability weighting function. *Cognitive psychology*, 38(1):129–166.

Greiner, B. (2015). Subject pool recruitment procedures: organizing experiments with ORSEE. *Journal of the Economic Science Association*, 1(1):114–125.

Grosskopf, B., Bereby-Meyer, Y., and Bazerman, M. (2007). On the robustness of the winner's curse phenomenon. *Theory and Decision*, 63(4):389–418.

Grosskopf, B., Erev, I., and Yechiam, E. (2006). Foregone with the wind: Indirect payoff information and its implications for choice. *International Journal of Game Theory*, 34(2):285–302.

Grund, C. and Sliwka, D. (2005). Envy and compassion in tournaments. *Journal of Economics & Management Strategy*, 14(1):187–207.

Hasselt, H. (2010). Double Q-learning. *Advances in neural information processing systems*, 23:2613–2621.

Herrmann, B. and Orzen, H. (2008). The appearance of homo rivalis: Social preferences and the nature of rent seeking. Working paper, CeDEx discussion paper series.

Hertwig, R., Barron, G., Weber, E. U., and Erev, I. (2004). Decisions from experience and the effect of rare events in risky choice. *Psychological science*, 15(8):534–539.

Hillman, A. L. and Katz, E. (1984). Risk-averse rent seekers and the social cost of monopoly power. *The Economic Journal*, 94(373):104–110.

Ho, T. H., Wang, X., and Camerer, C. F. (2008). Individual differences in EWA learning with partial payoff information. *The Economic Journal*, 118(525):37–59.

Hoffmann, M. and Kolmar, M. (2017). Distributional preferences in probabilistic and share contests. *Journal of Economic Behavior & Organization*, 142:120–139.

Hopkins, E. (2002). Two competing models of how people learn in games. *Econometrica*, 70(6):2141–2166.

Huck, S., Normann, H.-T., and Oechssler, J. (1999). Learning in cournot oligopoly–an experiment. *The Economic Journal*, 109(454):80–95.

Huttegger, S. M. (2013). Probe and adjust. *Biological Theory*, 8(2):195–200.

Jindapon, P. and Whaley, C. A. (2015). Risk lovers and the rent over-investment puzzle. *Public Choice*, 164(1-2):87–101.

Jindapon, P. and Yang, Z. (2017). Risk attitudes and heterogeneity in simultaneous and sequential contests. *Journal of Economic Behavior & Organization*, 138:69–84.

Kahneman, D. and Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, 47(2):263–292.

Kephart, C. and Friedman, D. (2015). Hotelling revisits the lab: equilibration in continuous and discrete time. *Journal of the Economic Science Association*, 1(2):132–145.

Kirchkamp, O., Reiss, J. P., and Sadrieh, A. (2010). A pure variation of risk in private-value auctions. Working paper, METEOR Research Memorandum No. 050.

Kong, X. (2008). Loss aversion and rent-seeking: An experimental study. Technical report, CeDEx discussion paper series.

Lehrer, E. and Kalai, E. (1993). Rational learning leads to Nash equilibrium. *Econometrica*, 61(5):1019–1045.

Li, J., Leider, S., Beil, D., and Duenyas, I. (2021). Running online experiments using web-conferencing software. *Journal of the Economic Science Association*, 7(2):167–183.

Li, S. (2017). Obviously strategy-proof mechanisms. *American Economic Review*, 107(11):3257–87.

Mago, S. D., Samak, A. C., and Sheremeta, R. M. (2016). Facing your opponents: Social identification and information feedback in contests. *Journal of Conflict Resolution*, 60(3):459–481.

Masiliūnas, A., Mengel, F., and Reiss, J. P. (2014). Behavioral variation in Tullock contests. Working paper, Working Paper Series in Economics, Karlsruher Institut für Technologie (KIT).

McKinney, Jr, C. N. and Van Huyck, J. B. (2013). Eureka learning: Heuristics and response time in perfect information games. *Games and Economic Behavior*, 79:223–232.

Millner, E. L. and Pratt, M. D. (1991). Risk aversion and rent-seeking: An extension and some experimental evidence. *Public Choice*, 69(1):81–92.

Myers, J. L. and Sadler, E. (1960). Effects of range of payoffs as a variable in risk taking. *Journal of Experimental Psychology*, 60(5):306.

Nax, H. H., Burton-Chellew, M. N., West, S. A., and Young, H. P. (2016). Learning in a black box. *Journal of Economic Behavior & Organization*, 127:1–15.

Ockenfels, A. and Selten, R. (2005). Impulse balance equilibrium and feedback in first price auctions. *Games and Economic Behavior*, 51(1):155–170.

Ockenfels, A. and Selten, R. (2014). Impulse balance in the newsvendor game. *Games and Economic Behavior*, 86:237–247.

Offerman, T., Potters, J., and Sonnemans, J. (2002). Imitation and belief learning in an oligopoly experiment. *The Review of Economic Studies*, 69(4):973–997.

Oprea, R., Henwood, K., and Friedman, D. (2011). Separating the hawks from the doves: Evidence from continuous time laboratory games. *Journal of Economic Theory*, 146(6):2206–2225.

Otto, A. R. and Love, B. C. (2010). You don't want to know what you're missing: When information about forgone rewards impedes dynamic decision making. *Judgment and Decision Making*, 5(1):1–10.

Parco, J. E., Rapoport, A., and Amaldoss, W. (2005). Two-stage contests with budget constraints: An experimental study. *Journal of Mathematical Psychology*, 49(4):320–338.

Price, C. R. and Sheremeta, R. M. (2011). Endowment effects in contests. *Economics Letters*, 111(3):217–219.

Price, C. R. and Sheremeta, R. M. (2015). Endowment origin, demographic effects, and individual preferences in contests. *Journal of Economics & Management Strategy*, 24(3):597–619.

Rakow, T., Newell, B. R., and Wright, L. (2015). Forgone but not forgotten: the effects of partial and full feedback in "harsh" and "kind" environments. *Psychonomic Bulletin & Review*, 22(6):1807–1813.

Riechmann, T. (2007). An analysis of rent-seeking games with relative-payoff maximizers. *Public Choice*, 133(1):147–155.

Roth, A. E. and Erev, I. (1995). Learning in extensive-form games: Experimental data and simple dynamic models in the intermediate term. *Games and Economic Behavior*, 8(1):164–212.

Salmon, T. C. (2001). An evaluation of econometric models of adaptive learning. *Econometrica*, 69(6):1597–1628.

Savikhin, A. C. and Sheremeta, R. M. (2013). Simultaneous decision-making in competitive and cooperative environments. *Economic Inquiry*, 51(2):1311–1323.

Schmitt, P., Shupp, R., Swope, K., and Cadigan, J. (2004). Multi-period rent-seeking contests with carryover: Theory and experimental evidence. *Economics of Governance*, 5(3):187–211.

Selten, R., Abbink, K., and Cox, R. (2005). Learning direction theory and the winner's curse. *Experimental Economics*, 8(1):5–20.

Selten, R. and Chmura, T. (2008). Stationary concepts for experimental 2x2-games. *American Economic Review*, 98(3):938–66.

Selten, R. and Stoecker, R. (1986). End behavior in sequences of finite prisoner's dilemma supergames a learning theory approach. *Journal of Economic Behavior & Organization*, 7(1):47–70.

Shaffer, S. (2006). Contests with interdependent preferences. *Applied Economics Letters*, 13(13):877–880.

Shafran, A. P. (2012). Learning in games with risky payoffs. *Games and Economic Behavior*, 75(1):354–371.

Sheremeta, R. M. (2010). Experimental comparison of multi-stage and one-stage contests. *Games and Economic Behavior*, 68(2):731–747.

Sheremeta, R. M. (2011). Contest design: An experimental investigation. *Economic Inquiry*, 49(2):573–590.

Sheremeta, R. M. (2013). Overbidding and heterogeneous behavior in contest experiments. *Journal of Economic Surveys*, 27(3):491–514.

Sheremeta, R. M. (2018). Impulsive behavior in competition: Testing theories of overbidding in rent-seeking contests. *Available at SSRN: https://ssrn.com/abstract=2676419.*

Sheremeta, R. M. and Zhang, J. (2010). Can groups solve the problem of over-bidding in contests? *Social Choice and Welfare*, 35(2):175–197.

Shupp, R., Sheremeta, R. M., Schmidt, D., and Walker, J. (2013). Resource allocation contests: Experimental evidence. *Journal of Economic Psychology*, 39:257–267.

Skaperdas, S. and Gan, L. (1995). Risk aversion in contests. *Economic Journal*, 105(431):951–962.

Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin*, 119(1):3.

Stephenson, D. (2019). Coordination and evolutionary dynamics: When are evolutionary models reliable? *Games and Economic Behavior*, 113:381–395.

Stephenson, D. G. and Brown, A. L. (2020). Characterizing persistent disequilibrium dynamics: Imitation or optimization. Working paper.

Stephenson, D. G. and Brown, A. L. (2021). Playing the field in all-pay auctions. *Experimental Economics*, 24(2):489–514.

Thrun, S. and Schwartz, A. (1993). Issues in using function approximation for reinforcement learning. In *Proceedings of the 1993 Connectionist Models Summer School Hillsdale, NJ. Lawrence Erlbaum*.

Tullock, G. (1980). Efficient rent seeking. *Toward a theory of the rent-seeking society*, 97:112.

Tversky, A. and Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, 5(4):297–323.

Weber, R. A. (2003). 'Learning' with no feedback in a competitive guessing game. *Games and Economic Behavior*, 44(1):134–144.

Weibull, J. W. (1997). *Evolutionary game theory*. MIT press.

Wilcox, N. T. (2006). Theories of learning in games and heterogeneity bias. *Econometrica*, 74(5):1271–1292.

Yechiam, E. and Busemeyer, J. R. (2006). The effect of foregone payoffs on underweighting small probability events. *Journal of Behavioral Decision Making*, 19(1):1–16.

Yechiam, E. and Rakow, T. (2012). The effect of foregone outcomes on choices from experience. *Experimental Psychology*, 59(2):55–67.

Young, H. P. (2009). Learning by trial and error. *Games and Economic Behavior*, 65(2):626–643.

Appendix

A Correlation between expected and accumulated earnings

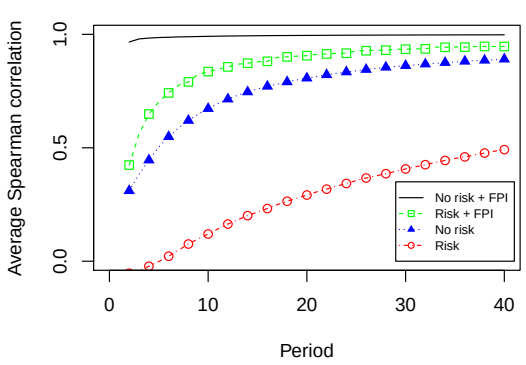
We ran a simulation to see how long players would need to play the rent-seeking contest to discover the payoff-maximizing action, if only learning from experience was possible. To simplify, assume that contest is played as a multi-armed bandit task, with each of the nine arms representing a contest investment level (from 0 to 8, as in our Study 1). When an arm is played, opponent's action is drawn from a stationary distribution (equal to the distribution of choices in our baseline experiment) and a profit is generated using a contest success function (with or without payoff risk).[26] Each round, one arm is chosen at random (uniformly) and the payoff is observed. The procedure is performed under one of the four conditions, corresponding to our experimental design: payoff risk is either present or not and FPI is either observed or not. For each condition, we calculate the average observed payoff that each arm generated over all previous rounds.[27] We then measure how well the average observed payoff reflects the expected payoff of that arm. First, we use Spearman's correlation coefficient to measure how well the ranking of arms in terms of average observed payoffs recovers the ranking based on expected payoffs. Panel (a) of Figure A.1 shows the average correlation from 10 000 simulations. In the regular contest, the correlation is very low and increases only to 0.5 after 40 rounds of experience. Correlation is increased if payoff risk is removed or if FPI is provided, and both manipulations combined produce a nearly perfect positive correlation after just one round. Second, we calculate how often the arm that maximizes average observed payoffs coincides with the arm that maximizes objective expected payoffs. This measure can be interpreted as the likelihood that a Bayesian player will correctly identify the expected payoff-maximizing action using all previous payoff realizations. Panel (b) in Figure A.1 shows that expected payoff maximization fails at least 50% of the time if either payoff risk is present or forgone payoffs are not observed, even after 40 rounds. In contrast, the maximization rate approaches 100% with no payoff risk and with FPI. In terms of both correlation and payoff maximization, feedback from one round with FPI and no payoff risk improves the decision quality more than 40 rounds of experience in either of the other three conditions. We ran additional simulations to see how soon the payoff maximization rate reaches 73%, the rate achieved in just one round with no payoff risk and with FPI. It takes about 500 rounds with payoff risk and FPI, 1500 rounds without payoff risk and without FPI and 15 000 rounds with payoff risk and with FPI.

Overall, the simulation shows that realized payoffs in contests are a noisy measure of expected payoffs,[28] and players who wanted to maximize expected payoffs would fail to do so if decisions were made only from experience.

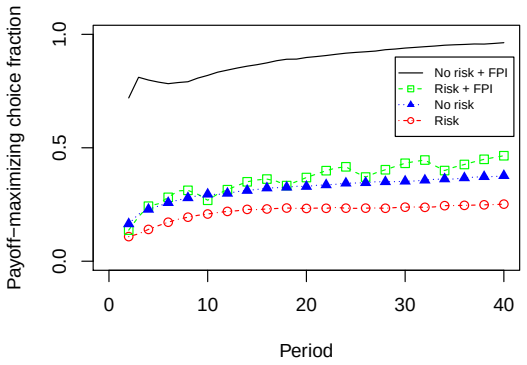
[26] A game with such structure has been implemented by Cox (2017), finding no difference at the aggregate level between this "robot" and standard "human" treatment.

[27] If no payoff is observed, we use the average expected payoff from the task, equal to 8.02.

[28] In the simulation, expected payoffs can be calculated because actions of opponents are drawn from a stationary distribution. In strategic games, expected payoffs could be calculated ex-post. Models such as reinforcement learning converge through iterated elimination of dominated strategies, therefore their speed depends on how accurately the realized payoff differences reflect differences in terms of expected payoffs.



(a) Average correlation between expected payoffs and average observed payoffs over prior rounds



(b) Expected payoff maximization, if choice is based on highest observed average payoff.

Figure A.1: Simulated recovery of expected payoffs from observed payoffs in a bandit task

B Nash equilibrium with preferences and probability weighting

RNNE is calculated assuming that expected utility is equal to the expected payoff and subjective probabilities are equal to objective probabilities. Here, we show how the predictions change when we relax these assumptions an allow other-regarding or risk preferences, non-monetary utility from winning and probability weighting.

