

NEUROMETRICALLY INFORMED MECHANISM DESIGN  
AND THE ROLE OF VISUAL FIXATIONS IN SIMPLE CHOICE

Thesis by

Ian Krajbich

In Partial Fulfillment of the Requirements

for the Degree of

Doctor of Philosophy

California Institute of Technology

Pasadena, California

2011

(Defended July 9, 2010)

(c) 2011

Ian Krajbich

All Rights Reserved

## ACKNOWLEDGEMENTS

If not for Colin Camerer, I would most likely be about halfway through a tedious Ph.D. in physics (my undergraduate major) or have lost interest by now and sold my soul to Wall Street or consulting, like many of my other Caltech undergraduate friends. But Colin introduced me to, in order, economics, behavioral economics, and finally neuroeconomics. He kindled a passion in me for a field of research that I, up to that point, knew and cared very little about. In fact, in high school I was so sure that I wanted nothing to do with economics that I arranged to take the course on “credit by exam” so that I would not have to sit through the semester-long class. I guess that I should have taken it as a sign, when I passed the exam having only skimmed the textbook’s chapter summaries that morning. Over the years, Colin has been there to advise, inspire, entertain, and provide all the support a student could ask for. For that, I will always be thankful.

Colin was the one who sold me on economics, but it was really Ralph Adolphs who piqued my interest in neuroscience. Simply put, Ralph has been an amazing figure in my graduate school career, and I can’t imagine what the last few years would have been like without him. I can’t thank him enough for including me in his lab and inspiring me to be not just a productive scientist, but a happy one.

On a similar note, I want to thank John Ledyard, both for his mentorship in my mechanism design work, and his dedication to keeping morale high for the graduate students in HSS. It was always a great comfort to me (and many others) to know that John would be at the Athenaeum every Friday afternoon, ready to share a pitcher and discuss whatever was on our minds, be it prelims, research ideas, or just the latest news in baseball.

I would also like to thank Peter Bossaerts, John O'Doherty, Preston McAfee, Ed McCaffery, Kim Border, Jean-Laurent Rosenthal, Charlie Plott, and Bob Sherman for their help and/or advice at various points along the way. Also, many thanks to Tiffany Kim, Laurel Auchampaugh, Melissa Slein, Sheryl Cobb, Barbara Estrada, and Susan Davis for help with day to day logistics.

I owe the greatest debt for this thesis to my advisor and mentor Antonio Rangel. Antonio's devotion to his craft is unparalleled in anyone that I have ever met. His passion for research and pure willpower to get things done has been a phenomenal motivating force throughout the time that I've known him. Working with Antonio is like pushing the fast-forward button on your life, and it has helped me (and others in the lab) be productive scientists in a rapidly changing field. Without his tireless work writing papers and grants to fund scanning hours, our awesomely powerful cluster, and conference trips, none of this would have been possible. I have to say that it is truly an honor to be Antonio's first Ph.D. student.

I owe thanks as well to all the members of the Rangel, Camerer, Ledyard, and Adolphs labs. Thank you all for making Caltech such a fun and stimulating place to be a student. I owe particular thanks to Todd Hare, Hilke Plassmann and Ming Hsu, who really helped me find my legs in fMRI.

Thanks to all my friends and family for their support throughout grad school. Thanks to my wonderful girlfriend, Hannah, for always being there and for putting up with me even on the bad days. In addition to my friends in the various labs, I also want to give special thanks to my good friends Chris and Dustin. And of course, thanks to my parents, who were 100% supportive of my choice to pursue neuroeconomics and who were always there to help out in rough times.

Finally I want to thank the National Science Foundation IGERT program for funding me through my Ph.D.

## ABSTRACT

The young field of neuroeconomics has already produced many important insights into the neurobiological underpinnings of decision making. However, at this early stage it is still unclear how much influence the field will have on mainstream economics. Here, I show how a neuroeconomics approach can shed light on two classic economic problems.

First, I show that it is possible to predict individuals' values for public goods, using functional magnetic resonance imaging (fMRI)-based pattern classification. With such predictions in hand, I demonstrate that it is possible to solve the free-rider problem, by taxing individuals based both on the values that they themselves report and on the predicted values (using fMRI). I go on to more generally prove that by using any informative signal of value, it is possible to overcome classic impossibility results in mechanism design. This allows us to construct mechanisms that simultaneously satisfy dominant strategy incentive compatibility, voluntary participation, budget-balance and social efficiency. Such mechanisms were previously thought to be impossible. I demonstrate how to construct such mechanisms, and test them in three different public goods experiments.

Second, I show that individuals' looking patterns are critical to the decision making process. When people make choices between options, they tend to look back and forth

between them. One might think that these “fixations” are an unimportant by-product of the choice process, but I demonstrate that they are in fact intimately tied to the comparison process. By using a variant of the drift-diffusion models from the perceptual decision making literature, I find that fixations seem to bias the accumulation of evidence towards the item that is being looked at. Therefore, if one spends more time looking at one item over the other, then one is more likely to choose that item. Critically, I am able to show that this effect is not due to subjects looking longer at preferred items. The model has deep implications for how looking patterns (treated as exogenous) should bias choices, and I confirm these predictions using eye-tracking data from subjects choosing between snack foods.

