A Behavioral Investigation of the FlipIt Game

ALAN NOCHENSON and JENS GROSSKLAGS, The Pennsylvania State University

Abstract: Security decisions are rarely made at a singular point in time. Though many models evaluate choice under uncertainty with regards to many options, fewer address the problem of *when* to act. The FlipIt game captures this temporal choice, only allowing players to choose when, and not how, to act. Decisions of this sort are encountered by managers choosing when to address fraud, computer users selecting when to update their software, and consumers deciding when to check their credit score. Recent investigations analyze the FlipIt game from a theoretical perspective, but the question remains - how do people actually act when given temporal decisions? To answer this question, we conduct a behavioral investigation of the FlipIt game through a Mechanical Turk experiment with over 300 participants. In our study, each participant is matched with a computerized opponent in several fast-paced rounds of the FlipIt game. We find that participant performance improves over time (however, older participants improve less than younger ones). Further, there are significant performance differences with regards to gender and an individual difference variable reflecting the extent to which individuals are inclined towards effortful cognitive activities (i.e., the need for cognition). We further vary the amount of information that participants have available about the actions of the computerized opponent with six different experimental treatments and find significant statistical effects.

Key Words: Behavioral Study, Mechanical Turk, FlipIt

1. INTRODUCTION

In the realm of security economics and policy-making, there are many factors that influence the choices that people make. Decisions such as how often to change a password or check a credit score are not without cost. People walk a fine line between performing these actions so often that they take a toll on everyday life and not performing them often enough, thus risking exposure of personal information and or identity theft. This type of problem is well-captured in the FlipIt game proposed by a team from RSA in 2012 [van Dijk et al. 2012].

In the FlipIt game, there is a red and a blue player, and a "board" that they are each interested in maintaining control of. For each unit of time that a player is in control of the board, he gains some benefit. And, conversely, when a player is not in control, he gains no benefit from the board (and may even incur some cost). Either player may "flip" the board to gain or maintain control, at some cost, as flipping while in control does not give the opponent control of the board. For example, in the case of a password reset scenario, the board is a password-protected account. Benefit is derived from using the account, and flipping the board is analogous to attempting a login (and if the login action fails it corresponds to resetting the password).

In this paper, we examine factors that affect behavior in the FlipIt game. With the data from an online experiment conducted on Amazon Mechanical Turk, we create a regression model of the relevant factors that predict outcomes for each round of the game. We examine the role of *a priori* information, learning, demographic differences, risk propensity, and Need For Cognition on behavior. We hypothesize that there is a learning effect (participants will do better in later rounds of the game), that more

Author addresses:
A. Nochenson, 309 IST Building, University Park, P.A. 16802; email: anochenson@psu.edu
J. Grossklags, 329A IST Building, University Park, P.A. 16802; email: jensg@ist.psu.edu

information will lead to better performance, and that individuals with a higher Need For Cognition (NFC) will perform better[1].

The structure of the paper is as follows. In Section 2, we present previous theoretical and behavioral work on non-cooperative games with continuous timing. We further discuss the use of Mechanical Turk for behavioral experimentation. In Section 3, we present our research methodology, experimental setup, and details about subject recruitment. The analysis of the experimental data is conducted in Section 4. We briefly discuss our results in Section 5, and conclude in Section 6. Finally, the Appendix contains the experimental instructions as they were presented to the study participants and selected additional information.

2. BACKGROUND

2.1 Games of timing in theory and experiment

Non-cooperative games with continuous timing and asynchronous decision-making have been studied intensively during the cold-war era. In the western hemisphere, scholars at the RAND corporation have been particularly active in this field starting in 1948 (e.g., Blackwell 1949). Similarly, many scientific papers have been published in Eastern European journals (and languages) in the following years (see, for example, Radzik and Orlowski 1982; Zhadan 1976). A large subsection of these studies on the inherent uncertainties of modern threats has been focused on zero-sum games called *games of timing*. The focus of these studies is not on what action to take from a pool of strategic options, but rather when the agent should take an action to get an advantage over the opponent. As such, games of timing are relevant for the study of tactical security problems.

A well-studied subclass of these games are *duels*. For example, in the Western duel, which has been dramatized in many movies, two agents start walking towards each other from relatively distant places while pointing guns at each other. The agents have a limited supply of ammunition and are aware of the fact that a shorter distance would mean a higher success likelihood in hitting the opponent. Each agent faces an obvious trade-off. When reducing the distance, accuracy is improved for both players. Waiting to shoot conserves ammunition, but one misses the opportunity to end the threatening situation early. A particularly reluctant duelist might therefore be shot by his opponent.

The theoretical contributions in this area have been surveyed and summarized by [Radzik 1996]. For our purposes, work on different information conditions is particularly relevant. Making an important distinction, [Karlin 1959] identifies two types of games of timing. The first class consists of games of complete information where the rules of the game are common knowledge and actions of the opponents are immediately known to everybody. The second class of games includes those with limited information (or to put it differently, all games not in the first set). The game we study falls into the second class.[2] More generally, in our work we are particularly interested in the influence of information about the rules of the interaction as well as the availability of information to the players at the start of the game. In our experimental treatments, we change the information that the human player has available at the start of the game about the behavior of the computerized agent.

In the domain of experimental and behavioral economics, there has been a renewed interest in non-cooperative games with continuous and asynchronous decision making. The roots of the behavioral research can be found in the 1970s. They concern a variety of games of timing, and duels. For example,

[1]NFC is a measure of "relative proclivity to process information" and "tendency to. . . enjoy thinking" [Wood and Swait 2002].

[2]As follow-up research to [van Dijk et al. 2012], the FlipIt game has been the subject of theoretical extensions. In particular, two generalizations of the FlipIt game have been proposed. Pham and Cid introduce security assessments for the defender [Pham and Cid 2012], while Laszka et al. propose a multi-resource model [Laszka et al. 2013].

Kahan and Rapoport studied duels with symmetric and asymmetric accuracy functions and number of bullets, respectively [Kahan and Rapoport 1974; Kahan and Rapoport 1975].

The study of human behavior in continuous time environments have also been undertaken in different contexts. For example, Friedman et al. observe that convergence in such environments (with limited information) can fail even when iterated deletion of dominated strategies would theoretically lead to the Nash equilibrium [Friedman et al. 2004]. Another recent investigation is provided by Rapoport and Murphy who study trust games with many players in continuous time environments [Rapoport and Murphy 2012]. Brunnermeier and Morgan study multi-player games where agents receive private signals about a payoff-relevant state variable. At the same time, an individualized desynchronized clock is started. To perform well in the game, agents have to predict other agents' clock times subject to different information conditions and the number of players [Brunnermeier and Morgan 2010].