Preferences are incorporated into the expected utility function by defining the lottery outcome $l \in \{1, \ldots, L\}$ for player i as $(\pi_i^l, \pi_j^l, W_i^l)$, where π_i^l is the payoff received by i, π_j^l is the payoff received by opponent j and W_i^l is the number of prizes that i has won in that round. Information about the opponent's payoff and the number of prizes won is required to calculate other-regarding preferences and non-monetary utility of winning.

Table B.1: A list of possible outcomes in each treatment

Treatment	**Contest**		**Non-contest**		**Payoff**		**Probability**
	i	j	i	j	i	j	
RR	W	L	W	W	$2V$	V	$p_i^c p_i^{nc} p_j^{nc}$
	W	L	W	L	$2V$	0	$p_i^c p_i^{nc}(1-p_j^{nc})$
	W	L	L	W	V	V	$p_i^c(1-p_i^{nc})p_j^{nc}$
	W	L	L	L	V	0	$p_i^c(1-p_i^{nc})(1-p_j^{nc})$
	L	W	W	W	V	$2V$	$(1-p_i^c)p_i^{nc}p_j^{nc}$
	L	W	W	L	V	V	$(1-p_i^c)p_i^{nc}(1-p_j^{nc})$
	L	W	L	W	0	$2V$	$(1-p_i^c)(1-p_i^{nc})p_j^{nc}$
	L	W	L	L	0	V	$(1-p_i^c)(1-p_i^{nc})(1-p_j^{nc})$
SR	W	L	–	–	$V+E-c_i$	$E-c_j$	p_i^c
	L	W	–	–	$E-c_i$	$V+E-c_j$	$1-p_i^c$
RS	–	–	W	W	$V+\frac{c_i}{c_i+c_j}V$	$V+\frac{c_j}{c_i+c_j}V$	$p_i^{nc}p_j^{nc}$
	–	–	W	L	$V+\frac{c_i}{c_i+c_j}V$	$\frac{c_j}{c_i+c_j}V$	$p_i^{nc}(1-p_j^{nc})$
	–	–	L	W	$\frac{c_i}{c_i+c_j}V$	$V+\frac{c_j}{c_i+c_j}V$	$(1-p_i^{nc})p_j^{nc}$
	–	–	L	L	$\frac{c_i}{c_i+c_j}V$	$\frac{c_j}{c_i+c_j}V$	$(1-p_i^{nc})(1-p_j^{nc})$
SS	–	–	–	–	$E-c_i+\frac{c_i}{c_i+c_j}V$	$E-c_j+\frac{c_j}{c_i+c_j}V$	1

Table B.1 lists all possible outcomes in each treatment. For each outcome, we specify (i) the payoffs obtained by player i and opponent j, (ii) whether each player won (W) or lost (L) the reward from the contest and non-contest lotteries and (iii) the probability of the outcome.

The expected utility of player i is calculated as a weighted average of utilities at each outcome. Utilities are weighted using function $w(p_l)$, where p_l is the objective probability that outcome l will occur.

$$E[u_i] = \sum_{l=1}^{L} w(p_l) u_i(\pi_i^l, \pi_j^l, W_i^l)$$

With standard preferences, $w(p_l) = p_l$ and $u_i(\pi_i^l, \pi_j^l, W_i^l) = \pi_i^l$, reducing expected utility to expected payoff.

Next, we specify the utility functions used to calculate utility at each outcome.

B.1 Risk preferences

Previous literature found that the failure of the risk neutrality assumption could explain the difference between theoretical predictions and experimental data in related games, such as auctions (Goeree et al., 2002, Cox et al., 1988). Contest investment has elements of both gambling and insurance, therefore it is not possible to identify the effect of risk preferences without assuming a specific utility function, usually with constant absolute risk aversion (CARA) or constant relative risk aversion (CRRA). Hillman and Katz (1984) show that risk aversion reduces investment if all participants use a logarithmic utility function (which exhibits CARA). Skaperdas and Gan (1995) assume exponential utility functions (which also exhibits CARA) and find that the more risk averse agent invests less than the opponent as long as both agents are sufficiently similar in terms of their risk preferences. Cornes and Hartley (2003) show that under exponential utility, total investment is always lower when all agents are either risk averse or risk neutral, compared to full risk neutrality. Jindapon and Yang (2017) show that with a generalized CARA utility function both risk averse and risk seeking players invest below the risk-neutral Nash equilibrium prediction in symmetric simultaneous two-player contests. Jindapon and Whaley (2015) use both exponential (CARA) and CRRA utility functions that exhibit risk-seeking preferences and show that in two-player contests with identical preferences imprudence increases over-investment, irrespective of the risk-aversion coefficient. Another strand of literature tests whether heterogeneity in choices can be explained by heterogeneity in risk preferences, elicited in experiments. It is typically found that risk averse participants on average invest less than risk seeking participants (Millner and Pratt, 1991, Herrmann and Orzen, 2008, Sheremeta and Zhang, 2010, Sheremeta, 2011, Price and Sheremeta, 2015).

We look at the predictions of a Nash equilibrium under the assumption of constant relative risk aversion (CRRA) or constant absolute risk aversion (CARA).

CRRA is measured using parameter r, which indicates risk aversion if $r > 0$ and risk seeking preferences if $r < 0$:

$$u_i(\pi_i, \pi_j, W_i) = \frac{\pi_i^{1-r}}{1-r} \tag{13}$$

We model CARA using an exponential function, in which parameter a determines whether preferences are risk seeking ($a < 0$), risk neutral ($a = 0$) or risk averse ($a > 0$):

$$u_i(\pi_i, \pi_j, W_i) = \begin{cases} \pi_i & \text{if } a = 0 \\ (1 - e^{-a\pi_i})/a & \text{otherwise} \end{cases} \tag{14}$$

B.2 Other-regarding preferences

Over-investment could be explained by spiteful preferences (Riechmann, 2007), aversion to disadvantageous inequality, or a preference for advantageous inequality (Herrmann and Orzen, 2008, Fonseca, 2009). Some studies elicit spitefulness or inequality aversion and correlate it with behavior in contests. Savikhin and Sheremeta (2013) find that participants who contribute more in a public goods game invest less in the contest. Herrmann and Orzen (2008) measure pro-sociality using a prisoner's dilemma and find the opposite result: pro-sociality is correlated with higher contest investment. An alternative to measuring social preferences is to eliminate their effect by removing the payoff consequences to the opponent. Herrmann and Orzen (2008) use a strategy method and find higher investment when playing against other participants compared to playing against no competitor. Masiliūnas et al. (2014) and Cox (2017) match players to computers that are programmed to play choices made by participants in the standard contest experiment. Both studies find no significant difference between computer and human treatments.

Players with other-regarding preferences could care about the distribution of expected payoffs (Hoffmann and Kolmar, 2017) or about realized payoffs (Herrmann and Orzen, 2008, Fonseca, 2009). Some studies assume preferences over the distribution of total earnings (Herrmann and Orzen, 2008, Fonseca, 2009, Mago et al., 2016 Shaffer, 2006), while others assume preferences only over the distribution of lottery earnings, net of investment costs (Grund and Sliwka, 2005, Hoffmann and Kolmar, 2017). We assume preferences over ex-post outcomes, including earnings from all sources. If we instead assumed preferences over ex-ante payoffs, predictions in all treatments would coincide with those in SS.

First, we model other-regarding preferences by assuming that in addition to own payoff, players care about opponent's payoff, weighted by s: preferences are altruistic if $s > 0$ and spiteful if $s < 0$.

$$u_i(\pi_i, \pi_j, W_i) = \pi_i + s\pi_j \tag{15}$$

Second, we model inequality aversion following Fehr and Schmidt (1999). The model has two parameters, α and β, which measure the weight of disadvantageous and advantageous inequality:

$$u_i(\pi_i, \pi_j, W_i) = \begin{cases} \pi_i - \alpha(\pi_j - \pi_i) & \text{if } \pi_i < \pi_j \\ \pi_i - \beta(\pi_i - \pi_j) & \text{if } \pi_i > \pi_j \end{cases}$$

B.3 Non-monetary utility of winning

It has been proposed that excessive investment in rent-seeking contests is driven by the non-monetary utility that participants receive from winning the contest (Schmitt et al., 2004). Sheremeta (2010) measures non-monetary utility of winning by asking players to compete for

a prize of zero value. It is found that investment in zero-value contests is correlated with investment in rent-seeking contests (Price and Sheremeta, 2011, Price and Sheremeta, 2015, Brookins and Ryvkin, 2014, Cason et al., 2020, Mago et al., 2016, Cox, 2017). Such correlation does not necessarily imply causality, as both could arise because of confusion, experimenter demand effect, mistrusting the experimenter or habit (see Sheremeta, 2010, and Sheremeta, 2013, for a discussion).

We follow Sheremeta (2010) and assume that players who win receive additional utility ω, in addition to the monetary prize value.

$$u_i(\pi_i, \pi_j, W_i) = \pi_i + \omega W_i \tag{16}$$

B.4 Non-linear probability weighting

It is well known that biases occur when choosing between risky gambles. In particular, it is often found that objective probabilities are weighted using an inverted S-shaped function (Tversky and Kahneman, 1992, Gonzalez and Wu, 1999). Such probability weighting could explain over-investment in contests with many players, but it predicts under-investment in two-player contests (Baharad and Nitzan, 2008) and fails to explain the data (Parco et al., 2005). Instead, over-investment could be rationalized by S-shaped probability weighting, which can originate from experience-based decision making (Hertwig et al., 2004).

We allow the weight placed on outcomes to differ from the objective probabilities. Specifically, we assume rank-dependent weighting as in the cumulative prospect theory (Tversky and Kahneman, 1992) with the following weighting function:

$$w(p) = \frac{p^\beta}{(p^\beta + (1-p)^\beta)^{1/\beta}} \tag{17}$$

Probability weighting is straightforward in SS (because there is only one outcome) and in SR (because there are two outcomes). In SR, the outcome in which the player receives the reward always provides a higher payoff, therefore it is weighted by $w(p_i^c)$, while the other outcome is weighted by $(1 - w(p_i^c))$. In RS, two pairs of outcomes are possible, but each pair provides identical payoffs for i. Cumulative prospect theory requires a strict order of outcomes, therefore all outcomes with identical outcomes must be combined. Unfortunately, outcomes cannot be easily combined because some outcomes will have identical payoffs but different utilities, because of social preferences. We therefore separately calculate the probability that each pair of outcomes will occur (p_i^{nc} and $(1 - p_i^{nc})$), weight these probabilities using $w(\cdot)$ and then divide the weights between outcomes in each pair proportionally to their (unweighted) probabilities. We could have weighted the probabilities within each pair too, but chose not to as it is unclear how probability weighting operates with the payoffs of the other participant. Our approach ensures that players with no social preferences always weight the outcomes exactly as in the cumulative prospect theory. We follow the same approach in RR, in which there are eight possible outcomes but only three distinct payoffs for i. We first calculate probabilities for each group of outcomes that generate these three payoff levels, and weight them. The outcome group with payoff $2V$ receives weight $w(p_i^c p_i^{nc})$. The outcome group with a payoff of V receives weight $w(p(2V) + p(V)) - w(p(2V)) = w(p_i^c p_i^{nc} + p_i^c(1 - p_i^{nc}) + (1 - p_i^c)p_i^{nc})) - w(p_i^c p_i^{nc})$. The outcome group with the payoff of 0 receives the remaining weight, ensuring that probabilities add up to 1. These weights are then divided within each outcome group proportionally to the (unweighted) probabilities of each individual outcome.

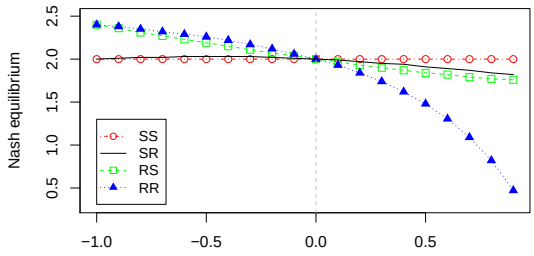
B.5 Impulse balance equilibrium with loss aversion

Impulse balance equilibrium (IBE) is based on ex-post rationality (Selten et al., 2005). IBE assumes that players receive an impulse to either decrease their action if it was above the ex-post rational action (downward impulse), or to increase it if the action was below the ex-post rational action (upward impulse). The size of an impulse is equal to the value of lost profit, multiplied by parameter θ if the profit falls below the reference level. It is commonly assumed that $\theta = 2$, in accordance with the prospect theory (Kahneman and Tversky, 1979), but we will permit a wider range of parameter values, including $\theta = 1$ (no loss aversion). IBE is defined as a point at which expected downward and upward impulses are equalized, therefore it is the rest point of dynamics modeled by the learning direction theory (Selten and Stoecker, 1986). The reliance on ex-post rationality makes IBE especially relevant for the current study because the treatment manipulations should change the likelihood of the two types of impulses and the ease with which the ex-post rational action can be identified. IBE has also been found to explain deviations from theoretical predictions in other settings, e.g., overbidding in auctions (Selten et al., 2005; Ockenfels and Selten, 2005) and "pull to the center" bias in the newsvendor game (Ockenfels and Selten, 2014), therefore it could potentially explain over-investment in contests. Since IBE has never been applied to rent-seeking contests, we derive the predictions for all the treatment variations in Appendix C.

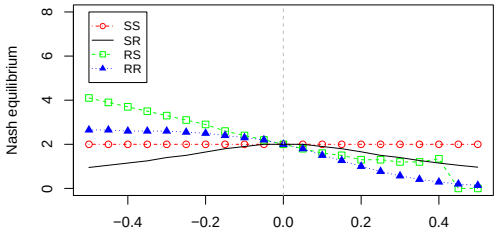
B.6 Predictions about treatment differences

Figure B.1 shows Nash equilibrium predictions for a range of parameter values.[29] Nash equilibrium with spitefulness ($s < 0$ in panel a) predicts over-investment in all treatments, and there is no predicted treatment difference for moderate levels of spitefulness. Risk-seeking preference ($r < 0$ in panel b) predicts very moderate levels of over-investment in SR, higher over-investment in RS and RR, and RNNE play in SS, where the risk is not present. Non-monetary utility from winning ($\omega > 0$ in panel c) predicts above-RNNE contest investment in SR, below-RNNE contest investment in RS and RNNE play in SS and RR. Inverse S-shaped probability weighting ($\beta < 1$ in panel d), commonly observed in previous studies, predicts below-RNNE contest investment. Over-investment in SR could be justified by S-shaped weighting function ($\beta > 1$ in panel d). This type of weighting predicts small over-investment in RR and under-investment in RS. Impulse balance equilibrium predicts over-investment in SR and under-investment in RS because the intensity of regret is highest when players could have received the prize by a slight increase in contest investment (in SR) or non-contest investment (in RS). This result holds for any value of θ, but higher loss aversion increases the intensity of impulses and the gap between SR and RS.