## TABLE OF CONTENTS

Acknowledgements.....	iii
Abstract.....	vi
Table of Contents.....	viii
Summary.....	10
Ch. 1: Using Neural Signals of Value to Solve the Public Goods Free-Rider Problem... 12	
Methods.....	24
Main Experiment: Behavioral Methods.....	24
Main Experiment: fMRI and Classification.....	29
Calibration Experiment: Methods.....	34
Calibration Experiment: fMRI and Decoding.....	36
Rules of the NIM.....	36
Key Properties of the NIM.....	40
Additional Properties of the NIM.....	53
Optimal Mechanism in the Absence of Informative Type Signals.....	54
References.....	58
Appendix.....	60
Instructions.....	80



Ch. 2: Neurometrically Informed Mechanism Design.....	93
Introduction.....	93
Theory.....	97
Review of Classical Mechanism Design Theory.....	97
Neurometrically Informed Mechanisms.....	102
Some Additional Observations.....	109
Experiments: Neurometrically Informed Public Goods.....	111
Experiment 1.....	112
Experiment 2.....	120
Final Remarks.....	130
References.....	134
Appendix.....	139
Ch. 3: Visual Fixations Guide the Computation and Comparison of Value in Simple Choice.....	161
Results.....	163
Discussion.....	181
Methods.....	186
References.....	196
Appendix.....	199

## SUMMARY

How can neuroscience inform economics? For many economists, this is the fundamental question for the new field of neuroeconomics. From a neuroscience perspective, there is no question of the value of understanding how the brain makes decisions. But from an economics perspective, it is unclear how knowledge of the brain can improve a field that is only concerned with revealed preference. In this thesis I tackle this important issue by presenting two examples of how neuroeconomics can help address traditional economics problems.

In the first two chapters I show how using neurometric signals of value can overcome famous impossibility results in mechanism design. In Chapter 1, I discuss the public goods problem. Every group needs to decide when to provide public goods and how to allocate the costs. In an ideal arrangement, individuals would reveal their values for the public good to the government, the socially optimal level of the good would be implemented, and the costs would be fully paid using fees that are proportional to individual benefits. Unfortunately, the economic theory of mechanism design has shown that this ideal solution is not possible when the government lacks knowledge about the individual valuations. I show that this impossibility result can be overcome in experimental settings by combining technologies for obtaining neural measures of value (functional magnetic resonance imaging-based pattern classification) with carefully designed economic incentives.

In Chapter 2, I extend the results from Chapter 1 to mechanism design in general. Several classic results have shown that it is impossible to design mechanisms that simultaneously satisfy efficiency, voluntary participation, and dominant strategy incentive compatibility. The results in the first chapter showed that it is possible to obtain noisy signals of subjects' preferences. Here I show that the availability of even mildly informative signaling technologies has a profound impact on the mechanism design problem. In quasi-linear environments, it is possible to construct "neurometrically informed mechanisms" that implement any desired allocation rule by using both subjects' reported preferences and their signals. In particular, there are neurometrically informed mechanisms that simultaneously satisfy efficiency, voluntary participation, and dominant strategy incentive compatibility. I go on to show, in two experiments, how to apply neurometrically informed mechanisms to complicated public goods games and how well these mechanisms perform in the presence of risk- and loss-aversion.

In the final chapter, I show how information about individuals' looking patterns can help shed light on the decision making process and better predict their choices. Most organisms facing a choice between multiple options will look repeatedly at them, presumably implementing a comparison process between the items' values. Little is known about the exact nature of the comparison process in value-based decision making, or about the role that the visual fixations play in this process. I propose a computational model in which fixations guide the comparison process in simple binary value-based choice and test it using eye-tracking. I show that the model is able to quantitatively predict complex relationships between fixation patterns and choices.

## CHAPTER 1

### **Using Neural Signals of Value to Solve the Public Goods Free-Rider Problem**

Public good allocation problems are pervasive in society. Examples in the government sector include the provision of national defense and environmental cleanups. Examples in the private sector include hiring a security guard or improving common areas in a condominium association. These examples highlight two key features of public goods. First, since their benefits are non-excludable, they are enjoyed by all members of the group, even those who do not help pay for them. Second, the optimal allocation of public goods depends on the group members' willingness-to-pay for them (1).

If the government (or group leadership) knew every individual's valuation for the good, the allocation problem would be straightforward: The government could compute the socially optimal level of the public good and then tax group members in proportion to the benefits that they receive in order to finance the cost of the good. In fact, in this case there are many possible fair rules for splitting the cost of the public good such that every individual's benefit from the public good is greater than his tax (2, 3). Unfortunately, individual valuations for public goods are not directly observable by the government, which makes the allocation problem challenging. In particular, self-interested individuals have an incentive to understate those values, if they are asked directly for their valuations and know that their share of the cost will increase with their reported values. This is known as the free-rider problem, and it makes it very difficult in practice to accurately

determine which public goods should be provided and how the costs should be shared. Countless experiments around the world have shown that the financial incentive to free-ride is pervasive and leads to allocations with a socially inefficient level of public good provision (4-6).