2.2 Mechanical Turk & Experiments

Amazon Mechanical Turk is a service that was launched in 2005 in order to allow "Requesters" to outsource Human Intelligence Tasks (HITs) to Mechanical Turk workers (Turkers). The service can be used for a variety of tasks including conducting surveys (e.g., Felt et al. 2012) and behavioral studies (e.g., Christin et al. 2012). Mechanical Turk is a useful tool for conducting research because the payment/quality ratio per-subject of participants is lower than it is typical with traditional laboratory studies and the demographic mix of participants is likely more diverse than university student convenience samples [Kam et al. 2007].

The payment/quality ratio being lower on Mechanical Turk means that researchers are able to obtain a large sample at a small cost (and generally in a short amount of time). The demographic structure of Amazon Mechanical Turk is a topic that has been investigated by a number of authors in the last few years (e.g., Ross et al. 2010; Ipeirotis 2009; Mason and Suri 2012). These studies have shown that the country-of-origin for Turkers is nearly half based in the United States and half based in India, with small representations from other countries. Slightly over half of workers are female and the median age of Turkers is around 30 years [Mason and Suri 2012]. This demographic diversity makes Mechanical Turk an attractive platform for recruiting subjects and conducting experiments.

A number of authors have conducted experimental economics experiments on Mechanical Turk in recent years. For example, Horton used Mechanical Turk to replicate three experiments that have been extensively studied in economic laboratories [Horton et al. 2011]. One of the replicated experiments was for participants to play a one-shot prisoner's dilemma game (with payments one-tenth the size online as in a physical lab). The authors found no significant differences in behavior between the traditional and online versions of the study. Each replication that was conducted was completed in fewer than 48 hours and cost less than $1 per subject on average. Despite the low stakes and extreme anonymity of MTurk, the subjects' behavior was consistent with findings from the standard laboratory [Horton et al. 2011].

3. METHODOLOGY

In this paper, we conduct an experiment in the tradition of experimental and behavioral economics [Grossklags 2007]. In contrast to many experiments, we did not bring experimental subjects into a physical laboratory.[3] Instead, we utilized Amazon's Mechanical Turk service to run an online experiment. The experiment was set up much like common laboratory experiments, and was approved by the

[3]See, for example, Grossklags et al. for a security-related experiment conducted in a physical laboratory [Grossklags et al. 2008].

University Office of Research Protection's Institutional Review Board.

At the beginning of the study, participants were presented with a consent form which detailed the procedures that they were to follow, the structure of payments, and a number of other pieces of pertinent information. After consenting to the terms of the experiment, participants were redirected to an instructions page. This page stated the rules to the game and also described an example game. Further, we included a more detailed description of the payment structure. The instructions page is included in the Appendix.

Once participants consented to take part in the experiment and read the rules of the game, they were redirected to a survey questionnaire with four parts. The first part of the survey asked participants basic demographic information, including their age, gender, level of education, and country of origin. The next three parts of the survey were presented in randomized order. One part was a set of integrity check questions which participants were required to correctly answer to continue to play the game.[4] The other two sections in the survey were psychometric scales[5] that assessed the level of risk propensity (from [Meertens and Lion 2008]) and "need for cognition" (from [Wood and Swait 2002]) of participants. Participants were required to answer every question on the survey before they were allowed to play the game.

After successfully completing the survey, participants were redirected to the main page of the experiment. On this page, there is a board showing the actions and results of the FlipIt game, buttons to start a round, and a button to "flip" the state of the board[6]. Additionally, participants were given the option to revisit the rules of the game.

3.1 Participant pool

We restricted the pool of participants to include only Mechanical Turk users based in the United States who had an approval rating of over 90%. We put these restrictions in place to ensure that the subject pool was as minimally contaminated as possible by variables other than those of interest (i.e., we did not aim to compare data for different countries of origin) and that there was minimal noise (from participants who frequently had their work rejected for poor quality). The experiment was run as a number of distinct Mechanical Turk Human Intelligence Tasks (HITs). Participants were not allowed to participate in multiple HITs.

[4]The types of questions that we call *integrity questions* are also sometimes referred to as "screeners" in the language of Berinsky et al. [2012]. They aim to ensure that participants are paying attention during the survey (and hopefully during the remainder of the experiment). According to Berinsky et al. [2012], many articles that use these types of checks simply exclude participants who fail them in an effort to reduce noise. Following the recommendations made in that paper, we decided not to throw out failed responses and to instead highlight to "shirkers" that they missed an integrity question. If a participant missed an integrity question they were redirected back to the survey and could not move on until the integrity questions were answered correctly.

[5]Psychometric scales measure psychological constructs, usually with regards to individual differences. The scales that were used in this paper were a series of Likert-style questions that were aggregated to yield a measure of the construct desired (here risk propensity and need for cognition). While there are a number of other ways to measure these constructs (such as the Iowa Gambling Task [Bechara et al. 1994] for Risk Propensity), we felt that introducing other non-survey activities into the process would distract from the task at hand. Therefore, despite the additional confidence that may have been found using these other types of tests, we opted for the simpler and less distracting scales.

[6]The system for playing the game was originally conceived by Ethan Heilman (https://github.com/EthanHeilman/flipIt) and was heavily modified and extended for the purposes of this paper.

3.2 Rounds

For the purposes of this experiment, participants played six rounds of the FlipIt game that lasted 20 seconds each. The first round was introduced as a "practice" round. In the practice round, participants were not eligible for a bonus payment. The start of the five experimental rounds was signaled to the participants with a pop-up message. The only difference between the practice round and the non-practice rounds is the presence of a possible bonus payment which was awarded to the participants according to their performance. Participants were informed of these rules on the instructions page. Further, we restated these facts to them at the beginning of the practice and experimental rounds, respectively.

Our experimental setup, involved each human participant in a relatively fast-paced version of the FlipIt game. However, we expected that the round length of 20 seconds would provide participants with enough time to develop an appropriate strategy against the computerized opponent. We used five paid rounds to allow players enough time to improve their performance in the game and a single practice round in which to experiment without penalty. In future work, we also intend to study the performance of human players in rounds with a longer duration and overall slower game-play (i.e., a slower computerized player).

3.3 Payment

Participants earned a show-up fee, a, for completing the study irrespective of their performance in the game. They played $n = 5$ experimental rounds numbered $1...n$ and a single practice round numbered 0. Participants were paid according to the point difference between the points awarded for their own performance and the points awarded to the computerized opponent for its performance. Let the point difference in round i be known as δ_i. That is, if a player won by 200 points in round 1 then $\delta_1 = 200$ and if he loses by 900 points in round 2 then $\delta_2 = -900$. Let e be the exchange rate for points (i.e. the monetary value of a single point in dollars). Let the per-round endowment be x_i. The purpose of the endowment was to allow participants to experience relative losses.