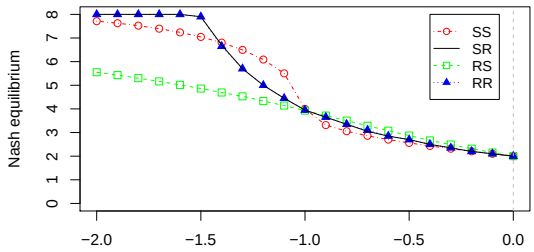
[29] All calculations are done for the parameters used in the experiment but allowing for a continuous strategy space. For the range of parameters that we consider, a pure strategy Nash equilibrium always exists. When multiple equilibria exist, we plot the average action in all equilibria.



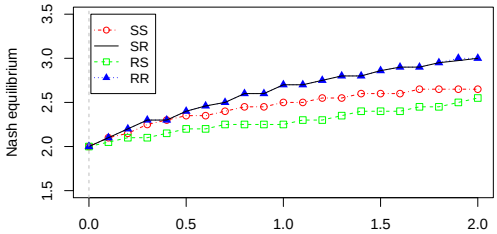
(a) CRRA, r



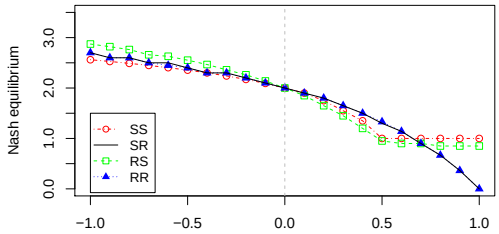
(b) CARA, a



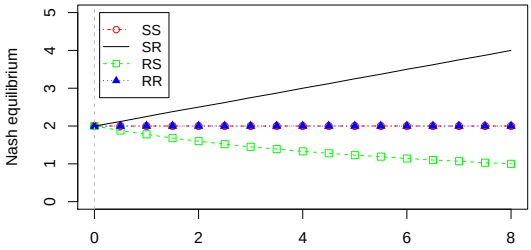
(c) Spitefulness, s



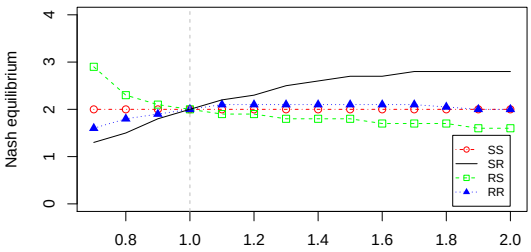
(d) Aversion to disadvantageous inequality, α



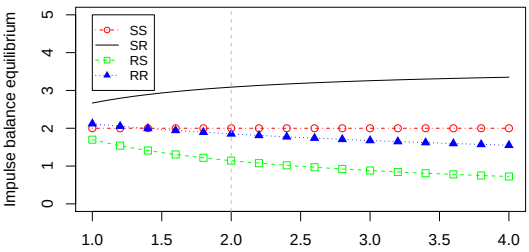
(e) Aversion to advantageous inequality, β



(f) Non-monetary utility of winning, ω



(g) Probability weighting, β



(h) Impulse balance equilibrium, θ

Figure B.1: Contest investment in Nash equilibrium and impulse balance equilibrium (panel h) for each treatment. Dashed gray line marks the standard parameter value.

C Impulse balance equilibrium

This appendix derives the predictions of the impulse balance equilibrium (IBE). Weighted IBE distinguishes between losses and gains, but since participants cannot lose money in our experiments, we need to make an assumption about the reference income level. We assume that the reference point is the initial endowment,[30] which also coincides with the maximin payoff, used as a reference point in IBE (e.g., Selten and Chmura, 2008). Whenever the realized payoff falls below the endowment, the strength of an impulse is multiplied by θ. In line with the prospect theory (Kahneman and Tversky, 1979), it is often assumed that $\theta = 2$ (Selten and Chmura, 2008), but we will study a wider range of values to distinguish between loss aversion and ex-post rationality. For simplicity, we assume that $E = V$, that is the endowment is equal to the value of the prize, as in our experiment and in most other studies.

For each action profile $\{c_i, c_j\}$, with $c_i, c_j \in [0, V]$, we calculate the expected upward ($E^+(c_i, c_j)$) and downward ($E^-(c_i, c_j)$) impulses. The symmetric weighted impulse balance equilibrium is an action profile $\{c^*, c^*\}$ in which upward and downward impulses are equalized, that is $E^+(c^*, c^*) = E^-(c^*, c^*)$.

The ex-post rational action will be determined by the lottery outcome in SR and RS, and by the outcome of two lotteries in RR. There are no IBE in which both players invest nothing into the contest because there would be no downward impulse and a positive expected upward impulse, in all treatments. We can therefore calculate the probability to win the contest by $p_c = \frac{c_i}{c_i + c_j}$.

C.1 SR

The probability that i receives the prize in profile $\{c_i, c_j\}$ is $p_c = \frac{c_i}{c_i + c_j}$. The prize is won if a random variable $r_c \in \mathcal{U}(0, 1)$ is below p_c. In our experimental design, r_c was generated once per round, therefore if a prize was won, it would have also been won with a higher contest investment level, and if the prize was not won, it would not have been won with a lower investment level. If the win is possible, it is ex-post optimal to choose an action such that the probability to win is exactly equal to the generated number: $r_c = \frac{c_i^*}{c_i^* + c_j}$, therefore the ex-post rational action is $c_i^* = \frac{r_c c_j}{1 - r_c}$. If the win is not possible ($r_c > \frac{V}{V + c_j}$), the ex-post rational contest investment is 0.

Table C.1 shows that there are three possible outcomes. In the first case, player suffers from wasted investment: the prize was won, but it could have been won by a lower investment level. If the second case, the player suffers from lost opportunity: the prize was not won, but it could have been won by a sufficiently higher contest investment. In the third case, winning was not possible because of a high generated number, therefore the ex-post optimal investment level is 0. In the latter two cases, players do not win the prize despite a positive investment, therefore a loss is incurred, and impulses are multiplied by θ. We calculate the expected upward and downward impulses by integrating over all possible realizations of r_c.

[30] The assumption that the reference point is equal to the endowment has been implicitly made in papers that study loss aversion in contests (Cornes et al., 2003; Kong, 2008).

Outcome	Condition	$\pi(c_i, c_j)$	$\pi(c_i^*, c_j)$	Impulse size	Direction	Loss
Wasted investment	$r_c \in (0, p_c)$	$V - c_i + V$	$V - \frac{r_c c_j}{1-r_c} + V$	$c_i - \frac{r_c c_j}{1-r_c}$	$\downarrow$	No
Lost opportunity	$r_c \in (p_c, \frac{V}{V+c_j}]$	$V - c_i$	$V - \frac{r_c c_j}{1-r_c} + V$	$c_i - \frac{r_c c_j}{1-r_c} + V$	$\uparrow$	Yes
Impossible win	$r_c \in (\frac{V}{V+c_j}, 1)$	$V - c_i$	V	c_i	$\downarrow$	Yes

Table C.1: Impulses in SR.

$$E^-(c_i, c_j) = \int_0^{p_c} c_i - \frac{r_c c_j}{1 - r_c} \, \mathrm{d}r_c + \theta \int_{\frac{V}{V+c_j}}^1 c_i \, \mathrm{d}r_c = \\ = c_i + c_j \log\left(\frac{c_j}{c_i + c_j}\right) + \frac{\theta c_i c_j}{V + c_j}$$

$$E^+(c_i, c_j) = \theta \int_{p_c}^{\frac{V}{V+c_j}} c_i - \frac{r_c c_j}{1 - r_c} + V \, \mathrm{d}r_c = \\ = \frac{(V + c_i + c_j)(V - c_i)\theta c_j}{(V + c_j)(c_i + c_j)} + \theta c_j \log\left(\frac{c_i + c_j}{V + c_j}\right)$$

Action profile $\{c^*, c^*\}$ is a symmetric weighted impulse balance equilibrium if it satisfies $E^-(c^*, c^*) = E^+(c^*, c^*)$. Simplifying and rearranging gives the following condition:

$$\frac{\theta V}{2c^*} + \frac{2\theta V}{V + c^*} - 2\theta - 1 - \log(0.5) + \theta \log\left(\frac{2c^*}{V + c^*}\right) = 0 \tag{18}$$

For the parameters used in the experiment ($V = 8$), IBE predicts over-investment even without loss aversion ($c^* = 2.66$ if $\theta = 1$), and predicted investment is increasing in loss aversion ($c^* = 3.09$ if $\theta = 2$).

C.2 RS

In RS, the probability that player i receives the prize is $p_{nc} = \frac{V - c_i}{V}$. The prize is received if $r_{nc} \leqslant p_{nc}$, where r_{nc} is drawn uniformly from $[0, 1]$. Then the range of p_{nc} is $[0, 1]$, and the ex-post rational action must satisfy $p_{nc} = r_{nc}$, therefore $c_i^* = V - r_{nc}V$.

Table C.2 shows that two types of impulses are possible. If $r_{nc} < p_{nc}$, the player suffers from a wasted investment: the non-contest prize was won, but it could have been won by investing more into the contest (upward impulse). If $r_{nc} > p_{nc}$, the player suffers from a lost opportunity: the prize was not won, but it could have been won if the contest investment was sufficiently reduced (downward impulse). In the case of a lost opportunity, earnings are below the initial endowment, therefore the downward impulse is multiplied by θ. Expected impulses are calculated by integrating over the possible values of r_{nc}:

$$E^+(c_i, c_j) = \int_0^{p_{nc}} \left(\frac{V - r_{nc}V}{V - r_{nc}V + c_j} - \frac{c_i}{c_i + c_j}\right) V \, \mathrm{d}r_{nc} = \\ = (V - c_i)\frac{c_j}{c_i + c_j} + c_j \log\left(\frac{c_i + c_j}{V + c_j}\right)$$

Outcome	Condition	$\pi(c_i, c_j)$	$\pi(c_i^*, c_j)$	Impulse size	Direction	Loss
Wasted investment	$r_{nc} \in (0, p_{nc})$	$\frac{c_i}{c_i+c_j}V + V$	$\frac{V-r_{nc}V}{V-r_{nc}V+c_j}V + V$	$(\frac{V-r_{nc}V}{V-r_{nc}V+c_j} - \frac{c_i}{c_i+c_j})V$	$\uparrow$	No
Lost opportunity	$r_{nc} \in (p_{nc}, 1)$	$\frac{c_i}{c_i+c_j}V$	$\frac{V-r_{nc}V}{V-r_{nc}V+c_j}V + V$	$(\frac{V-r_{nc}V}{V-r_{nc}V+c_j} + \frac{c_j}{c_i+c_j})V$	$\downarrow$	Yes

Table C.2: Impulses in RS.

$$E^-(c_i, c_j) = \theta \int_{p_{nc}}^{1} \left(\frac{V - r_{nc}V}{V - r_{nc}V + c_j} + \frac{c_j}{c_i + c_j} \right) V \, \mathrm{d}r_{nc} = \\ = \frac{\theta c_i (c_i + 2c_j)}{c_i + c_j} + \theta c_j \log \left(\frac{c_j}{c_i + c_j} \right)$$

Setting $E^-(c^*, c^*) = E^+(c^*, c^*)$ and simplifying gives the following condition:

$$c^*(0.5 + (1.5 + \log(0.5))\theta) - c^* \log \left(\frac{2c^*}{V + c^*} \right) - 0.5V = 0 \tag{19}$$

If $V = 8$ and $\theta = 1$, $c^* = 1.7$. Loss aversion reduces contest investment even further, because it increases the downward impulse.

C.3 RR

In RR, players face a tradeoff between the probability to win the contest prize ($p_c = \frac{c_i}{c_i+c_j}$) and the probability to win the non-contest prize ($p_{nc} = \frac{V-c_i}{V}$). Figure C.1 illustrates this tradeoff using the "budget line" (plotted in blue), calculated as $p_c = \frac{V(1-p_{nc})}{V(1-p_{nc})+c_j}$. If nothing is invested into the contest, $p_{nc} = 1$ and $p_c = 0$ (bottom right corner); if the entire endowment is invested into contest, $p_{nc} = 0$ and $p_c = \frac{V}{V+c_j}$ (top left corner). Lottery outcomes are determined by numbers $r_c, r_{nc} \in \mathcal{U}(0, 1)$. Lottery outcome can be represented by a point on a unit square, and the corresponding prize is won if the generated number is closer to the origin than the probability to win.

Figure C.1 illustrates impulses for the RNNE action, marked on the line. If the lottery outcome falls in area a, both prizes are won, therefore the chosen action maximizes ex-post payoffs. If the lottery outcome is in area b, the player receives only the non-contest prize, and there is a higher contest investment level that would have resulted in winning both prizes, therefore an upward impulse of size V is received. If the lottery outcome is in area c, only the contest prize is received, while both prizes could have been won by choosing a lower contest investment level, therefore a downward impulse of size V is received. Only the non-contest prize is received in areas d and e, and only the contest prize is received in area f; these areas lie outside of the budget set, therefore both prizes could not have been won, and no impulse is received. In area g, no prize is received, but a prize could have been won by either a lower or a higher contest investment. We assume that no impulse is received in that case. In area h, no prize is received, but a non-contest prize could have been received by a sufficiently lower contest investment, therefore a downward impulse is received. If the lottery outcome falls in area h, payoffs fall below the endowment, therefore the impulse is multiplied by the loss aversion parameter θ.