Social scientists have explored two different ways to limit the problems caused by free-riding. One approach investigates whether pro-social motives can be used to overcome the financial incentive to free ride. For example, pre-play communication and costly punishment of free-riders have been shown to ameliorate the problem in laboratory settings (7, 8). Although the full capabilities of these types of institutions are not yet known, the body of evidence (4, 5) suggests that pro-social motives are not always sufficient to eliminate free-riding behavior in all cultures (9). It is also unknown if these motives are strong and pervasive enough to solve large-scale problems of practical interest.

The second approach has focused on designing institutions (known as “mechanisms”) that make it advantageous for self-interested individuals to reveal their true values. A mechanism is a set of rules specifying the information that is collected from the group members and how that information is used to decide how much of the public good to produce and how to split the costs. The number of potential mechanisms for public good problems is very large. Fortunately, the mechanism design problem is greatly simplified by a result, known as the revelation principle (10-12), which states that for every mechanism with a desirable set of properties, there is a related mechanism that achieves

the same outcomes but in which individuals are simply asked to reveal their values. This result is useful and important because it limits the search space to direct revelation mechanisms: if a desirable solution does not exist within this class, then it does not exist at all.

A large body of work in economics has sought to design revelation mechanisms satisfying four desirable properties. The first is social efficiency (SE), which requires that the optimal amount of the public good always be produced, meaning that the net benefit to the group is maximized. The second property is dominant strategy incentive compatibility (DSIC), which requires that the wealth-maximizing strategy for each member of the group is to reveal his true value, regardless of others' values or behavior. This property is desirable because truthful reporting is essential for determining the socially efficient level of the public good, and DSIC ensures that every subject has a financial incentive to do so regardless of his beliefs about the other group members. The third property is budget-balance (BB), which requires that the cost of the public good be completely covered by the members of the group. This property is desirable because it rules out the need for outside sources of funding. The fourth property is voluntary participation (VP), which requires that the expected value from participating in the mechanism be non-negative for each individual, so that members do not have to be coerced into participating. A central result in economic theory is that there is no set of rules satisfying all four desired criteria (SE, DSIC, BB, and VP) simultaneously (13). In response to this fundamental impossibility result, theorists and experimenters have explored mechanisms that relax some of the criteria, but those mechanisms constitute a

less-than-ideal solution to the problem (14, 15).

A key assumption behind the impossibility result is that the information used by the mechanisms is restricted to voluntarily reported values. However, a growing body of work in neuroscience has shown that it is possible to read subjective states with ranging degrees of accuracy (commonly 60-90%) using technology such as functional magnetic resonance imaging (fMRI) (16-23). This technology opens the door for a new class of mechanisms in which outcomes and payments depend both on individuals' reported values and on neural readings about their values. We refer to this new class of institutions as Neurometrically Informed Mechanisms (NIMs).

To explore the technological feasibility of NIMs, we studied the public good allocation problem in a simple experimental setting. In each trial subjects were randomly assigned to a group of size  $N=5, 10, 15, 20,$  or  $25$  and were assigned either a *Low* (\$0-2) or *High* (\$8-10) induced value for an abstract public good. The cost of this good was fixed at  $\$5 \times N$ . As is common in experimental economics, subjects were paid based on their payoffs in the experiment. Therefore, subjects were paid an amount equal to their value for the public good if it was produced, and zero otherwise. Subjects made decisions in 50 different trials and were paid based on their average payoff from all trials. The overall payoff for each trial depended on the subject's value, the tax he had to pay (described below), and whether or not the public good was produced. Under the NIM the public good was produced only when the sum of the reported values was greater than its cost. The true values were independently and identically drawn from a uniform distribution so

that on average it was efficient to produce the public good in only half of the trials.

The experimental task procedure and rules of the NIM were as follows. First, subjects were shown the parameters of the decision problem in the sequential order depicted in Figure 1A while undergoing whole-brain fMRI. Their trial-specific value for the public good was shown in isolation during an initial screen, which allowed us to use a non-linear support-vector-machine classifier (SVM) to predict subjects' values (*High* or *Low*) based only on their pattern of neural responses to the value screen. After seeing the group size and the total cost of the public good, subjects chose whether to report their true value for the public good (*High* or *Low*). If the public good was produced, the NIM then used both the classifier predictions and the reported values to determine the taxes paid by each individual, as depicted in Figure 1B. Note that subjects are penalized with a higher tax when their reported value differs from the classifier's prediction. Furthermore, the higher the prediction accuracy, the more likely it is that a lie will be detected.