The payment function for a single participant is as follows:

$$\text{Bonus payment} = e \sum_{i=1}^{n} \max(x_i + \delta_i, 0) \tag{1}$$

$$\text{Total payment} = a + \text{Bonus Payment} \tag{2}$$

The practice rounds are not included in the payments above. In this experiment, we set the exchange rate $e = 0.0001$ which corresponds to 100 points $= \$0.01$. The per-round endowment was set at $x_i = 1000$ points $\forall\ 1 \leq i \leq n$. I.e., in a tied game the participant would earn \$0.10.

The total payment above was paid to the participant in two installments. Participants first accepted a Mechanical Turk Human Intelligence Task (HIT) that paid \$0.50 upon successful completion. This corresponds to the show-up fee a above. After the experiment was completed, participants were paid a bonus payment through the Mechanical Turk system equal to the remainder of the total payment (Equation (1) above).

3.4 Strategy of the Computer Player

Regardless of the treatment that a participant was assigned to, each participant faced an opponent playing a fixed (non-adaptive) periodic strategy. The opponent's flip rate and time of first flip (anchor) changed in every round of the game, but the overall strategy of the opponent did not. Both values were drawn from uniform distributions before a new round started. Flip rates ranged from 1 to 5 seconds, and the anchor ranged form 0.1 to 4.1 seconds.

In the instructions, we did not inform the human partipants that they would be paired with a computerized player. From the information treatments (which we discuss below), the human participants could have, however, infered that they are playing against a non-human opponent. Previous research has shown that varying the information about the type of opponent player (i.e., human or computerized) can impact the strategies and outcomes in a competitive game (see, in particular, Grossklags and Schmidt 2006). We did not explicitly vary such information in our experiment.

3.5 Treatments

The treatments that we used in this experiment varied the specific amount of *a priori* information about the computer player's strategy given to the human player. When participants began the practice round, they were randomly assigned to a treatment that persisted throughout all rounds of the game.[7] We list the six treatments below:

(1) Participants are given no information about the game. They are told that their opponent may adapt to moves made.
(2) Participants are told that their opponent is non-adaptive (i.e., it is playing a fixed strategy).
(3) Participants are told that their opponent is playing a periodic strategy.
(4) Participants are told that their opponent is non-adaptive and they are told his average flip rate in each round (however, they are not told the computer is playing a periodic strategy).
(5) Participants are told that their opponent is playing a periodic strategy and they are told his flip rate in each round.
(6) Participants are told that their opponent is playing a periodic strategy and they are told his flip rate and time of first flip in each round.

While the treatment assignment was permanent for each participant, the flip rate of the computer player and time of first flip (anchor) were not. In each round of the game, a new flip rate and anchor were drawn from uniform distributions. If the treatment required this information, it was shown to the participant under the heading "Important Information About Your Opponent".

In their paper, van Dijk et al. [2012] support the hypothesis that information conditions might have an important effect on play. They discuss two types of pre-game information that can be given to players, known as Rate of Play [RP] and Knowledge of Strategy [KS]. RP is given in information treatments 4, 5 and 6. KS is given (at least partially) in all treatments except for 1. They suggest that these pre-game information conditions are "meaningfully applicable" only to players that are unable to estimate the rate of play or the strategy of their opponents reasonably quickly during the game.

Through the different experimental treatment conditions, we aim to identify the differences in average play by human participants given the pre-play information given to them. We keep the duration of

[7] A tabular version of these treatments can be found in Table 12 in the appendix.

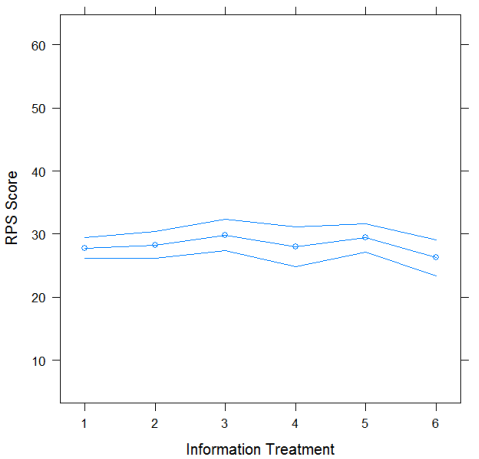


Fig. 1. Risk Propensity across Information Treatments

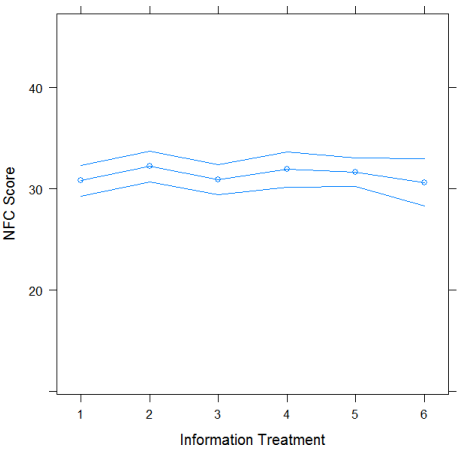


Fig. 2. Need for Cognition across Information Treatments

each round constant across all treatments and participants (i.e., 20 seconds). We expect that in longer games the play of human participants would almost always converge to the optimal response strategy (i.e., to flip just after the periodic play of the computer player). However, we defer this investigation to future work.

4. RESULTS

4.1 Descriptive results

Participation in our study was restricted to Mechanical Turk users from the United States. We ended up with 310 unique participants who played a total of 1860 rounds. Male participants made up 214 (69%) of the participants, while 96 (31%) were female. The mean age of participants was 29.5 (sd = 9.8), with less than 15% of participants being older than 40 years. 36.77% of participants had completed "Some college" and 39.03% had obtained a four year college degree. Risk propensity scores (RPS) were normally distributed (Shapiro-Wilk test, $p < 0.001$) around the mean of 28 with sd$= 10.18$ (the scale's maximum is 63). The need for cognition (NFC) scores of the participants were normally distributed (Shapiro-Wilk test, $p < 0.001$) around the mean of 31 and sd$= 7.12$.

This demographic data is somewhat different compared to the results reported by previous studies conducted on Mechanical Turk (see Ross et al. [2010] and Ipeirotis [2009]); i.e., less women participated in our study. But they are similar with respect to age and education levels. The reasons for the disagreement may be that the title or description of the task enticed more male Turkers to participate. Alternatively, the times at which the Mechanical Turk HITs were posted (usually late night) may be times that tend to draw more male participants.