Players receive an upward impulse if the outcome is in area b, and a downward impulse if it is in c or h. We calculate these probabilities by integrating the area under the budget line:

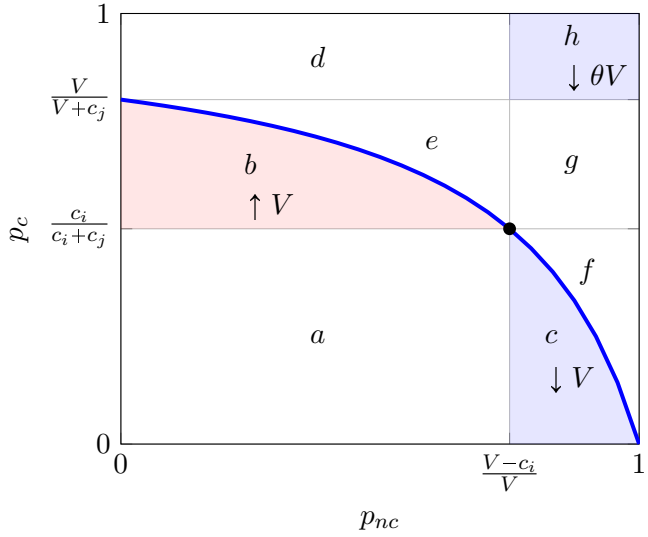


Figure C.1: Impulses in RR. The marked location and impulses are plotted for the RNNE action ($c_i = V/4$). The "budget line", marked in blue, shows the boundary of all attainable combinations of probabilities, if the opponent chooses the RNNE action ($c_j = V/4$).

$$b = \int_0^{\frac{V-c_i}{V}} \frac{V(1-p_{nc})}{V(1-p_{nc})+c_j} - \frac{c_i}{c_i+c_j}\,\mathrm{d}p_{nc} = \frac{c_j}{V}\log\left(\frac{c_i+c_j}{V+c_j}\right) + \frac{c_j(V-c_i)}{V(c_i+c_j)}$$

$$c = \int_{\frac{V-c_i}{V}}^1 \frac{V(1-p_{nc})}{V(1-p_{nc})+c_j}\,\mathrm{d}p_{nc} = \frac{c_j}{V}\log\left(\frac{c_j}{c_i+c_j}\right) + \frac{c_i}{V}$$

$$h = \left(1-\frac{V-c_i}{V}\right)\left(1-\frac{V}{V+c_j}\right) = \frac{c_i c_j}{V(V+c_j)}$$

Expected impulse is a product of impulse probability and strength: $E^+(c_i,c_j) = bV$, $E^-(c_i,c_j) = cV + h\theta V$. A symmetric impulse balance equilibrium $\{c^*,c^*\}$ must equalize upward and downward impulses. Substituting the definition of b, c and h and simplifying gives the following condition that c^* must satisfy:

$$\log\left(\frac{4c^*}{V+c^*}\right) - \frac{3c^*-V}{2c^*} - \theta\frac{c^*}{V+c^*} = 0 \tag{20}$$

If $V = 8$ and $\theta = 1$, $c^* = 2.12$, but increased loss aversion increases the downward impulse, therefore IBE falls below RNNE when loss aversion is sufficiently strong.

In SS, the ex-post rational action is always equal to the best-response, thus no player receives an impulse only if the actions of both players are mutual best-responses. Therefore, IBE in the SS treatment coincides with the RNNE.

Each of the conditions (18), (19) and (20) are satisfied by a unique c^*, which depends on the loss aversion parameter θ. These IBE predictions are compared in panel (h) of Figure B.1. Without loss aversion, IBE is above RNNE in SR, but below it in RS. This difference is a

result of a strong impulse after failing to win the prize. Expectation of this impulse induces players to over-invest into activity that can potentially provide the prize. Loss aversion further increases the intensity of lost opportunity, and therefore increases the gap between SR and RS. RR treatment falls between SR and RS, because ex-post rational actions can lie on either side of the chosen action.

D Supplementary results for Study 1

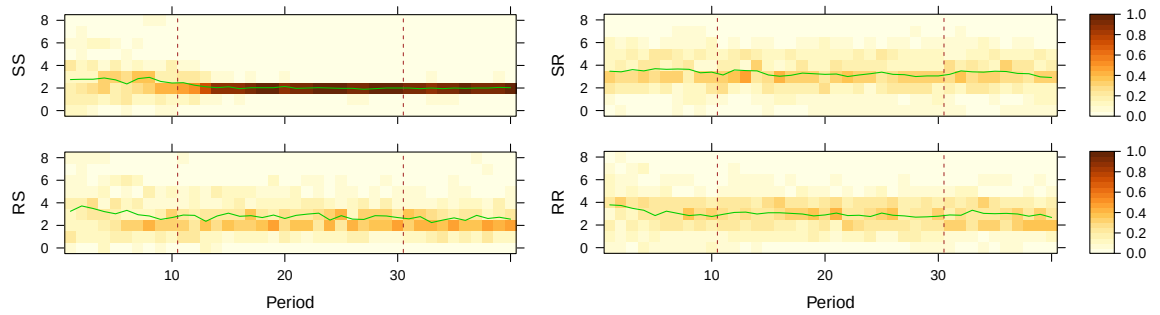


Figure D.1: Density of contest investment and its average (green line) in each treatment. FPI was available from round 11 to round 30.

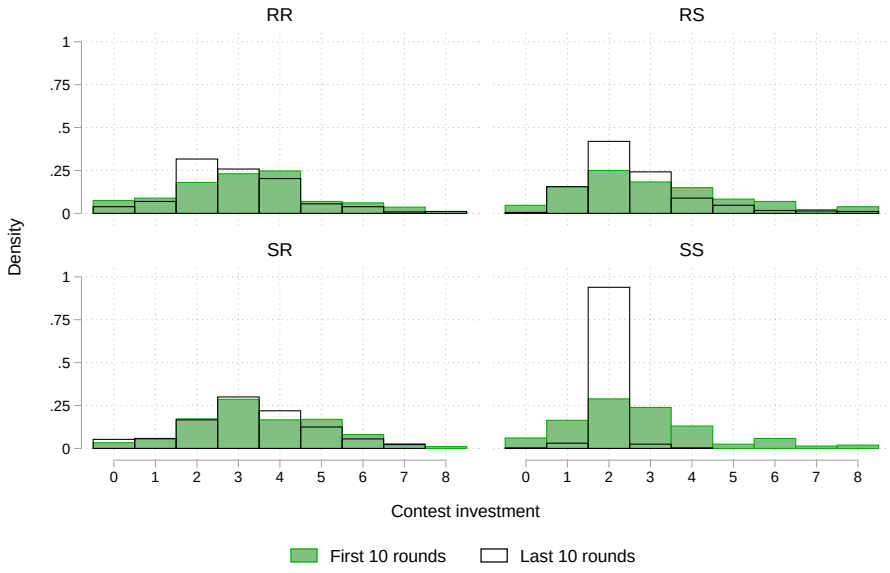


Figure D.2: Distribution of choices in the first 10 and in the last 10 rounds, by treatment.

Table D.1: Aggregate outcomes in rounds 1-10 in Study 1. Outcomes are average contest investment, standard deviation of contest investment, frequency with which both matched participants choose the RNNE action, frequency of dominated strategies and average absolute deviation from the RNNE prediction. All measures are aggregated on the treatment level. Statistical significance is evaluated using two-sided p-values of a Mann-Whitney U test, when data is averaged on the matching group level.

	SS	SR	p-value	RS	p (vs SS)	RR	p (vs SS)
Contest investment	2.71	3.54	0.0104	3.11	0.2971	3.19	0.1093
SD	1.70	1.64	0.6310	1.93	0.7488	1.77	0.8728
RNNE investment	8%	2%	0.3099	9%	0.8072	3%	0.1990
RNNE deviation	1.28	1.77	0.0303	1.61	0.6310	1.66	0.1093
Dominated	49%	74%	0.0103	55%	0.6874	66%	0.1081

Table D.2: Aggregate outcomes in rounds 11-30 in Study 1. Outcomes are average contest investment, standard deviation of contest investment, frequency with which both matched participants choose the RNNE action, frequency of dominated strategies and average absolute deviation from the RNNE prediction. All measures are aggregated on the treatment level. Statistical significance is evaluated using two-sided p-values of a Mann-Whitney U test, when data is averaged on the matching group level.

	SS	SR	p-value	RS	p (vs SS)	RR	p (vs SS)
Contest investment	2.05	3.21	0.0039	2.79	0.0039	2.92	0.0039
SD	0.44	1.55	0.0039	1.42	0.0039	1.42	0.0039
RNNE investment	81%	4%	0.0037	13%	0.0038	5%	0.0037
RNNE deviation	0.14	1.55	0.0039	1.12	0.0039	1.33	0.0039
Dominated	7%	70%	0.0038	51%	0.0038	62%	0.0039

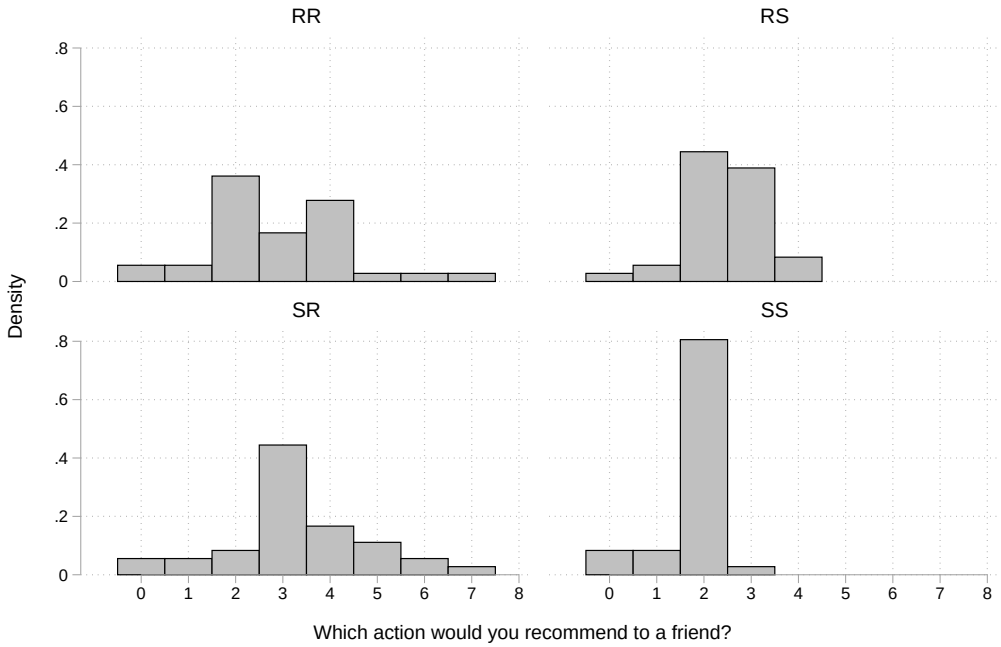


Figure D.3: Distribution of actions that players would recommend a friend to play.

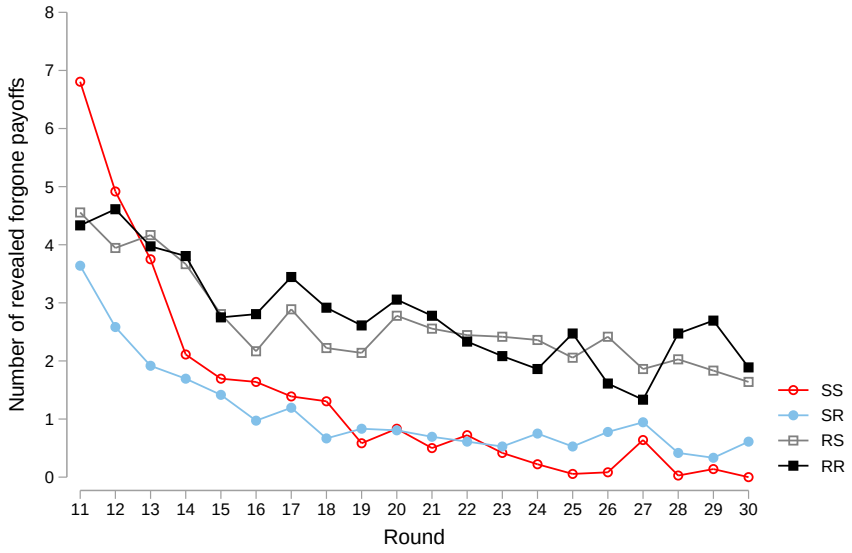
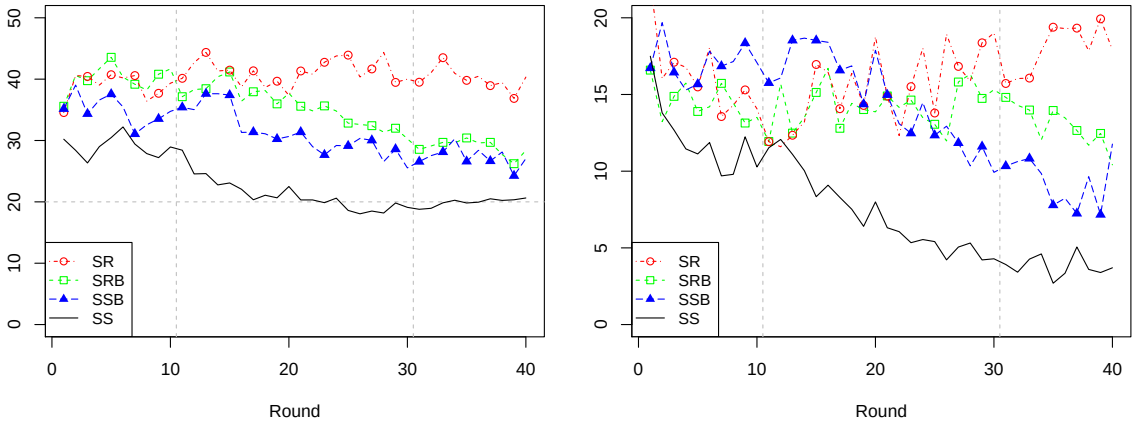


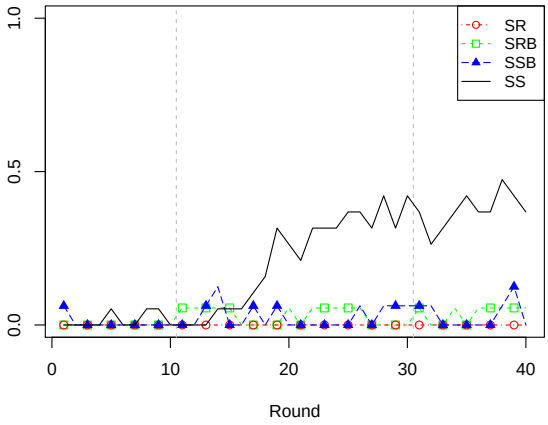
Figure D.4: Average pieces of foregone payoff information that participants chose to reveal in the rounds where revelation was possible (11-30). Participants could reveal between 0 and 8 pieces of information.

E Supplementary results for Study 2

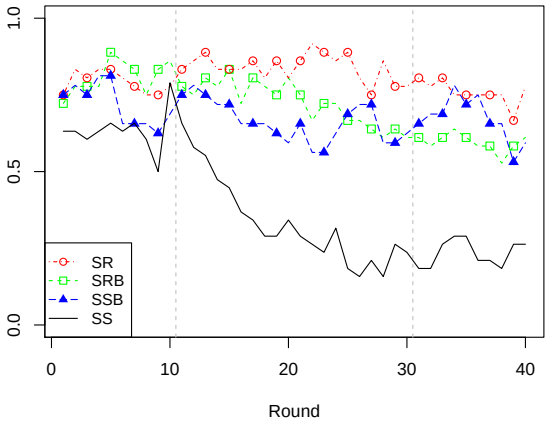


(a) Average contest investment

(b) Standard deviation of investment



(c) Fraction of pairs in RNNE



(d) Fraction of dominated strategies

Figure E.1: Dynamics in Study 2. FPI was available in SS and SR treatments from round 11 to round 30. FPI was never available in SSB and SRB.