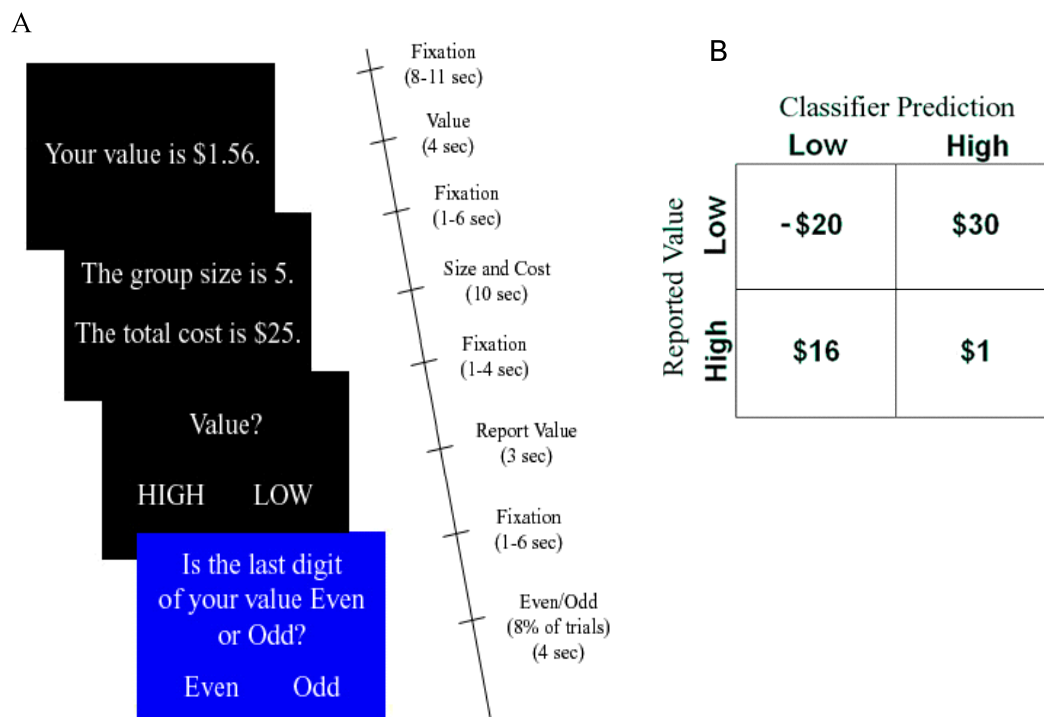
In the Methods section we show that the NIM satisfies SE, DSIC, BB, and VP. Since the public good is produced only when the reported values exceed the cost, SE requires that every individual reveal his true value. Subjects' incentives to reveal their true values depend on what they believe the accuracy of the classifier to be. Figure 2A depicts the difference in expected payoff between truth-telling and lying as a function of the classifier's accuracy. Note that *Low*-value types are strictly better off revealing their true value for any classification rate between 50% (i.e., no decoding) and 100% (i.e., full decoding). In contrast, *High*-value types are strictly better off revealing their true value



for any accuracy rate above 55%, but have an incentive to lie for rates between 50% and 55%. This provides an intuition for why the mechanism satisfies DSIC, and thus SE, for classification rates above 55%. Figure 2B shows the total expected payoff from reporting truthfully in the NIM at different classifier accuracies, assuming that the other subjects are reporting truthfully and that they have *High* values 50% of the time. The expected payoff is positive for both value-types at classification rates above 60%, which illustrates why VP is satisfied. Finally, BB is satisfied because by design the NIM distributes any financial surplus or deficit evenly between the players.

There was no feedback during the experiment and subjects' values were classified afterwards to determine their payments. Therefore, subjects made decisions based solely on their beliefs about the classifier's accuracy, which were assessed at the end of the experiment by debriefing. The rules of the NIM were explained to the subjects beforehand. In particular, they were told that in a previous experiment the same classification algorithm used here was able to predict values with an accuracy of 60%. Clear instructions about how the mechanism works are necessary to guard against comprehension mistakes that would cloud interpretation of the results, and are considered a requirement by mechanism designers (24). The  $60 \pm 2\%$  (SEM) estimate for the classifier accuracy was based on an actual preliminary calibration experiment in which 14 subjects played a simple version of the NIM. In this experiment the classified values played no role on outcomes and the subjects did not know that their values were being predicted (Fig. 3A).

**Figure 1.** **A)** Timing of the experimental trials (top to bottom). **B)** Tax paid by the subject in each trial as a function of the classifier's prediction and his reported type. Negative numbers denote transfers to the subjects.



**Figure 2.** **A)** Expected benefit of truth-telling as a function of value type and classifier accuracy. For a particular classifier accuracy, the value of the curve indicates the difference in expected payoff between reporting truthfully and lying. The arrow denotes the payoffs at the 60% accuracy rate used to describe the mechanism. Note that a subject's decision is based on their beliefs about what the accuracy of the classifier will be, and not on the realized accuracy after the experiment. **B)** Total expected payoffs as a function of the actual classification accuracy of the mechanism for a subject who reveals his true type. For a particular classifier accuracy, the value of the curve indicates how much the subject can expect to earn on average, if he reports his type truthfully. A positive value means that VP is satisfied; a negative value means VP is violated. Note that since the function is increasing with the accuracy rate, subjects have an incentive to cooperate with the experimenter to make the classifier as accurate as possible.