We found that male participants tend to have a higher RPS score (One-Tailed Wilcoxon Rank Sum Test; $p < 0.001$), indicating that they are more risk-seeking. In contrast, female participants had a higher (but not significant) NFC score which suggests that they tend to enjoy problem solving and information processing on average slightly more than their male counterparts. This can also be seen

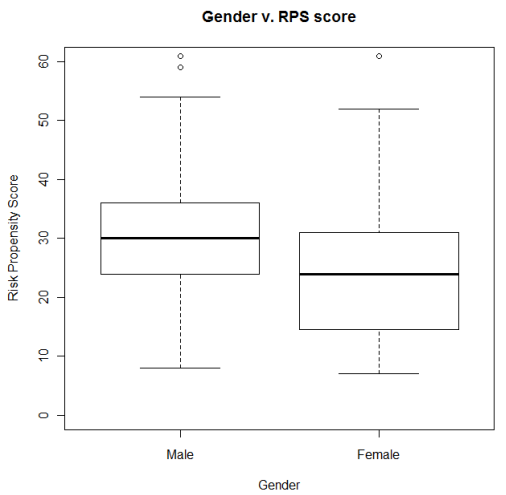


Fig. 3. Risk Propensity broken down by Gender

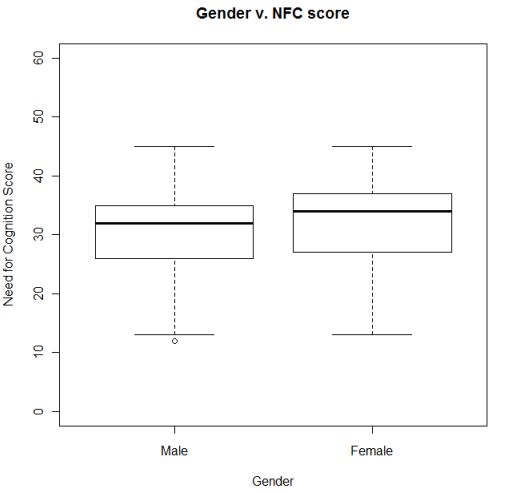


Fig. 4. Need for Cognition broken down by Gender

in Figs. 3 & 4.

Total bonuses paid to participants had a mean value of $0.58 (sd=$0.14). The minimum bonus that any participant earned was $0.10 and the maximum was $1.29. For reference, by tying in every round with the computerized opponent, a participant would have earned $0.50. Per-round payments had a mean of $0.11 and standard deviation of $0.05. The minimum round payment that any participant earned was $0.00 and the maximum was $0.28. For reference, the initial endowment for each round was $0.10, meaning that a tie in a given round paid a round payment of $0.10. Participants could exhaust their endowments but were not permitted to earn negative money in any round. Therefore, if a participants score difference was less than the per-round endowment in a round (i.e $x_i + \delta_i < 0$) then a participant earned $\$0.00$ in that round. Out of the 1860 rounds, 68 rounds (3.66%) resulted in earnings of $\$0.00$.

In order to better illustrate the behavior of our participants, we discuss here the case of two players; a “good” one and a “bad” one.

The good player that we look at is a 34 year old college-educated male with a high need for cognition (41). He was assigned to information treatment 4, meaning that he knew his opponent is non-adaptive and he was given the average flip rate of the opponent in every round. Despite facing an opponent whose flip rate varied from flipping every 1.61 seconds to every 3.64 seconds, this player managed to earn at least 0.20 in every round (i.e., he beat the computer player by 1000 points or more in every round).

The bad player on the other hand had much less success. This individual is a 24 year old female with some college education and a low need for cognition (20). She lost in every round that she played, and even lost by more than 1000 points in 2 of the non-practice rounds.

Abbreviation	Description	Estimate	Std. Error	t value
RP	Round Payment (in $)			
	(Intercept)	$9.213 * 10^{-2}$	$1.615 * 10^{-2}$	5.704
NFC	Need for Cognition (NFC) score	$4.149 * 10^{-4}$	$1.981 * 10^{-4}$	2.095
FEM	Gender (1=Female, 0=Male)	$-4.194 * 10^{-3}$	$3.014 * 10^{-3}$	-1.392
$LAGE$	Age (in years) (log transformed)	$-1.129 * 10^{-2}$	$4.769 * 10^{-3}$	-2.367
PER	Periodicity of computer player (how often he flips in 1/100 s)	$1.188 * 10^{-4}$	$8.859 * 10^{-6}$	13.416
ANC	Time of computer first click (anchor; in 1/100 s)	$4.598 * 10^{-5}$	$8.754 * 10^{-6}$	5.252
TRE_2	Is treatment 2? (1=Yes, 0=No)	$3.011 * 10^{-3}$	$4.440 * 10^{-3}$	0.678
TRE_3	Is treatment 3?	$1.453 * 10^{-2}$	$4.081 * 10^{-3}$	3.560
TRE_4	Is treatment 4?	$1.252 * 10^{-2}$	$5.028 * 10^{-3}$	2.490
TRE_5	Is treatment 5?	$5.750 * 10^{-3}$	$4.288 * 10^{-3}$	1.341
TRE_6	Is treatment 6?	$1.445 * 10^{-2}$	$5.288 * 10^{-3}$	2.733
$TRE_1 * RNC$	Round effect in Treatment 1	$1.942 * 10^{-2}$	$7.969 * 10^{-3}$	2.437
$TRE_2 * RNC$	Round effect in Treatment 2	$1.949 * 10^{-2}$	$8.043 * 10^{-3}$	2.423
$TRE_3 * RNC$	Round effect in Treatment 3	$1.768 * 10^{-2}$	$7.984 * 10^{-3}$	2.214
$TRE_4 * RNC$	Round effect in Treatment 4	$2.122 * 10^{-2}$	$8.070 * 10^{-3}$	2.630
$TRE_5 * RNC$	Round effect in Treatment 5	$2.076 * 10^{-2}$	$8.136 * 10^{-3}$	2.552
$TRE_6 * RNC$	Round effect in Treatment 6	$2.493 * 10^{-2}$	$8.207 * 10^{-3}$	3.038
$LAGE * RNC$	Round effect by log(Age)	$-5.401 * 10^{-3}$	$2.365 * 10^{-3}$	-2.284
RNC	Round Number (0 is practice, 5 is last round) (centered around the average)			

Fig. 5. Regression results along with abbreviations and their descriptions.

4.2 Regression model

After exploratory data analysis, we developed the following regression model:

$$RP = \beta_0 + \beta_1 * NFC + \beta_2 * FEM + \beta_3 * LAGE + \beta_4 * PER + \beta_5 * ANC + \\ + \beta_6(LAGE * RNC) + \beta_{TRE=t}(TRE_t) + \beta_{RNC,TRE=t}(TRE_t * RNC) + \epsilon \quad (3)$$

In this model, we fitted a random intercept and a random slope grouped by session id (which is a unique identifier for each set of rounds a participant played) as well as multiple fixed effects to explain variance in the per-round payment. Since this was a linear mixed-effects model[8], it is difficult to estimate the number of degrees of freedom, and therefore p-values are not reported in the above table. However, a rule of thumb is that t-values of greater than $|2|$ indicate significance at a 0.05 level, and greater than $|2.576|$ indicate significance at a 0.01 level, respectively [Koufteros 1999].