Table E.1: Aggregate outcomes in rounds 11-30 in Study 2. Outcomes are average contest investment, standard deviation of contest investment, frequency with which both matched participants choose the RNNE action, frequency of dominated strategies and average absolute deviation from the RNNE prediction. All measures are aggregated on the treatment level. Statistical significance is evaluated using two-sided p-values of a Mann-Whitney U test, when data is averaged on the matching group level.

	SR	SS	p-value	SRB	p (vs SR)	SSB	p (vs SS)
Contest investment	41.2	21.2	0.0029	35.7	0.5286	31.3	0.0009
SD	17.6	9.3	0.0015	19.4	0.7527	16.2	0.0172
RNNE investment	0%	22%	0.0008	3%	0.1441	3%	0.0078
RNNE deviation	24.2	5.9	0.0005	20.1	0.5995	13.4	0.0009
Dominated	84%	33%	0.0005	73%	0.8747	66%	0.0008

Table E.2: Aggregate outcomes in rounds 1-10 in Study 2. Outcomes are average contest investment, standard deviation of contest investment, frequency with which both matched participants choose the RNNE action, frequency of dominated strategies and average absolute deviation from the RNNE prediction. All measures are aggregated on the treatment level. Statistical significance is evaluated using two-sided p-values of a Mann-Whitney U test, when data is averaged on the matching group level.

	SR	SRB	p-value	SS	SSB	p-value
Contest investment	39.0	40.1	0.4622	29.0	35.0	0.2664
SD	18.4	19.8	0.2076	15.0	17.8	0.0390
RNNE investment	0%	0%	–	2%	1%	0.5868
RNNE deviation	22.9	23.3	0.5286	13.8	17.5	0.4273
Dominated	79%	81%	0.4929	63%	72%	0.8321

Table E.3: Comparison of SSB and SSB treatments in Study 2. FPI was not available in either treatment. Statistical significance is evaluated using two-sided p-values of a Mann-Whitney U test, when data is averaged on the matching group level.

	All rounds			Rounds 31-40		
	SRB	SSB	p-value	SRB	SSB	p-value
Contest investment	35.1	31.2	0.4179	28.9	27.4	0.8621
SD	19.7	15.7	1.0	18.5	10.9	0.4179
RNNE investment	3%	2%	0.0667	4%	3%	0.2378
RNNE deviation	19.9	13.5	0.0826	16.0	9.6	0.1052
Dominated	71%	68%	0.4175	59%	67%	0.6012

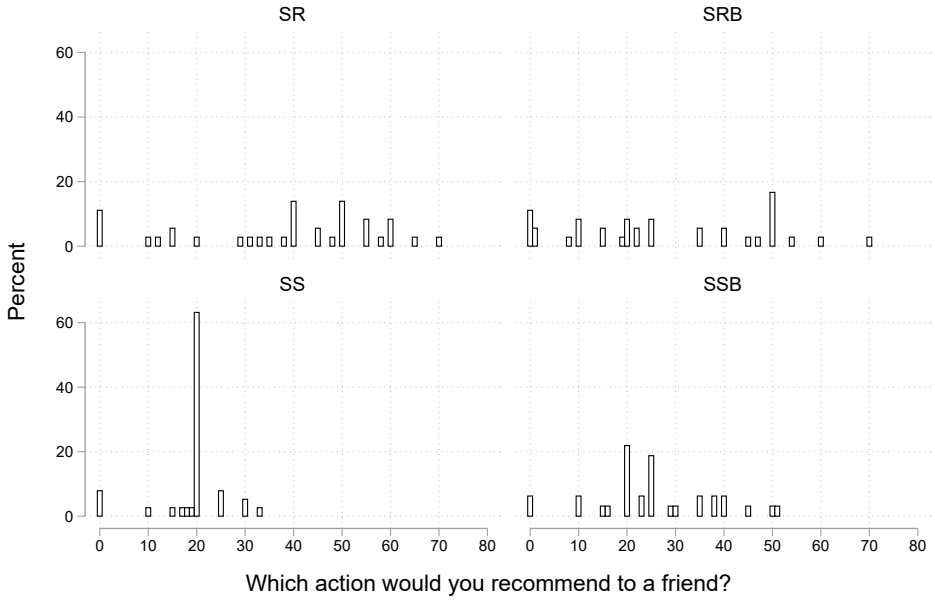
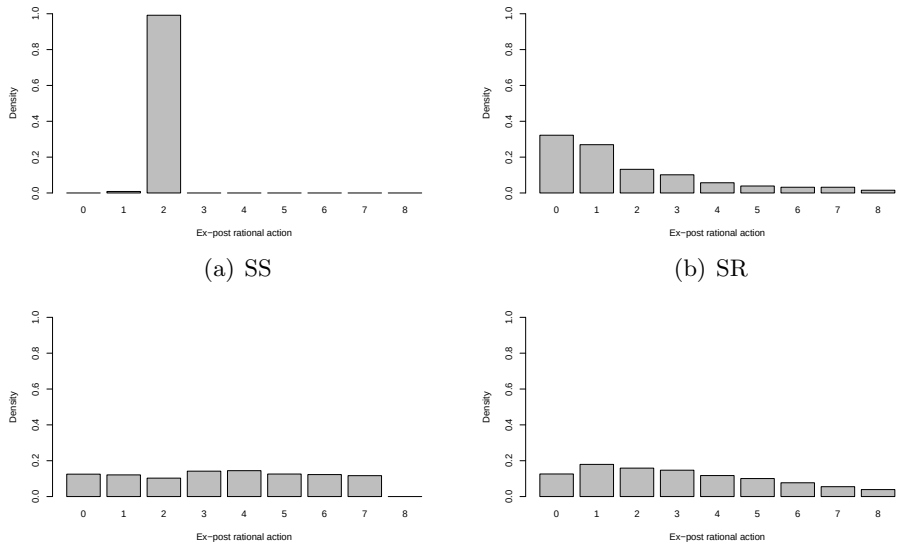


Figure E.2: Distribution of contest investment that participants would recommend a friend to choose in Study 2.

F Feedback and adaptation in Study 1

We have shown that choices converge to RNNE in SS, but not in the other treatments. To understand the mechanism behind the treatment effect, we compare the obtained feedback and its effect on subsequent choices.



(a) SS (b) SR

(c) RS (d) RR

Figure F.1: Distributions of ex-post payoff-maximizing actions in rounds with FPI.

To measure what players could learn from the game, for each round we calculate the action that would have provided the highest payoff.[31] We use data from rounds with FPI, which allows the ex-post payoff-maximizing action to be known. Figure F.1 shows that the RNNE action ($c_i^* = 2$) almost always generates the highest payoff in SS, but not in the other treatments. In SR, payoffs are typically maximized by investing 0 or 1, which are optimal when winning is either impossible or guaranteed by choosing the smallest investment level. In RS and RR, all actions maximize payoffs with similar frequency, thus the observed payoff information does not allow to identify the expected payoff-maximizing action. We conclude that the feedback from a single round in SR, RS and RR treatments reveals little information that could improve the quality of subsequent decisions.

Next, we study how the observed feedback affects the direction in which participants subsequently adjust their choices. We compare the predictions of the following theories:

- **Cournot best-response** predicts that players choose the action that maximized the expected payoffs in the previous round. Cournot best-response would converge to RNNE in all treatments.
- **Imitate-the-best** predicts that players adapt towards the action of the player who received highest earnings. In a two player game, players either change actions in the

[31] We calculate the payoffs associated with each action by holding constant opponent's action and the draw of a random number, used to determine the lottery outcome. These payoffs are identical to those observed by participants. When the payoff-maximizing action is unique, it receives a weight of 1. When multiple actions generate the same maximal payoff, the weight is divided equally among those actions.

direction of opponent's action (if opponent received a higher payoff), or make no change. In SS, imitate-the-best would converge to the relative payoff maximization point (Fallucchi et al., 2013), equal to 4 with the parameters of our experiment.

- **"Chasing" hypothesis** predicts that players adapt towards the action that provided the highest foregone payoff (Ert and Erev, 2007). This adaptation rule ignores the magnitude of payoffs and therefore may converge to the action that outperforms other actions most of the time, instead of the one that maximizes expected payoffs.[32] In contests, "chasing" converges to RNNE only in SS, while in other treatments it oscillates depending on the lottery outcomes. We expect that "chasing" will be used only in rounds with FPI. Since we have information about which FPI was actually observed, we will also test if players choose actions with the highest payoff from the set of actions with observed FPI.

Theories are compared by estimating whether the change in investment can be explained by observed feedback. We use a model adapted from Huck et al. (1999):

$$c_i^t - c_i^{t-1} = \beta_0 + \beta_{RP}(c_{RP}^{t-1} - c_i^{t-1}) + \beta_{OP}(c_{OP}^{t-1} - c_i^{t-1}) + \beta_{EP}(c_{EP}^{t-1} - c_i^{t-1}) + \beta_I(c_I^{t-1} - c_i^{t-1}) + \varepsilon_i^t$$

where c_{RP}^{t-1} is the action with the highest *realized payoff* in round $t-1$; c_{OP}^{t-1} is the action with the highest *observed realized payoff* in round $t-1$; c_{EP}^{t-1} is the action that would have maximized the *expected payoff*; c_I^{t-1} is the action chosen by the player with the highest payoff. The "chasing" hypothesis predicts adaptation towards c_{OP}^{t-1}, Cournot best-response predicts adaptation towards c_{EP}^{t-1} and imitate-the-best predicts adaption towards c_I^{t-1}.

We start by estimating the parameter values from rounds in which FPI was available. First, we calculate the action that maximized foregone payoffs, regardless of whether they were observed or not.[33] Table F.1 displays the estimated parameter values from all treatments pooled together. In the first model, all coefficients are significantly different from zero, but the largest coefficient belongs to the expected payoff-maximizing action. If parameters are estimated separately for each treatment, adjustment towards expected payoff-maximizing action is significant in all treatments, imitation is not significant in RR and SS, and adjustment towards foregone payoff-maximizing action is not significant in RR.

Since we have information about which foregone payoffs are observed, we test a hypothesis that players adapt in the direction of the action with the highest observed payoff. Model (2) shows that this variable (β_{OP}) is highly significant, while β_{RP} is no longer significant. Results are very similar if we set $\beta_{RP} = 0$ or if we remove rounds in which FPI was not observed (although the p-value of β_{OP} increases to 0.063), or if we additionally remove rounds in which multiple actions provide identical maximal payoffs.

Next, we look at rounds in which FPI was not available (1-10 and 31-40). The software calculated foregone payoffs in all rounds, but since this information was withheld, we expect it to play no role in the adjustment process. If the parameter is significantly different from

[32] It has been shown that "chasing" can reduce efficiency if there is another option that has higher long-run benefits (Otto and Love, 2010) or if the option that is usually better has a lower expected value (Yechiam and Busemeyer, 2006).

[33] If several actions provide the same maximal payoff, we calculate their average. We will use the average value in all specifications so as not to discard any observations, but the results are very similar if observations with multiple payoff-maximizing actions are discarded.

Table F.1: Random-effects GLS regression with a matching-group-specific random component. Independent variable is the change in contest investment. Standard errors are clustered on a matching group level.

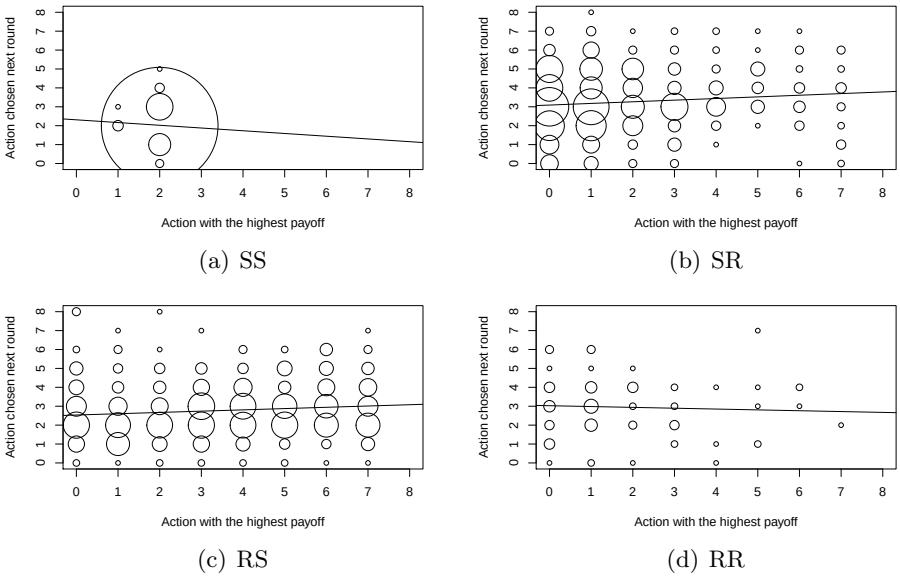
	(1) Rounds with FPI	(2) Rounds with FPI	(3) Rounds without FPI
β_{RP}	0.0586***	0.00699	0.0187
	(3.94)	(0.41)	(1.02)
β_{OP}		0.163***	
		(6.12)	
β_{EP}	0.223***	0.225***	0.327***
	(5.17)	(5.30)	(9.61)
β_I	0.166***	0.160***	0.206***
	(5.99)	(5.85)	(5.59)
Constant	0.161***	0.146***	0.307***
	(3.35)	(3.25)	(4.61)
Overall R^2	0.205	0.228	0.244
N. of observations	2736	2736	2592

t statistics in parentheses

* p<0.10, ** p<0.05, *** p<0.01

zero, adjustment towards the foregone payoff-maximizing action would be caused by some other reason. Column 3 in Table F.1 shows that it is not: the estimated value of β_{RP} is small and not significantly different from zero, while the parameters for imitation and Cournot-best-response are highly significant. When we estimate the model separately for each game and either the first or the last block of 10 rounds, we find that foregone payoffs do no play any role in all games except for SS, where its predictions coincide with those of expected payoff maximization. Adaptation towards expected payoff-maximizing action is highly significant in all treatments and blocks, while imitation is significant in the first block for SS, SR and RS treatments (p-values are at most 0.006), but not significant in the last block (p-values are at least 0.152). This decrease over time suggests that imitation is replaced by more rational adaptation rules once players gain more experience.