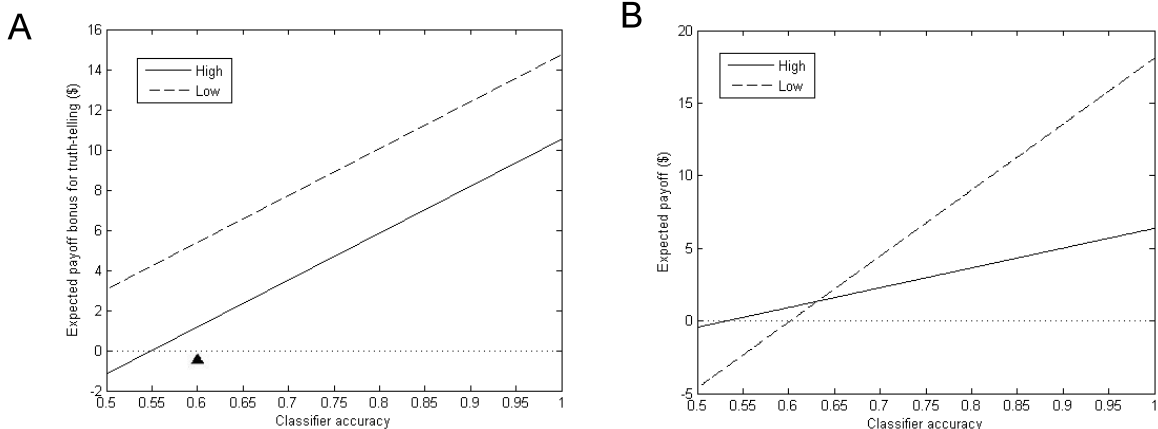
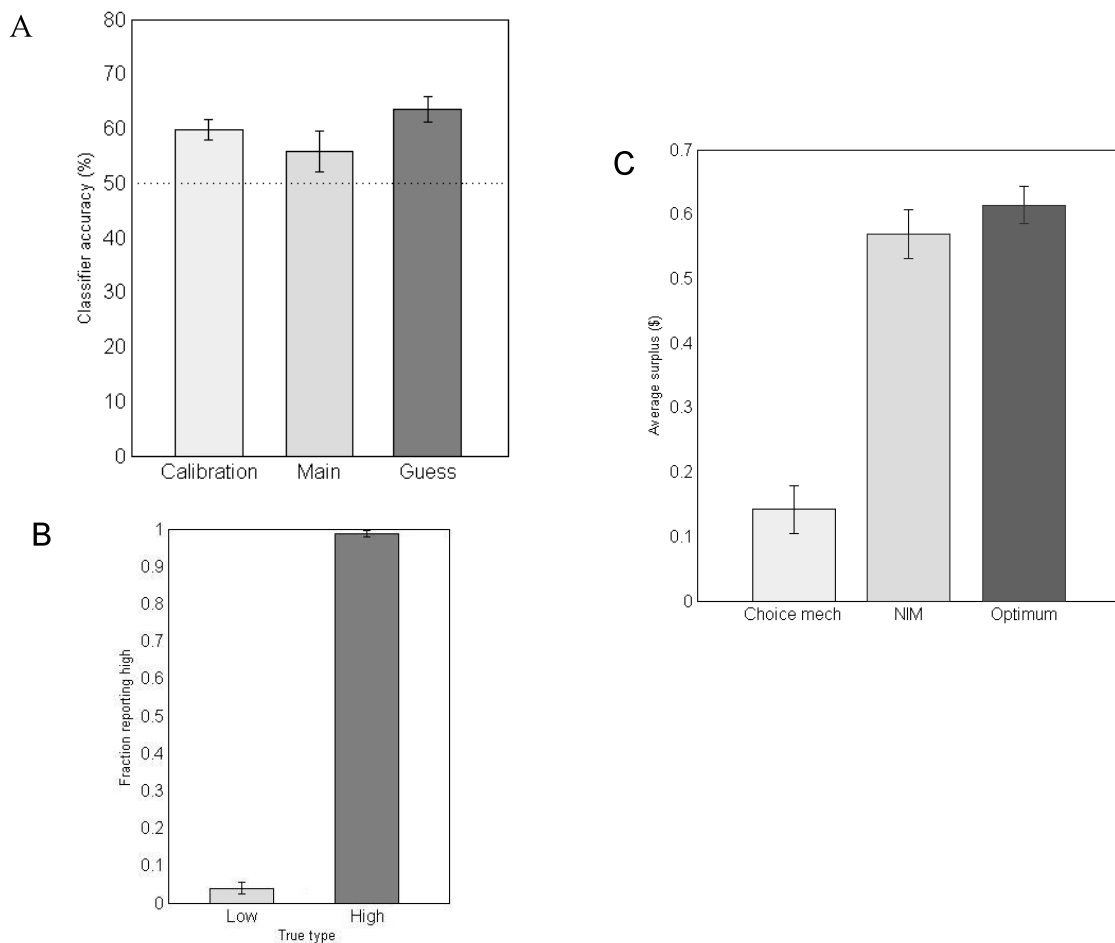