In this regression, we excluded all data from the practice round (round 0), since it did not count towards the participants' earnings and they were aware of this fact. Therefore, the practice round was a safe place for participants to experiment without fear of negative outcomes. (Comparing the results of the regression with the practice round removed to the results with the practice round included did not have an impact on which variables and interactions were significant.)

In order to examine the regression from a somewhat neutral perspective, we took the following steps: 1) we removed a single participant that did not specify his age and 2) we removed rounds where pay-

[8]See [Bates 2010] for more details on the construction and use of linear mixed-effects models. The R code used to generate this model is located in the appendix.

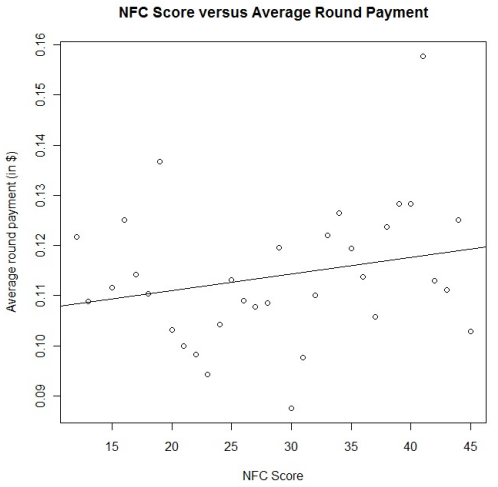


Fig. 6. Need for cognition score versus average round payment

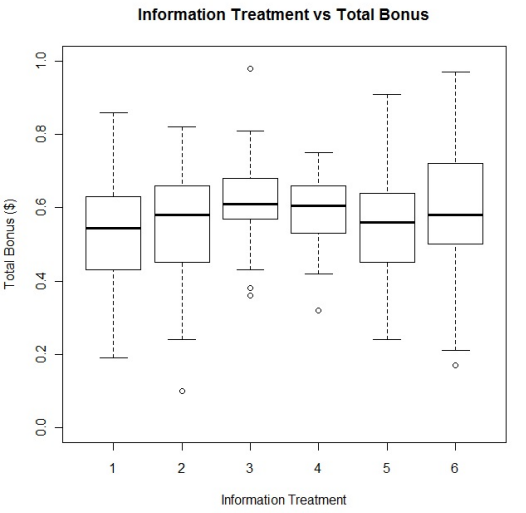


Fig. 7. Information Treatment v Total Bonus.

ment was $0.00, since this is not a true measure of performance ($0.00 is the value assigned when a participant loses their initial endowment and possibly more), and we centered the round number variable (RNC).[9] After centering the variable, the covariance among the variables decreased due to the apparent centering issue (i.e., it shows that the observed problem was not an issue of covariance).

In the following, we discuss the findings from our regression analysis.

4.3 Higher NFC leads to higher round payments

The need for cognition score (NFC) is one of the factors that we found to be a significant predictor of performance in a round, as measured through the round payment (see Fig. 6). The results show that individuals with higher NFC scores tend to do better than individuals with lower NFC scores. The regression coefficient for the NFC score is $4.149 * 10^{-4}$, which means that the effect is weak (each point higher leads to an increase in 4/100 of 1 ¢), but significant. This effect is consistent with our prediction that individuals who like to think more will do better in a given round. However, since this model also includes round payments for early rounds, the effect is small. We would expect to see a larger effect of NFC if we had restricted data to only the last round of the game (i.e., once the high NFC participants had time to actualize their learning potential).

4.4 Risk propensity has no impact on round payment

The measure of risk aversion (the RPS scale) is derived from a scale by Meertens and Lion [2008]. Ex ante, we believed that risk propensity should be of little relevance to the current version of the game, since the computer player was following a fixed periodic strategy. We decided not to include risk propensity in the regression model above due to the size of the effect compared to other effects. However, this does not mean that there is not a risk effect in performance and we have followed this path in a related paper conducting a secondary analysis of the data. In an in-depth study of indi-

[9]Centering here means moving the data around the mean, such that $x_{new} = x_{old} - x_{avg}$, where x here refers to the round number.

viduals who performed well in the experiment (i.e., who made positive earnings above the per-round endowments), we found that risk seeking affects participants timing in early and late rounds of the game [Reitter et al. 2013]. We also expect that risk propensity will be important in explaining behavior in a game with human players against computer players with non-deterministic or adaptive strategies.

4.5 More information is generally better

As we can see from the regression results above, there is a somewhat positive trend between being placed in a higher treatment number and earning an increased payoff in a given round. However, the additional information provided in the different treatments is not unambiguously rankable in a straightforward ordinal fashion, and this is reflected in the performance by the participants (see Fig. 7). When inspecting the nature of the information given in the treatments more closely, we can determine that an information treatment i has unambiguously more information than information treatment $i-2$. However, it is not necessarily the case that it gives the user more information than treatment $i-1$. Therefore, we would not expect a linear increase, but a near-linear one, which is what the regression and Fig. 7 show.

4.6 Male participants perform better than female partipants

The regression output above shows that there is a marginally significant and weak gender effect on round payment. That is, women make 0.42¢ less than men per round holding all else equal. However, in aggregate these differences add up. Examining the same effect for the average total bonus we observe that men earn significantly more than women, with \$0.5621 and \$0.5249, respectively (one-tailed t-test $p < 0.05$).

4.7 Learning effect

We expected that participants would perform better in later rounds. We observe that the learning effect is quite variable between rounds, but generally positive. This can be seen in Figure 8. To further illustrate this observation, we included in the regression model a variable for the round in which data was collected and further studied the interaction between performance in later rounds and the treatment conditions.

4.7.1 *The learning effect is dependent on information treatment.* In addition to the overall learning effect, we observe that the effectiveness of learning also depends on the treatment condition. Specifically, learning in treatments 1 and 2 is about the same. There is less learning in treatment 3. Treatments 4 and 5 have about the same amount of learning (but greater than 1-3) and treatment 6 has the greatest amount of learning. Therefore, there is a near-linear increase in the amount of learning as we provide more information to the participants, which was expected. That is, having more *a priori* information about the computerized player should help individuals to better understand the impact of their actions and to develop a more optimal timing of their flips. See Fig. 9 for a graphical representation of the data.

4.7.2 *Older participants learn less well than younger ones.* As we see from the regression coefficients, with increasing age participants earn less of a bonus in the game. Further, older participants learn less from one round to another as shown by the interaction effect between age and the round numbers. This can also be observed in Figs. 10 & 11.