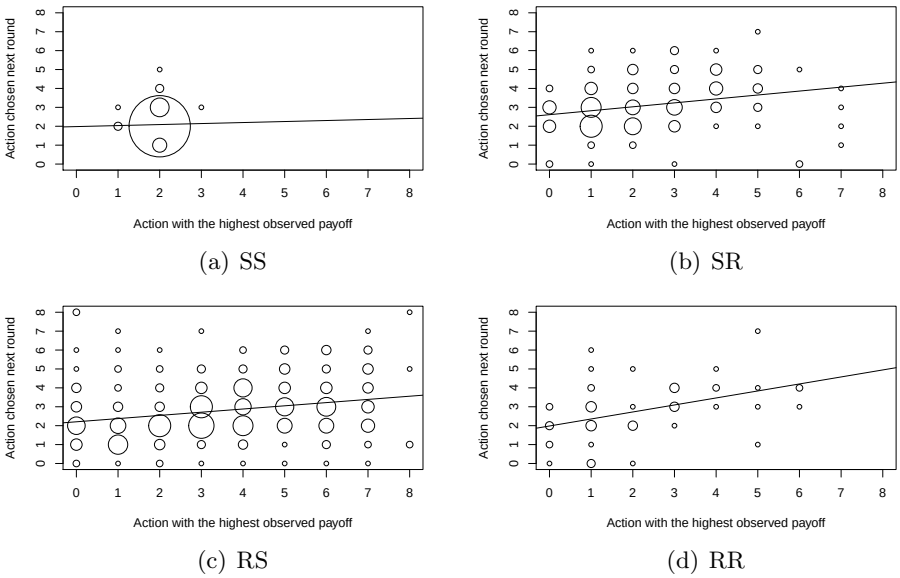
In terms of the adaptation rules, we find that players choose the action that maximized ex-post payoffs, but only when these payoffs are observed. We calculate the action which would have maximized ex-post payoffs in the previous round using data from rounds in which FPI was available and there was a single payoff-maximizing action. Figure F.2 displays the joint distributions of ex-post payoff-maximizing investment levels and the investment chosen the following period (area of a bubble is proportional to the number of observations). There is little evidence that players choose actions that have performed best in the previous round, and the slope of a regression line (displayed in the graph) is not significantly different from zero, in all treatments. However, in Study 1, information was initially hidden and participants might never have learned the ex-post payoff-maximizing actions if they did not reveal this information. Information acquisition was costless, but few players acquired information about all actions (see Figure D.4). Since information acquisition was tracked, we can identify the action that provided the highest payoff from the set of actions whose foregone payoffs were observed. Figure F.3 shows that players do choose actions that were observed providing high



(a) SS (b) SR

(c) RS (d) RR

Figure F.2: Ex-post payoff-maximizing action, and the action chosen next period. Only data from rounds with FPI in which one action maximizes ex-post payoffs. Area of a bubble is proportional to the number of observations. Regression lines are added to all graphs.



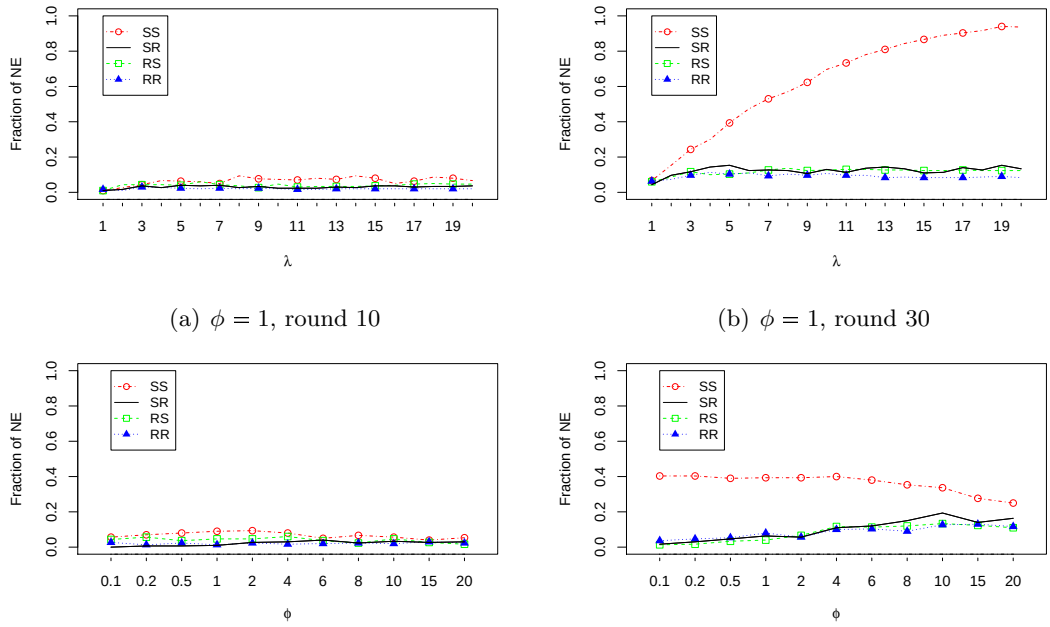
(a) SS (b) SR

(c) RS (d) RR

Figure F.3: Observed ex-post payoff-maximizing action, and the action chosen next period. Only data from rounds with FPI in which at least one foregone payoff was observed and one action maximized ex-post payoffs. Area of the bubble is proportional to the number of observations. Regression lines are added to all graphs.

payoffs, and the slope of the regression line is significantly higher than zero in SS and RR (p-values respectively 0.055 and 0.026) and marginally significant in SR ($p = 0.11$).

G Supplementary figures



(a) $\phi = 1$, round 10

(b) $\phi = 1$, round 30

(c) $\lambda = 10$, round 10

(d) $\lambda = 10$, round 30

Figure G.1: Fraction of RNNE pairs in reinforcement learning simulations. Data only from round 10 and round 30.

H Screenshots from Study 1

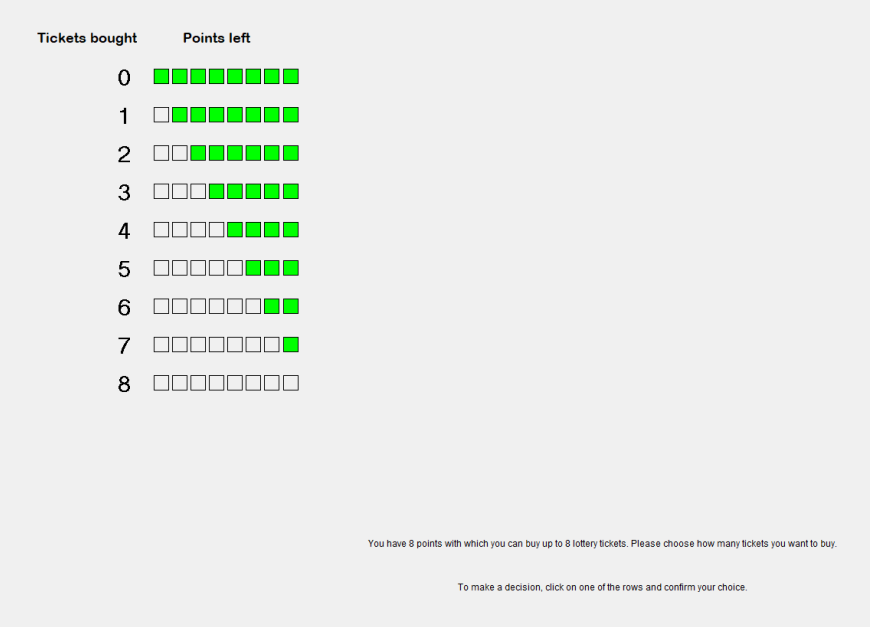


Figure H.1: Decision screen in SR treatment.

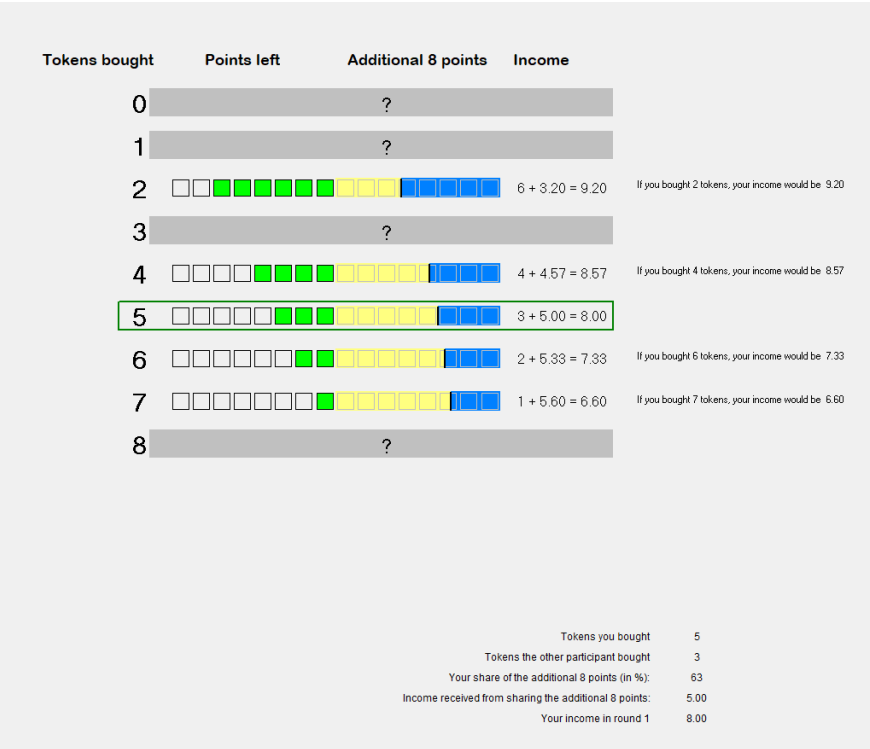
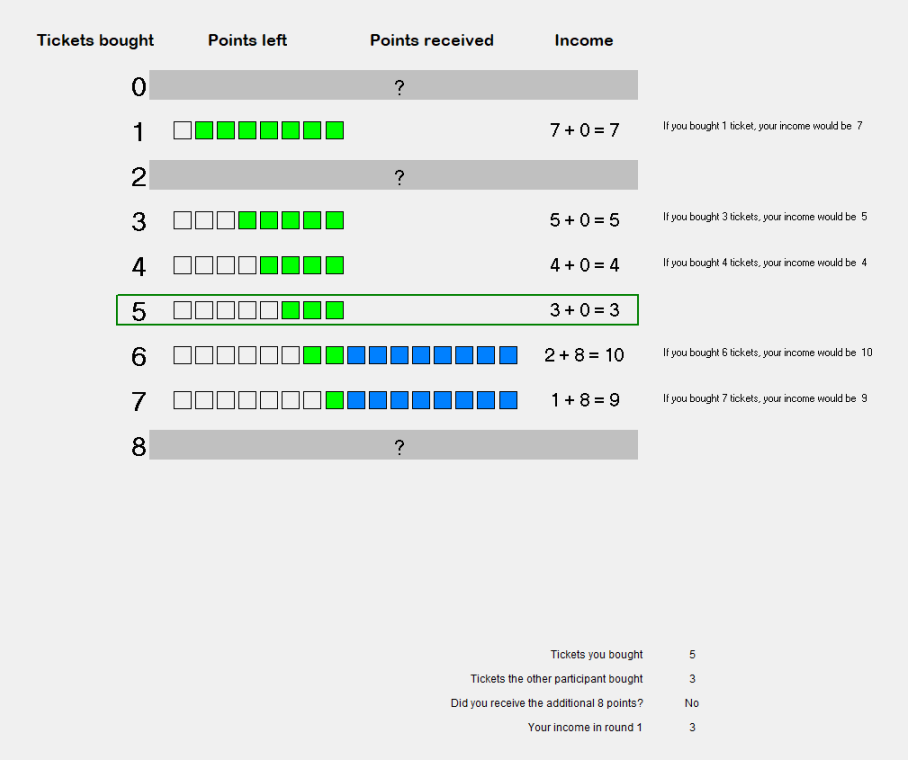


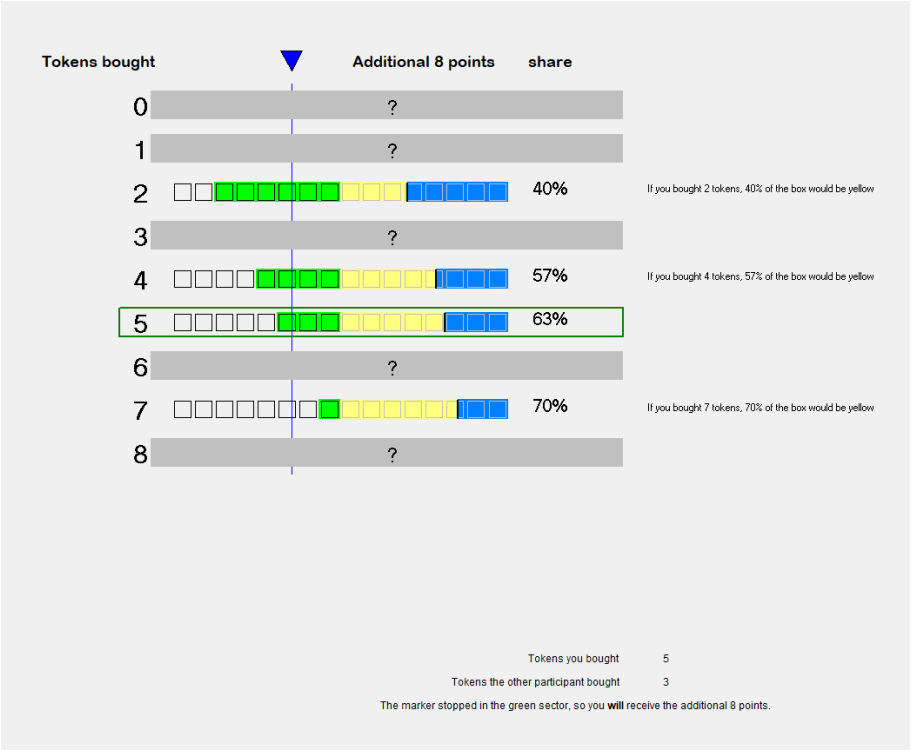
Figure H.2: Feedback in SS treatment with FPI.