Figure 3 depicts the results of the experiment. The average classification accuracy was  $56 \pm 4\%$  (SEM), insignificantly below the stated 60% rate (two-tailed  $p=0.33$ ; Fig. 3A). We tested subjects' belief in the accuracy of the classifying technology by asking them to predict the classifier prediction rate for their own data and rewarding them based on the accuracy of their guess. During the debriefing period subjects predicted a classification rate of  $64 \pm 2\%$  (SEM), which is insignificantly different from the actual classification rate (two-tailed  $p=0.10$ ; Fig. 3A). Most importantly, subjects revealed their true values nearly 100% of the time, consistent with the properties of the NIM at the subjects' predicted classification rates (Fig. 3B). Figure S12 shows that the frequency of truth-telling did not change during the experiment. Figure 3C compares the social surplus generated by the NIM, which is a measure of social efficiency, with two important benchmarks: (1) the social optimum that could be achieved if the government had full information and thus could always choose the socially efficient allocation, and (2) the theoretical average outcome generated by the best non-NIM mechanism satisfying BB, VP, and DSIC. The NIM generated 93% of the full-information social optimum, as compared to 23% for the best theoretical non-NIM mechanism.

**Figure 3. A)** Mean accuracy rate of the classifier in the calibration ( $N=14$ ) and main experiments ( $N=10$ ), as well as the classification rate guessed by the subjects ( $N=10$ ), with standard error bars. Two-sided p-values: calibration vs. main = 0.41, calibration vs. guess = 0.20, main vs. guess = 0.10. **B)** Individual reports as a function of true type ( $N=10$ ), with standard error bars clustered by subject. *Low* types misreported *High*  $3.5 \pm 1.6\%$  (SEM) of the time, while *High* types reported *High*  $99.5 \pm 0.5\%$  (SEM) of the time. **C)** Average social surplus per individual ( $N=489$ ), a monetary measure of social efficiency, in the best non-neural mechanism ( $0.142 \pm 0.037$  (SEM)), the NIM ( $0.569 \pm 0.038$  (SEM)), and the best possible allocation under conditions of full information ( $0.614 \pm 0.029$  (SEM)). One-sided p-values are the following: Choice vs. NIM =  $10^{-18}$ , Choice vs. Optimum =  $10^{-23}$ , NIM vs. Optimum = 0.20.



This study establishes the viability of NIMs in a simple experimental setting with two types and with experimentally induced valuations. Since NIMs constitute a significant departure from previous institutions used to solve the public goods allocation problem, it is worth highlighting several of their key properties.

First, NIMs advance the theory and practice of mechanism design by combining economic theory with neural measurement technology. In the past, economists have considered mechanisms that only use the reported values from each group member to determine if the public good is produced and how the costs are shared. Here we show that it is possible to do substantially better by also employing fMRI measures that are reliably correlated with value.

Furthermore, the use of NIMs is not limited to fMRI technology. As shown in detail in the Methods section, all that is needed for the NIM to work is the existence of some signal of value that is known to be informative, whatever its source. Thus, simple physiological measures (e.g., pupil dilation or facial electromyography) might be feasible as well.

Another attractive property of NIMs is that they do not depend on beliefs about the types or behavior of the other group members. Truth-telling and voluntary participation are both dominant strategies with these mechanisms. The only requirement is that subjects believe that their values can be predicted with sufficient accuracy by the technology. Therefore, NIMs might not be viable if subjects could interfere with the technology.

Fortunately, NIMs have a built-in incentive for subjects to make the classifier predictions as accurate as possible, since subjects' expected payoffs are increasing with the prediction accuracy (Fig. 2B).

Finally, VP is an attractive feature of the NIMs because it ensures that the public good makes every individual better off, so the entire group has an incentive to support the use of the NIM. Mechanisms are deliberately required to satisfy this VP property to bolster widespread acceptance. Note however, that VP can be harder to satisfy when individuals have substantial amounts of risk- or loss-aversion (see Appendix), although the problem is substantially reduced as the accuracy of the neural measurements improves. Thus, future technological advances should alleviate this problem.

To summarize, the free-rider problem has been a challenge for economics, public policy, and political science since the work of Adam Smith (25). The field of mechanism design made substantial progress during the 20th century. Unfortunately, a major contribution of the theory was to show that an ideal solution is not possible when institutions rely only on revealed values. We have shown that this problem can be overcome in simple public good settings by using fMRI to obtain informative signals of individuals' values, and using those signals to induce truthful reporting. Our results take the first step in combining physiological measurements with carefully designed mechanisms to create better institutions for collective decision making. Future theory and experiments will be needed to take this technology to more practical applications.

## Methods

### I. Main experiment: Behavioral methods

*Subjects.* 10 subjects (5 male, mean age 25) were recruited from the Caltech community. Caltech's Human Subjects Internal Review Board approved the experiment.