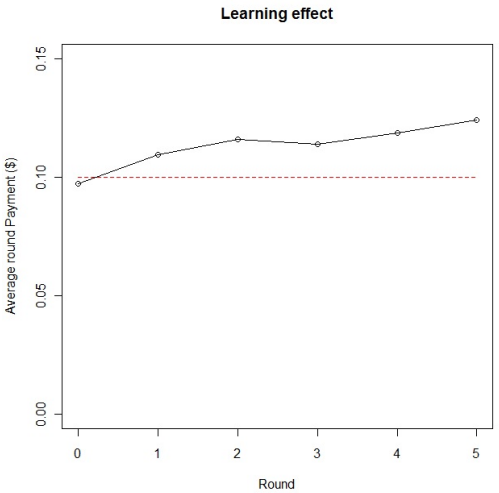


Fig. 8. Learning effect (aggregated over information treatments)

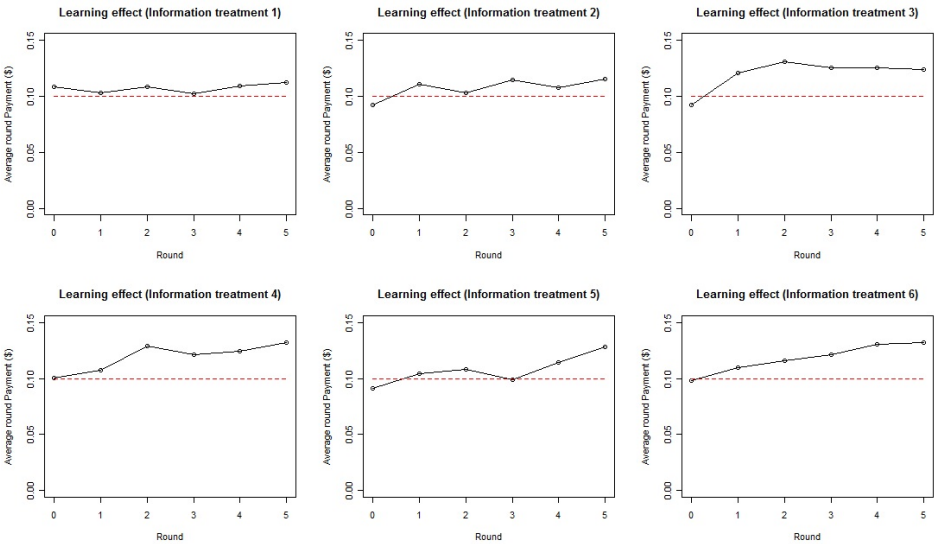


Fig. 9. Learning effect in different information treatments.

5. DISCUSSION

Rand [2012] raised a number of concerns about using Mechanical Turk for experimental studies, including concerns about participant attention, trust in experimental instructions, non-random attrition, and independence of observations. To address the problem of participant attention, we included integrity check questions in the pre-experiment survey, and while this does not completely address the problem, it at least ensures that participants are not being careless. The concern over trust in experimental instructions is one that we attempt to address by including a consent form with our names and university associations in order to establish our credibility. However, as we found in this exper-

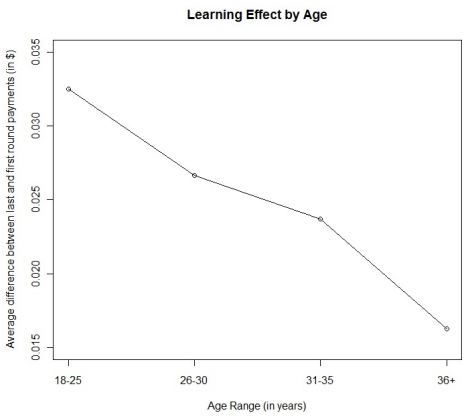


Fig. 10. Learning effect in different age groups. Age ranges were chosen to balance number of participants in each age range.

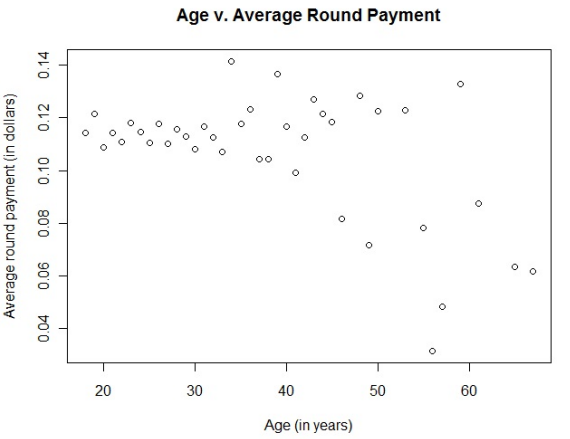


Fig. 11. Age vs. Average Round Payment. Note: Older participants are less well represented.

iment, it may be the case that participants do not fully use the instructions that are given to them. The problem of attrition was non-existent in this experiment. And, while we cannot be entirely sure that participants were unique (and thus observations independent), we did require that a user with a specific Mechanical Turk ID could not perform the task more than once.

We hope that this experiment does not only yield insights into the performance of humans in the FlipIt game, but that it also shows some best practices for conducting behavioral research on Mechanical Turk. We further received very positive feedback from our participants through a number of channels, including a comment box on the Mechanical Turk HIT and posts on Reddit[10]. These Reddit posts serve as a medium for Mechnical Turk workers to discuss tasks and requesters (i.e., experimenters) that they like or disklike. Comments on the posts included, "really good one," and "[t]his was awesome." Overall, participants seemed to both enjoy the game and the compensation they received.

6. CONCLUSION

The FlipIt game is a valuable way to evaluate performance of people in situations such as checking their accounts for security compromises or checking their credit score. We examined a number of factors that we anticipated to have an impact on performance in this game. Younger participants, male participants, and those with a higher NFC score performed better. Risk propensity did not have a significant observable impact considering the overall sample of participants, however, in a follow-up secondary data analysis we found that successful participants were impacted by their risk-seeking behavior [Reitter et al. 2013]. There was a learning effect, and the learning effect was different depending on the amount of pre-play information available to the participants. Information treatments had a significant, but somewhat non-linear, effect.

[10] http://www.reddit.com/r/HITsWorthTurkingFor/comments/1965fp/us_play_a_game_jens_grossklags_50_bonus5min_90/ as well as http://www.reddit.com/r/HITsWorthTurkingFor/comments/197ma1/usplay_a_game_and_and_earn_050200jens/

There are a number of areas for further research. One next step would be to have participants play against opponents with adaptive and/or non-deterministic strategies. Additionally, we are in the process of using the data collected from this experiment to train a cognitive model to play the FlipIt game (see [Reitter et al. 2013] for initial results). There is also space to conduct these types of experiments with other security games, both online (through Mechanical Turk) and offline in a traditional laboratory.

7. ACKNOWLEDGMENTS

We would like to thank David Reitter for helpful discussions and technical assistance with the regression model. In addition, we appreciate the detailed feedback from the anonymous reviewers.