(a) First screen

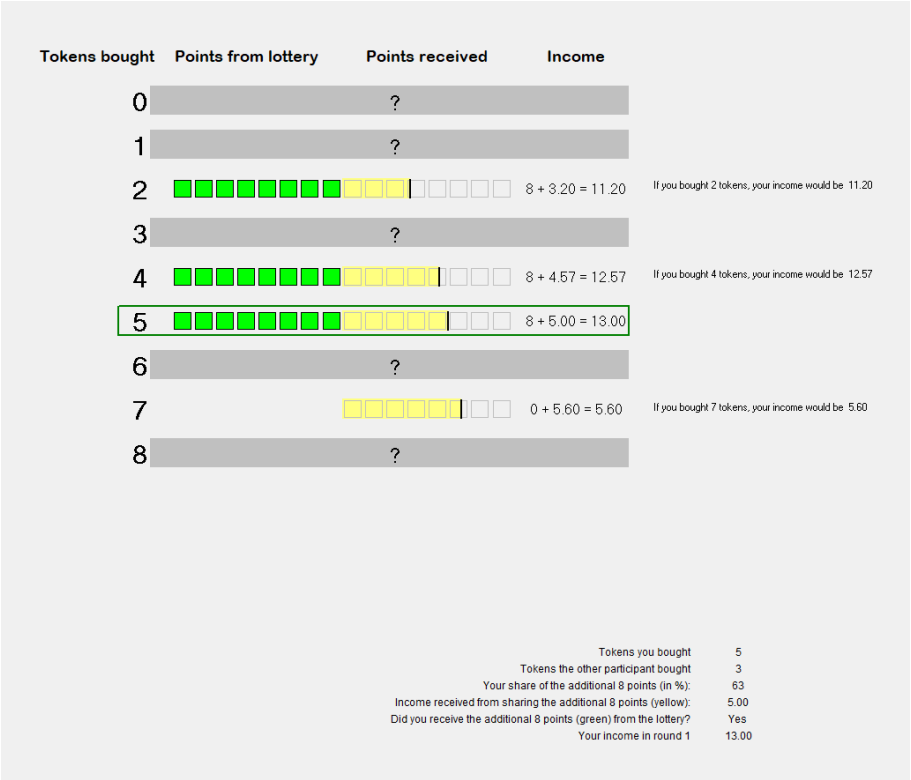


(b) Second screen

Figure H.3: Feedback in SR treatment with FPI.

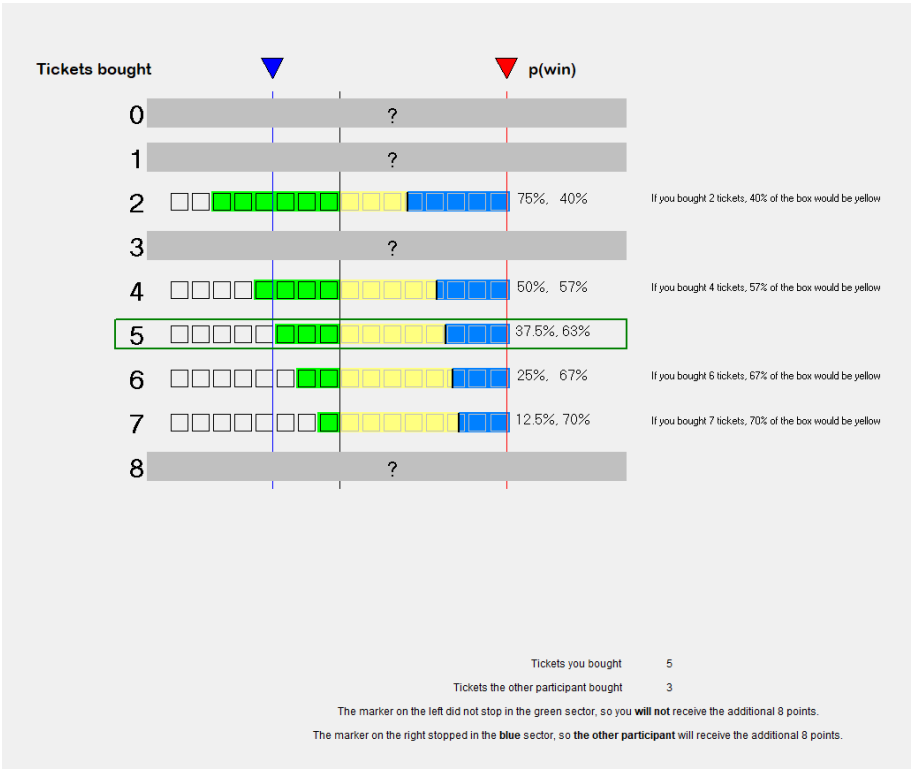


(a) First screen

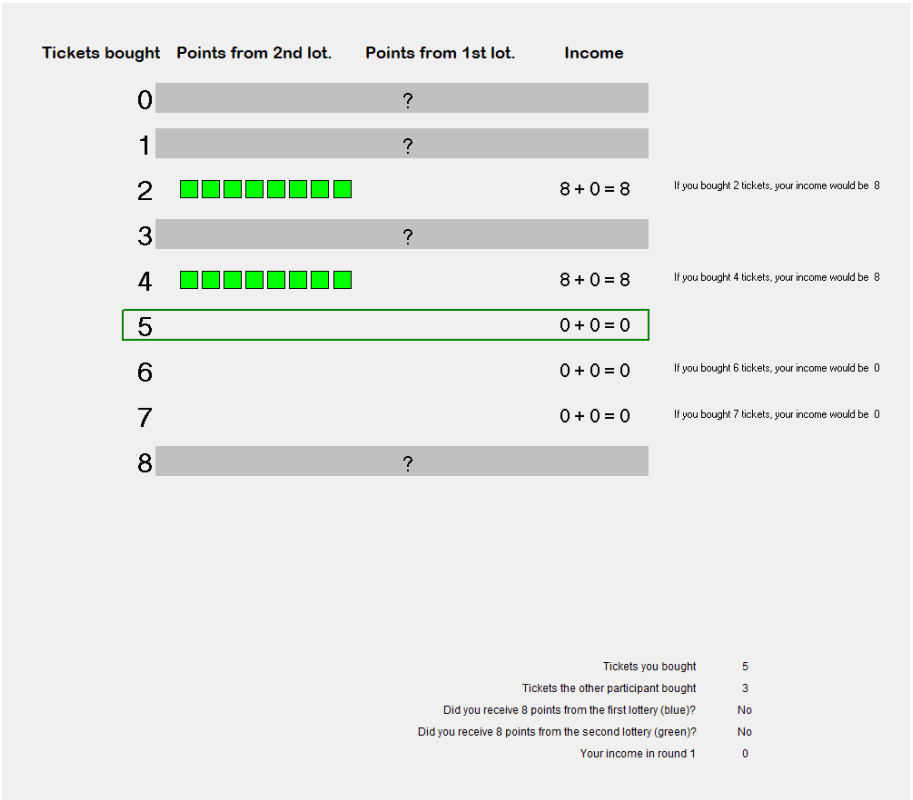


(b) Second screen

Figure H.4: Feedback in RS treatment with FPI.

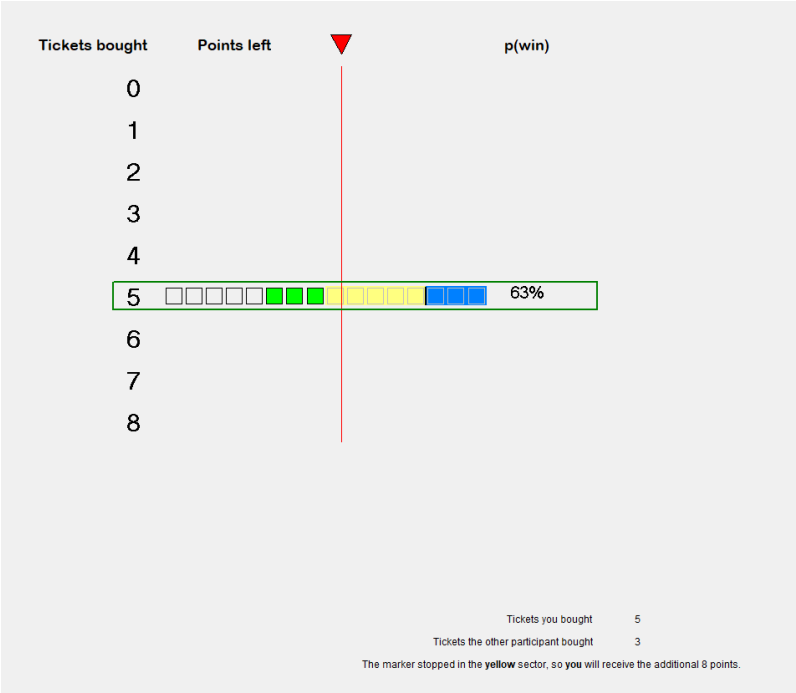


(a) First screen

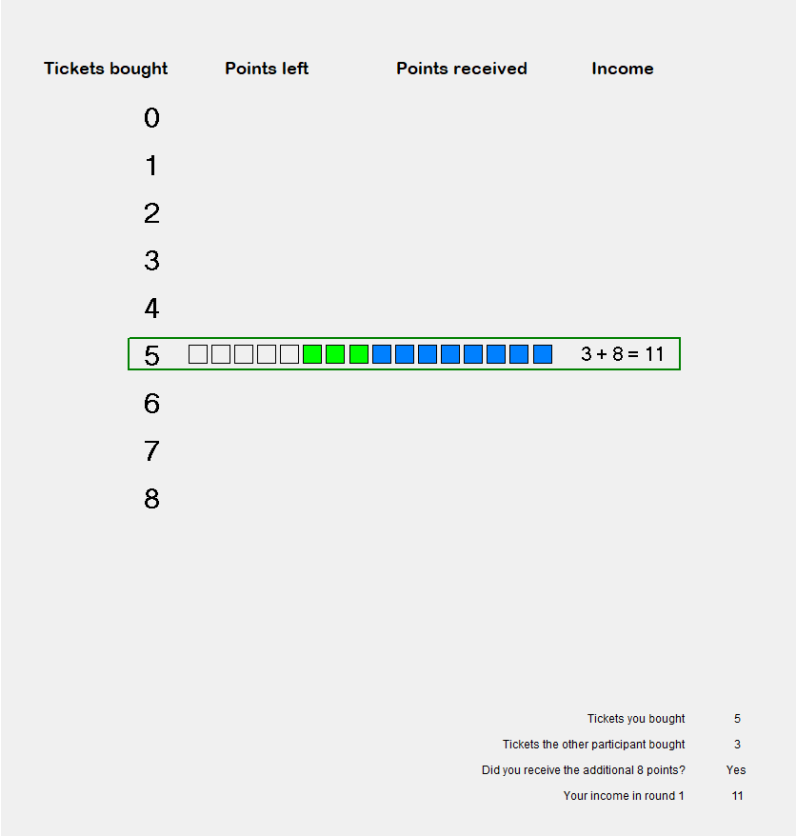


(b) Second screen

Figure H.5: Feedback in RR treatment with FPI.



(a) First screen



(b) Second screen

Figure H.6: Feedback in SR treatment with no FPI.

I Instructions for Study 1

Below we reproduce the instructions for SR treatment, with changes in other treatments marked in brackets.

INSTRUCTIONS

Welcome to the experiment. Please read these instructions carefully. They are identical for all the participants with whom you will interact during this experiment. If you have any questions please raise your hand. One of the experimenters will come to you and answer your questions. From now on communication with other participants is not allowed. If you do not conform to these rules you will be excluded from the experiment with no payment. Please also switch off your mobile phone at this moment.

In this experiment you can earn some money. How much you earn depends on your decisions and the decisions of the other participants. During the experiment we will refer to points instead of euros. The total amount of points that you will have earned during the experiment will be converted into euros at the end of the experiment and paid to you in cash confidentially. The conversion rate that will be used to convert your points into your cash payment will be **1 point = 0.45 euros.**

At the end of the experiment you will see how many points you earned in each round. One round from each 10 rounds will be randomly chosen for payment: the first round will be chosen from rounds 1-10, the second round from rounds 11-20 and so on. Your earnings from the selected rounds will be added, converted into euros and paid to you in private at the end of the experiment.

The experiment will contain several parts. Now we will describe you the task that you will be doing in Part 1. At the start of each other part additional instructions will be displayed on your computer screen. Please read those instructions before you continue.

The task in Part 1

In Part 1 there will be 10 rounds. At the beginning of each round the computer will randomly match you with another participant in this room. The participant you are matched with will be changed randomly each round. All participants with whom you will interact will receive the same information and will face exactly the same task.

Decision screen

In each round you will receive an endowment of 8 points. You can use these points to buy "lottery tickets" [*SS, RS:* "tokens"]. Each ticket [*SS, RS:* token] you buy costs 1 point, so you can buy up to 8 tickets [*SS, RS: tokens*] each round. Any points that you do not spend on tickets [*SS, RS: tokens*] will be added to your round income [*RS, RR:* will determine the probability to receive points in a second lottery.].

How your decision screen will look like is shown in Figure 1. The 8 points that you receive at the start of each round will be represented by 8 green squares. You will have to choose how many of these points to use to buy lottery tickets [*SS, RS:* tokens]. Every additional ticket [*SS, RS:* token] you buy will take away one green square. To make a decision, click on one of the rows and confirm your choice.

[*SR, RR:* **The lottery**]

[*SR, RR*: After you and the other participant have chosen how many tickets to buy, either you or the other participant will receive additional 8 points. Who receives these additional 8 points will be determined by a lottery. The lottery will be implemented the following way: after you have chosen how many tickets to buy, a box containing 8 squares will be displayed

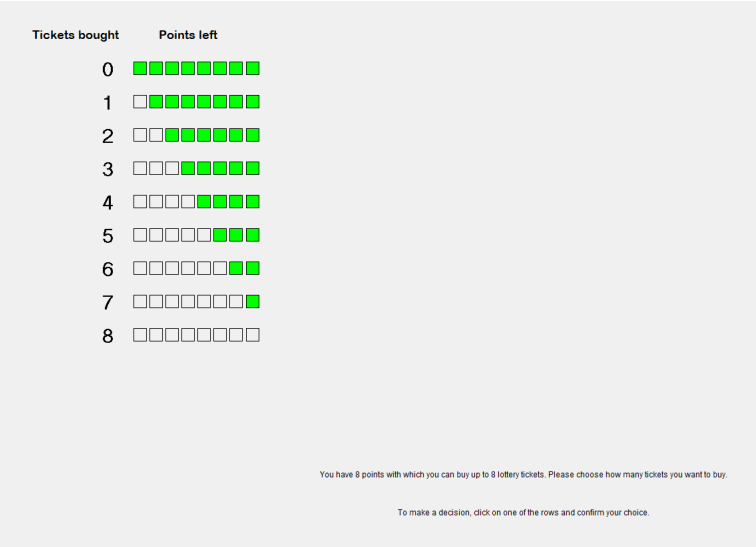


Figure 1. Screenshot of the decision screen. [In RS and SS figure showed "Tokens bought"]

on your screen. This box represents the additional 8 points that you may receive. The box will be divided into two sectors, yellow and blue. The yellow sector belongs to you and the blue sector belongs to the other participant. Once you start the lottery, a marker will move around the box and stop at one place at random. The marker is equally likely to stop at any place. If the marker stops in the yellow sector, you will receive the additional 8 points. If the marker stops in the blue sector, the other participant will receive the additional 8 points. The size of the yellow sector represents the probability that you will receive the additional 8 points: if $x\%$ of the box is colored yellow, the probability that you will receive the 8 additional points is $x\%$.]