*Economic setting.* In every trial subjects make decisions in a group of size  $N$ . The group needs to decide whether or not to produce a public good that generates benefits to the group members, but costs  $\$C$  to produce. The value of the public good is different for each subject, and the value for subject  $i$  is denoted by  $V_i$ .  $V_i$  is drawn randomly and independently each trial from a mixture of two uniform distributions, one with support  $\$0$ - $\$2$ , and one with support  $\$8$ - $\$10$ . The value is drawn by first selecting one of the two distributions with equal probability and then drawing it from the selected distribution. If  $V_i$  is drawn from  $\$0$ - $2$ , then we say that the subject has a *Low* value, and if  $V_i$  is drawn from  $\$8$ - $10$ , then we say that the subject has a *High* value. We use 5 different values for  $N = \{5, 10, 15, 20, 25\}$  with corresponding costs  $C = \{\$25, \$50, \$75, \$100, \$125\}$ . Note that  $C/N = \$5$  in every case. As a result, on average it is efficient to produce the good only half of the time. Each subject played 10 trials for each value of  $N$ , and trials were presented in a random order.

*Task.* Subjects make decisions in 50 different trials. In each one, they are asked to report their value type to the experimenter: *High* or *Low*. As described in Figure 1A, they are



given 3 seconds to indicate their choice. If they do not respond in the allotted time, then the subject's payoff for that trial is a loss of \$2 regardless of the actions of the other subjects with whom he is matched. It is important to emphasize that the subjects need not report their true values: when their value is *High* they can report *Low*, and vice-versa.

As shown in Figure 1A, on 8% of the trials subjects are asked to report whether the last digit of their value is even or odd. The appearance of this question is fixed so that it occurs in the same trials for every subject (2 in the first session and 2 in the second session). The goal of this "test" screen is to ensure that subjects are paying attention to their values in each round. The mean (standard error) percentage of correct answers was 92.5% (5.3%).

Subjects' payoffs for every trial are determined by the rules of the NIM (described in the main text and in Section V below), which maps the reports and classifier predictions from each subject in the group to a decision regarding the public good (i.e., whether or not to produce it) and to a list of tax payments for each subject.

A subject's total payment from participating in the task equaled a fixed show-up fee, plus \$5 for each correct even/odd screen (up to \$20), plus five times their average payoff on the 50 trials, plus a payoff for how accurate their guess was about the classifier accuracy (up to \$10). The factor of five was used to place more emphasis on the payoffs generated

by their decisions relative to the show-up fees. The average total payoff was \$73.

Even though subjects' aggregate decisions were used to compute payoffs, only one subject participated at a time because they were being scanned during the task. After all the subjects were scanned and the classification was performed, groups were formed ex-post for every subject-trial pair by matching them with decisions made by a random set of other subjects in trials of the same type. Thus, for example, to compute the outcome for a trial with  $N=10$  for a given subject, we randomly selected nine more trials from the other subjects for the case  $N=10$ , and their joint decisions determined the outcome for the group following the rules of the NIM. Note that since subjects have to choose their report without knowing the decisions made by others, and since no feedback is provided in the experiment, it makes no difference from a strategic point of view if a series of subjects make decisions sequentially (as they did in our experiment) or if all subjects choose simultaneously (as in most behavioral experiments of this type). Their incentives and information are identical in both cases.

No deception was used in the experiment. All the information about the decision structure and the method for determining payoffs was communicated to the subjects during the instruction period prior to the experiment. Because of the complexity of the instructions, we held a separate instruction session 1-2 days before the experiment to give subjects a chance to think in detail about the rules of the NIM and to reflect about their optimal

strategy. During this session the instructions were read out loud and then the subjects were asked to complete a quiz about them.

It is important to note that we explained that the NIM was set up so that their unique payoff-maximizing strategy was to report the true value every trial and gave them detailed (and correct) calculations showing their expected payoffs for lying and for truth-telling. They also went through calculations describing their expected payoffs in cases where the classification algorithm could predict their values at rates that were lower or higher than the 60% rate described in the instructions.

However, subjects were told that they were free to misreport their values and to try to control their brain activity. The only requirement imposed on them was to keep their eyes open and to look at the information on each screen.

The reason for explicitly calculating expected payoffs from different reporting strategies is to enhance internal validity. Without such instructions, we would not know whether any misreporting is due to a desire to misreport (perhaps doubting the accuracy of the classifier) or a miscomprehension of the financial consequences of misreporting. This kind of instruction is often used in experimental economics when the economic consequences of particular choices are not easy to calculate, and when differences in the ability of participants to make such calculations are not a variable of interest. For

example, in Becker-DeGroot-Marschak mechanisms used to reveal maximum willingness-to-pay values, it is common to explain to subjects how truthfully revealing values leads to the highest expected payoff (26-30).

On the day of the experiment subjects answered a short quiz to ensure that they understood the experiment. This experimental session was the second time they had seen the instructions and quiz, but both were repeated to ensure that the subjects remembered and comprehended the details of the experiment. Subjects were then run through two practice trials of the experiment on a laptop outside of the fMRI scanner to familiarize them with the stimuli.