REFERENCES

Douglas Bates. 2010. *lme4: Mixed-effects modeling with R*. http://lme4.r-forge.r-project.org/book/Ch1.pdf

Antoine Bechara, Antonio Damasio, Hanna Damasio, and Steven Anderson. 1994. Insensitivity to future consequences following damage to human prefrontal cortex. *Cognition* 50, 1–3 (April-June 1994), 7–15.

Adam Berinsky, Michele Margolis, and Michael Sances. 2012. Separating the Shirkers from the Workers? Making Sure Respondents Pay Attention on Internet Surveys. In *NYU CESS 5th Annual Experimental Political Science Conference*.

David Blackwell. 1949. *The noisy duel, one bullet each, arbitrary accuracy*. Technical Report. The RAND Corporation, D-442.

Markus Brunnermeier and John Morgan. 2010. Clock games: Theory and experiments. *Games and Economic Behavior* 68, 2 (2010), 532 – 550.

Nicolas Christin, Serge Egelman, Timothy Vidas, and Jens Grossklags. 2012. It's all about the Benjamins: An empirical study on incentivizing users to ignore security advice. *Financial Cryptography and Data Security* (2012), 16–30.

Adrienne Porter Felt, Serge Egelman, and David Wagner. 2012. *I've got 99 problems, but vibration ain't one: A survey of smartphone users' concerns*. Technical Report. Technical Report UCB/EECS-2012-70.

Eric Friedman, Mikhael Shor, Scott Shenker, and Barry Sopher. 2004. An experiment on learning with limited information: Nonconvergence, experimentation cascades, and the advantage of being slow. *Games and Economic Behavior* 47, 2 (May 2004), 325–352.

Jens Grossklags. 2007. Experimental Economics and Experimental Computer Science: A Survey. In *Proceedings of the Workshop on Experimental Computer Science (ExpCS'07)*.

Jens Grossklags, Nicolas Christin, and John Chuang. 2008. Predicted and observed user behavior in the weakest-link security game. In *Proceedings of the 2008 USENIX Workshop on Usability, Psychology, and Security (UPSEC'08)*.

Jens Grossklags and Carsten Schmidt. 2006. Software Agents and Market (In)Efficiency - A Human Trader Experiment. *IEEE Transactions on System, Man, and Cybernetics: Part C* 36, 1 (Jan. 2006), 56–67.

John Horton, David Rand, and Richard Zeckhauser. 2011. The online laboratory: Conducting experiments in a real labor market. *Experimental Economics* 14, 3 (Sept. 2011), 399–425.

Panagiotis Ipeirotis. 2009. Turker Demographics vs. Internet Demographics. (2009). http://www.behind-the-enemy-lines.com/2009/03/turker-demographics-vs-internet.html

James Kahan and Amnon Rapoport. 1974. Decisions of timing in bipolarized conflict situations with complete information. *Acta Psychologica* 38 (1974), 183–203.

James Kahan and Amnon Rapoport. 1975. Decisions of timing in conflict situations of unequal power between opponents. *Journal of Conflict Resolution* 19 (1975), 250–270.

Cindy Kam, Jennifer Wilking, and Elizabeth Zechmeister. 2007. Beyond the "Narrow Data Base": Another Convenience Sample for Experimental Research. *Political Behavior* 29, 4 (2007), 415–440.

Samuel Karlin. 1959. *Mathetical methods and theory in games, programming, and economics*. Addison-Wesley. 250–270 pages.

Xenophon Koufteros. 1999. Testing a model of pull production: A paradigm for manufacturing research using structural equation modeling. *Journal of Operations Management* 17, 4 (1999), 467–488.

Aron Laszka, Gabor Horvath, Mark Felegyhazi, and Levente Buttyan. 2013. *FlipThem: Modeling Targeted Attacks with FlipIt for Multiple Resources*. Technical Report. Budapest University of Technology and Economics.

Winter Mason and Siddharth Suri. 2012. Conducting behavioral research on Amazon's Mechanical Turk. *Behavior Research Methods* 44, 1 (2012), 1–23.

Ree Meertens and Rene Lion. 2008. Measuring an Individual's Tendency to Take Risks: The Risk Propensity Scale. *Journal of Applied Social Psychology* 38, 6 (2008), 1506–1520.

Viet Pham and Carlos Cid. 2012. Are We Compromised? Modelling Security Assessment Games. In *Proceedings of the Fourth Conference on Decision and Game Theory for Security (GameSec)*. 234–247.

Tadeusz Radzik. 1996. Results and Problems in Games of Timing. *Lecture Notes-Monograph Series, Statistics, Probability and Game Theory: Papers in Honor of David Blackwell* 30 (1996), 269–292.

Tadeusz Radzik and Krzysztof Orlowski. 1982. A mixed game of timing: Investigation of strategies. *Zastosowania Matematyki* 17, 3 (1982), 409–430.

David Rand. 2012. The promise of Mechanical Turk: How online labor markets can help theorists run behavioral experiments. *Journal of Theoretical Biology* 299 (April 2012), 172–179.

Amnon Rapoport and Ryan Murphy. 2012. Evolution and Breakdown of Trust in Continuous Time. *The Oxford Handbook of Economic Conflict Resolution* (2012).

David Reitter, Jens Grossklags, and Alan Nochenson. 2013. Risk-Seeking in a Continuous Game of Timing. In *Proceedings of the 12th International Conference on Cognitive Modeling (ICCM)*.

Joel Ross, Lilly Irani, Six Silberman, Andrew Zaldivar, and Bill Tomlinson. 2010. Who are the crowdworkers?: Shifting demographics in Mechanical Turk. In *Proceedings of the 28th of the International Conference on Human Factors in Computing Systems (Extended Abstracts)*. 2863–2872.

Marten van Dijk, Ari Juels, Alina Oprea, and Ronald Rivest. 2012. *Flipit: The game of "stealthy takeover"*. Technical Report. Cryptology ePrint Archive, Report 2012/103, 2012. http://eprint. iacr. org.

Stacy Wood and Joffre Swait. 2002. Psychological indicators of innovation adoption: Cross-classification based need for cognition and need for change. *Journal of Consumer Psychology* 12, 1 (2002), 1–13.

Vitaliy Zhadan. 1976. Noisy duels with arbitrary accuracy functions. *Issledovanye Operacity* 5 (1976), 156–177.