[*SS, RS*: After you and the other participant have chosen how many tokens to buy, you and the other participant will share additional 8 points. These additional 8 points will be displayed as a box, divided into two sectors, yellow and blue. The yellow sector represents your share and the blue sector represents the share of the other participant.]

The sizes of yellow and blue sectors are proportional to the number of lottery tickets [*SS, RS:* tokens] that you and the other participant buy. For example, if you buy the same number of tickets [*SS, RS:* tokens] as the other participant, the yellow sector will be of the same size as the blue sector. If you buy twice as many tickets [*SS, RS:* tokens] as the other participant, the yellow sector will be two times larger than the blue sector. Overall, the probability that you will receive the additional 8 points [*SS, RS:* the share of the additional 8 points that you will receive], represented by the size of the yellow sector, is calculated as follows:

[*SR, RR:*]

$$\text{Probability of receiving the additional 8 points} = \frac{\text{Number of tickets you bought}}{\text{Number of tickets you bought + Number of tickets the other participant bought}} \times 100\%$$

[*SS, RS:*]

$$\text{Share of the additional 8 points} = \frac{\text{Number of tokens you bought}}{\text{Number of tokens you bought + Number of tokens the other participant bought}} \times 100\%$$

If nobody buys any tickets [*SS, RS: tokens*], each of you will have a 50% probability to receive the additional 8 points [*SS, RS: will receive 50% of the additional 8 points, that is 4 points*].

[*RS:* **The lottery**] [*RR:* **The second lottery**]

[*RS, RR:* In addition to the share of 8 points, a lottery [*RR:* In addition to the first lottery, a second lottery] will be carried out in which you will have a chance to receive additional 8 points. In this lottery, the probability that you receive the additional 8 points will be determined by the number of points that were not used to buy tokens. On the computer screen these remaining points will be represented by green squares. Each remaining point increases the probability to receive the additional 8 points by 12.5%, as shown in the table:

Points remaining	0	1	2	3	4	5	6	7	8
Probability to receive 8 points	0%	12.5%	25%	37.5%	50%	62.5%	75%	87.5%	100%

The lottery will be implemented as follows: a marker will move and stop at random either on a green square, or on an empty square. If the marker stops on the green square, you will receive additional 8 points. If it stops on an empty square, you will receive 0 points, and the additional 8 points will not be given to anyone. If the lottery determines that you receive additional 8 points, 8 green squares will be added to your round income.] [*RS:* If the lottery determines that you receive additional 8 points, 8 green squares will be added to your round income.]

[*RR:* The two lotteries that are played are independent, so you may receive a total of 16 points, 8 points or 0 points.]

Round income

[*SR:* If the lottery determines that **you** receive the additional 8 points, 8 blue squares from the box will be added to your round income. Your round income will be equal to the number of points that you did not spend buying tickets (green squares) plus the additional 8 points (blue squares). If the **other participant** receives the additional 8 points, the squares from the blue box will disappear and your round income will consist of the points that you did not spend on buying tickets (green squares).]

[*RR:* If the first lottery determines that **you** receive the additional 8 points, 8 blue squares from the box will be added to your round income. Your round income will be equal to the number of points that you did not spend buying tickets (green squares) plus the additional 8 points (blue squares). If the first lottery determines that the **other participant** receives the additional 8 points, the squares from the blue box will disappear. and your round income will consist of the points that you did not spend on buying tickets (green squares). If the second lottery determines that you receive additional 8 points, 8 green squares will be added to your round income.]

[*SR:*

- If you receive the additional 8 points, your round income is:

  Round Income = 8 —- Number of purchased tickets + 8

- If you do not receive the additional 8 points, your round income is:

  Round Income = 8 —- Number of purchased tickets]

[*RS:*

- If you receive the additional 8 points from a lottery, your round income is:

  Round Income = (Share of the additional 8 points) * 8 + 8

- If you do not receive the additional 8 points from a lottery, your round income is:

  Round Income = (Share of the additional 8 points) * 8]

[*RR:*

- If you receive the additional 8 points in the first lottery and in the second lottery, your round income will be 16 points.
- If you receive the additional 8 points only in one of the lotteries, your round income will be 8 points.
- If you do not receive the additional 8 points from either lottery, your round income will be 0 points.]

[*SS:* Your round income will be equal to the number of points that you did not spend buying tokens (green squares), plus the share of the additional 8 points (yellow sector):

Round income = 8 —- Number of purchased tokens + (Share of the additional 8 points) * 8]

At the end of a round you will see the number of tickets [*SS, RS: tokens*] the other participant bought, your probability to receive [*SS, RS: your share of*] the additional 8 points, [*SR, RS:* the outcome of the lottery] [*RR:* the outcome of the lotteries] and your round income.

End of the experiment

At the end of the experiment you will be informed about your income in the rounds that were randomly selected for payment. Income from these rounds will be added, converted into euros and paid in private once you complete a short questionnaire. Please stay seated until we ask you to come to receive the earnings.

If you have any further questions, please raise your hand now. If you have read the instructions and have no further questions, please click "Start the Experiment" on your computer screen.

Before starting Part 1 of the experiment we will ask you to complete two trial rounds. These two rounds will be the same as the task in Part 1, but income in these two rounds will not affect your final earnings.

Notes:

- The participant you are matched with will be randomly changed each round.
- One round from each 10 rounds will be randomly selected for payment.
- Part 1 will have 10 rounds. The number of rounds in other parts and additional instructions will be displayed on your computer screen.
- You can buy tickets [*SS, RS:* tokens] only using your endowment, which is equal to 8 points in each round.

I.1 Additional on-screen instructions

At the start of block 2, participants saw additional instructions on the computer screen, informing about the availability of foregone payoff information:

This is the start of **Part 2**.

In Part 2 there will be **20** rounds.

Two rounds from these 20 will be randomly chosen for payment at the end of the experiment.

The task you do will be the same as the task you did in Part 1. The only difference is that at the end of each round you will have an option to reveal information about what would have happened if you had bought a different number of [*SR, RR*: lottery tickets; *SS, RS*: tokens]. To uncover this information, click on any grey box marked with "?". Then the box will disappear and you will see what would have been the sizes of yellow and blue sectors [*SS:* and your round income] if you had bought a different number of [*SR, RR*: tickets; *SS, RS*: tokens]. [*SR, RS, RR:* You will also see what would have been the outcome of the lottery [*RR*: two lotteries] and whether or not you would have received the additional 8 points.] This additional information will have no effect on your earnings.

If you want to consult these instructions during the experiment, please raise your hand and the experimenter will bring you a paper copy.

J Instructions for Study 2

We reproduce the instructions for the SR and SRB treatments, which were identical. Changes in the SS and SSB treatments were equivalent to the changes in Study 1.

INSTRUCTIONS

Welcome to the experiment. Please read these instructions carefully. They are identical for all the participants with whom you will interact during this experiment. If you have any questions, please ask the experimenter on Zoom.

In this experiment you can earn some money. How much you earn depends on your decisions and the decisions of the other participants. During the experiment we will refer to points instead of dollars. The total amount of points that you will have earned during the experiment will be converted into dollars at the end of the experiment. The conversion rate that will be used to convert your points into your cash payment will be **1 point = $0.05**. You will also receive a $4 show-up fee.

At the end of the experiment you will see how many points you earned in each round. One round from each 10 rounds will be randomly chosen for payment: the first round will be chosen from rounds 1-10, the second round from rounds 11-20 and so on. Your earnings from the selected rounds will be added and converted into dollars.

The experiment will contain several parts. Now we will describe you the task that you will be doing in Part 1. At the start of each other part additional instructions will be displayed on your computer screen. Please read those instructions before you continue.

The task in Part 1

In Part 1 there will be 10 rounds. At the beginning of each round the computer will randomly match you with another participant in this session. The participant you are matched

with will be changed randomly each round. All participants with whom you will interact will receive the same information and will face exactly the same task.

Decision screen

In each round you will receive an endowment of 80 points. You can use these points to buy "lottery tickets". Each ticket you buy costs 1 point, so you can buy up to 80 tickets each round. Any points that you do not spend on tickets will be added to your round income.

How your decision screen will look like is shown in Figure 1. The 80 points that you receive at the start of each round will be represented by 8 green squares. You will have to choose how many of these points to use to buy lottery tickets. Every 10 additional tickets you buy will take away one green square. To make a decision, first click on the row with the closest multiple of 10 and then select the exact number of tickets. For example, if you want to buy 63 tickets, first click on the row for "60 tickets", then the screen will display all options between 56 and 65 tickets, where you can select 63.

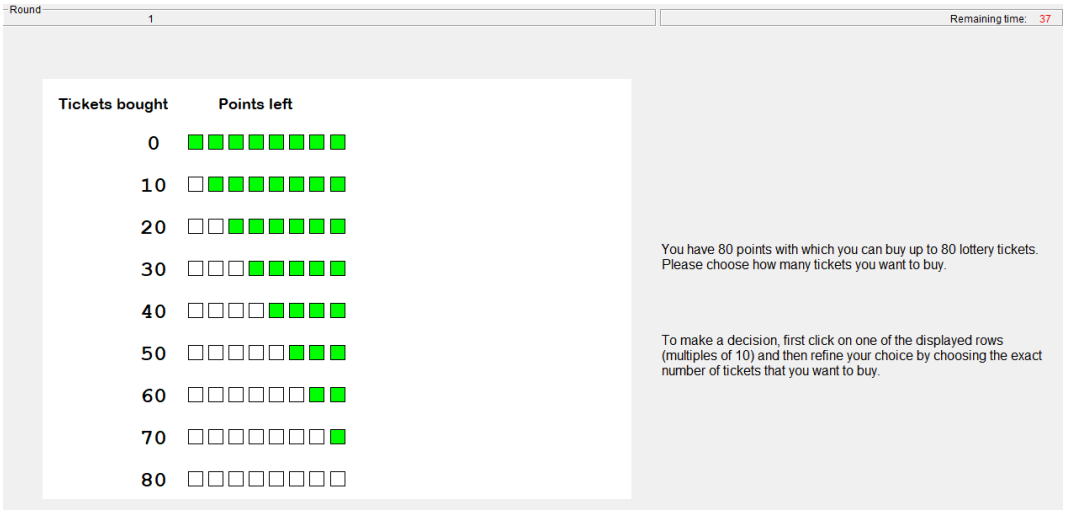


Figure 1. Screenshot of the decision screen.

The lottery

After you and the other participant have chosen how many tickets to buy, either you or the other participant will receive additional 80 points. Who receives these additional 80 points will be determined by a lottery. The lottery will be implemented the following way: after you have chosen how many tickets to buy, a box containing 8 squares will be displayed on your screen. This box represents the additional 80 points that you may receive. The box will be divided into two sectors, yellow and blue. The yellow sector belongs to you and the blue sector belongs to the other participant. Once you start the lottery, a marker will move around the box and stop at one place at random. The marker is equally likely to stop at any place. If the marker stops in the yellow sector, you will receive the additional 80 points. If the marker stops in the blue sector, the other participant will receive the additional 80 points. The size of the yellow sector represents the probability that you will receive the additional 80 points: if $x\%$ of the box is colored yellow, the probability that you will receive the 80 additional points is $x\%$.

The sizes of yellow and blue sectors are proportional to the number of lottery tickets that you and the other participant buy. For example, if you buy the same number of tickets as

the other participant, the yellow sector will be of the same size as the blue sector. If you buy twice as many tickets as the other participant, the yellow sector will be two times larger than the blue sector. Overall, the probability that you will receive the additional 80 points, represented by the size of the yellow sector, is calculated as follows:

$$\text{Probability of receiving the additional 80 points} = \frac{\text{Number of tickets you bought}}{\text{Number of tickets you bought + Number of tickets the other participant bought}} \times 100\%$$

If nobody buys any tickets, each of you will have a 50% probability to receive the additional 80 points.

Round income

If the lottery determines that **you** receive the additional 80 points, 8 blue squares that represent 80 points will be added to your round income. Your round income will be equal to the number of points that you did not spend buying tickets (green squares) plus the additional 80 points (blue squares). If the **other participant** receives the additional 80 points, the squares from the blue box will disappear and your round income will consist of the points that you did not spend on buying tickets (green squares).

- If you receive the additional 80 points, your round income is:

  Round Income = 80 —- Number of purchased tickets + 80

- If you do not receive the additional 80 points, your round income is:

  Round Income = 80 —- Number of purchased tickets

At the end of a round you will see the number of tickets the other participant bought, your probability to receive the additional 80 points, the outcome of the lottery and your round income.

Time limit

Each decision has to be made within a certain time limit to ensure that the experiment will finish on time. You will have 45 seconds to make a decision and 45 seconds to view the feedback screen. If you fail to make a decision within the allocated time, you will be assigned the same choice you made in the previous round (in the first round, a choice would be made at random). The remaining time will always be shown in the top right corner of your screen.

End of the experiment

At the end of the experiment you will be informed about your income in the rounds that were randomly selected for payment.

If you have any further questions, please ask them now. If you have read the instructions and have no further questions, please open the link that you received on Zoom and click "OK" on the screen.

Before starting Part 1 of the experiment we will ask you to complete two trial rounds. These two rounds will be the same as the task in Part 1, but income in these two rounds will not affect your final earnings.

Notes:

- The participant you are matched with will be randomly changed each round.

- One round from each 10 rounds will be randomly selected for payment.
- Part 1 will have 10 rounds. The number of rounds in other parts and additional instructions will be displayed on your computer screen.
- You can buy tickets only using your endowment, which is equal to 80 points in each round.

Notes about the online interface:

- Once you have finished reading these instructions, please go back to the Zoom call and click on the link that you received and start the experiment.
- If you accidentally close the window during the experiment or disconnect from the internet, you can reconnect by using the same link.
- For optimal experience, enter the fullscreen mode on your browser (press F11). Do not navigate to other screens during the experiment.
- Each time you see a screen with a button, you have to press the button to move to the next stage.
- Since the experiment is conducted online, you might have to wait for others to make their decisions before you move to the next stage.
- Please do not close the browser window until you complete the entire experiment and see a message saying that the experiment is finished.