APPENDIX

Fig. 12. Information treatments and the information given in each treatment

Treatment	I1 Non-adaptive	I2 Periodic	I3 Average rate of play (α)	I4 Anchor
1				
2	x			
3	x	x		
4	x		x	
5	x	x	x	
6	x	x	x	x

R code used for generating the regression model

```
data <- read.csv('data.csv')
library(lme4)
lmer(round_payment~(nfctotal+gender+log(age)+tick+anchor+(treatment_id*centerData(round_num))
 + (log(age)*centerData(round_num)) - centerData(round_num) + ((1+treatment_id)|session_id)),
 data=subset(data, age>0 & round_num>0 & round_payment>0))
```

Derivation of when flips are useless

Need to find when the score difference (Δ) between the player and computer is the same for playing flip (and maintaining control for some period of time) as not playing flip (and not incurring the flip cost).
Let x be the number of 1/100 seconds that a player is considering flipping before the opponent's next flip.
Utility functions are only for a given period, that is from one opponent flip to the next.

$$\begin{aligned}
\Delta(\text{do nothing}) &= \Delta(\text{Flip x seconds before opponent's next flip}) \\
u_{opp}(\text{no flip}) - u_{player}(\text{no flip}) &= u_{opp}(\text{player flip}) - u_{player}(\text{player flip}) \\
(\text{flip period} - \text{flip cost}) - 0 &= (\text{flip period} - \text{flip cost} - x) - (x - \text{flip cost}) \\
\text{flip period} - 100 &= (\text{flip period} - 100 - x) - (x - 100) \\
\text{flip period} - 100 &= \text{flip period} - 100 - 2x + 100 \\
-100 &= -2x \\
x &= 50
\end{aligned}$$

Therefore, regardless of the flip period, a player does not gain anything from flipping half a second before the opponent's next flip.

Experimental Instructions

Basic Rules

You will be playing multiple rounds of a two-player game called FlipIt. The objective of FlipIt is to gain and maintain possession of the game board. Until you take an action, the state of possession of the game board is hidden from your view. In this state, the board is shown in gray color.
The only action you have available is to 'flip' the game board. When you flip the board, it will be shown to you who had possession of the game board until this very moment. This information will only be shown to you and not your opponent. At the same time, you also gain possession of the board, or maintain possession if you already owned the board.
The same rules apply to your opponent. That is, you cannot observe if and when the opponent flipped the board in the past, until you take the action to flip the board yourself.
Below, we break down the rules in more detail.

Detailed Rules

- **Points**

  You gain **100** points per second that you are in control.
  You earn **0** points while your opponent is in control.
  You pay **100** points when you play 'flip'.

  You begin with **0** points. Scores are updated when you play a 'flip' and at the end of the game.

- **Moves**

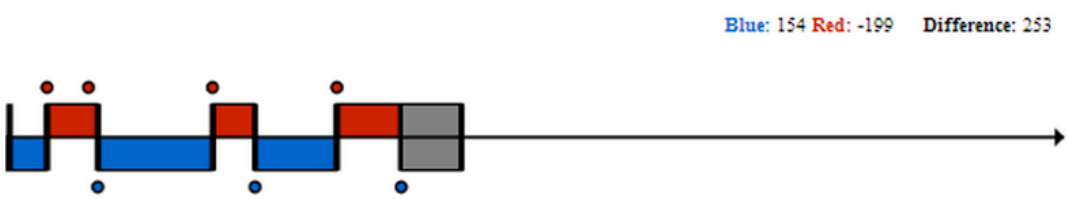
  Your only move is to play 'flip'. If you are in control and you play 'flip' you remain in control. If you are not in control and you play 'flip' you regain control. Only one player can be in control at a time.

- **The Board**

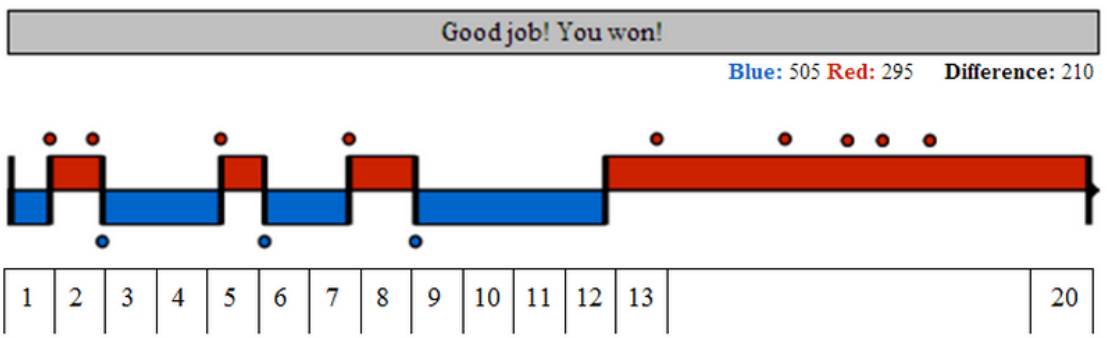
  The board displays the current known information about the game, including your points, the points of the red player, and the difference between your points and the points of the red player. Each 'flip' played is marked with a dot. You can only see information that was revealed by your flips. Blue rectangles represent periods of time in which you, the blue player, had control. Red rectangles represent periods of time in which the red player was in control.

An example game:

The game in progress:



The game when finished:



Time (in seconds)

Let's examine the moves made in the game given above.

- *1st second*: The blue player starts in control.
- *2nd second:* The red player plays 'flip' and gains control. The red player plays 'flip' again less than a second later and remains in control.
- *3rd second:* The blue player plays 'flip' and regains control. He maintains control for a bit over 2 seconds.
- *5th second:* The red player plays 'flip' and regains control. He keeps control for less than a second.
- *6th second:* The blue player plays 'flip' and regains control. He keeps control for about 2 seconds.
- *7th second:* The red player plays 'flip' and regains control. He keeps control for about 1 second.
- *9th second:* The blue player plays 'flip' and regains control. He maintains control for 4 seconds.
- *12th second:* The red player plays 'flip' and regains control. He maintains control for the rest of the game. He makes a number of flips in which he maintains control.
- *20th second:* The game ends.

The blue player was in control for **8.05** seconds earning **805** points, and played 'flip' 3 times, costing **300** points. This gives him a total score of **505** points.

The red player was in control for **11.95** seconds earning **1195** points, and played 'flip' 9 times, costing **900** points. This gives him a total score of **295** points.

The blue player has more points than the red player and thus wins.

- ## Payment

You will be compensated according to your performance in this study. For completing the study, you are guaranteed the amount listed on the Mechanical Turk HIT that you have accepted, and you will be paid an additional sum based on your performance.
You will participate in multiple rounds of the game. At first, you will participate in a practice round without a bonus payment to familiarize yourself with the interface. Then, you will participate in several additional rounds. You can receive a bonus payment for your performance in each of those rounds.
You can increase your bonus payment in a given round by performing well compared to the red player. If you lose by more than 1000 points, however, you will receive no bonus payment for that round.
The exchange rate for points into the bonus payment is 1 cent for 100 points. For example, you would earn a bonus payment of 10 cents by gaining exactly as many points as the red player. If you outperform your opponent by 500 points you would earn 15 cents. If you underperform you opponent by 500 points you would earn 5 cents.

I understand the above rules