Immediately following the fMRI experiment, subjects were asked to answer the following question: “The algorithm will be able to guess my value \_\_\_\_ % of the time.” Their responses gave us a measure of how well they believed the classifier could predict their values. Note that this information was extracted in an incentive-compatible way: subjects received a maximum of \$10 if they guessed the prediction rate correctly, but their payoff decreased by \$1 for each 2%-step deviation from the true percentage (in either direction). For example, if a subject guessed that we could predict their value 60% of the time, but we could in fact predict it at a 70% rate, then they received  $\$10 - \$1 * |(70 - 60) / 2| = \$5$ .

## II. Main experiment: fMRI and classification

*fMRI scanning procedure.* Scans were acquired using the 3 Tesla Siemens Trio scanner at Caltech's Broad Imaging Center. Anatomical images (high resolution, 1x1x1mm, T1-weighted) were acquired first. Functional (T2-weighted) images were then acquired using the following parameters: TR= 2750ms, TE = 30ms, in-plane resolution, and slice thickness = 3mm, 44 slices. Horizontal slices were acquired approximately 15 degrees clockwise of the anterior-posterior commissure (AC-PC) axis to allow for complete brain coverage, and were collected in an interleaved ascending manner. The onsets of the value screens were time-locked to the beginning of TRs. The experiment lasted approximately 28 min, broken into two sessions of 14 min each.

*Data preprocessing.* The fMRI analysis proceeded in several steps. Images were corrected for slice acquisition time within each volume, motion corrected with realignment to the last volume, and spatially normalized to the standard Montreal Neurological Institute EPI template. Intensity normalization and high-pass temporal filtering (using a filter width of 128 s) were also applied to the data. We did not apply any spatial smoothing, since spatial smoothing reduces the information that can be extracted from patterns of activation.

We then ran 50 general linear models (GLMs) for every subject, one for each subset of 49 trials out of the total 50. In each of these GLMs we constructed regressors for the

following events: *High* value screen, *Low* value screen, left-out-trial value screen, group information screen, response screen, and even/odd screens. All of these events were modeled as stick functions, except for the group information screen, which was modeled as a boxcar with a duration equal to that of the event. The regressors were convolved with canonical hemodynamic response functions. For each of these 50 GLMs and for every subject we then constructed a contrast of *High* – *Low* value trials at the onset of the value screen. Based on these contrasts, we selected voxels-of-interest (VOIs) by taking the 100 voxels with the largest t-statistics in the whole brain. Univariate feature selection is a common solution to the problem of reducing the number of variables used as inputs to multivariate classifiers such as support vector machines (SVMs) (31). Note that our procedure avoids a peeking problem that has been pervasive in previous fMRI classification exercises: we pick VOIs based on the responsivity to *High* vs. *Low* values without using information from the 50<sup>th</sup> test trial in which we want to *predict* the subject's value. For more detail on this leave-one-out procedure, see the Classification Analysis section below.

After whitening the data (AR(1) process) and removing movement-associated noise, we then extracted the BOLD time-courses for each VOI and created a more continuous signal by interpolating the data at 12Hz. We then computed the average BOLD signal for the 3 to 7 second time interval after each value screen onset, for each of the VOIs.

*Classification analysis.* The process described above generated a vector of 100 measures

of BOLD activity for each trial, which served as the input to the classification analysis. This data set was analyzed using non-linear SVMs with radial basis functions. We used the Matlab SVM toolbox from LIBSVM to run these analyses.

The first step of the SVM analysis was to normalize each voxel's activity across trials to lie in the interval  $[0,1]$ , as suggested by the LIBSVM documentation (32).

The data was then run through a cross-validation procedure with two nested cycles (33). In this procedure, we iteratively attempted to predict the value for each trial by training our learning algorithm on the other 49 trials and then applying the resulting model to the test trial, generating a prediction of *High* or *Low*. This procedure is known as “leave-one-out” classification. Note that for each leave-one-out iteration we used a different set of 100 voxels to predict, as described above.

The radial basis function used by SVM contains two free parameters, *gamma* and *C*. *Gamma* is essentially the degree of non-linearity in the SVM. *C* is a parameter in the cost function that determines the model's tolerance for counterexamples (34,35). In order to select the parameters with the best predictive power, we performed a Five-fold cross-validation on the 49 training trials, searching over all combinations of *gamma* and  $C = \{2^5, 2^{-4}, 2^{-3}, 2^{-2}, 2^{-1}, 2^0, 2^1, 2^2, 2^3, 2^4, 2^5\}$ . Five-fold cross-validation is similar to leave-one-out except that it randomly splits the 49 training trials into 5 groups (9-10 trials each), trains the SVM on 4 of those 5 groups, and then predicts on the remaining “validation”

group. This is repeated 5 times, each time using a different validation set from the 5 possible groups. After the five-fold cross-validation, we picked the combination of parameters that yielded the highest prediction accuracy on the validation sets. We then fixed these parameters, retrained the SVM on all 49 training trials, and finally ran the resulting model on the remaining test trial to generate a prediction of *High* or *Low*.

We ran through this whole analysis for each of the 50 trials, each time leaving one trial out and training the model on the remaining 49 trials. We used these 50 predictions to determine the classifier's accuracy, i.e., the percentage of the time that we correctly predict the subject's value.

We conducted a permutation test to ensure that our classifier was not biased towards reporting *High* or *Low*. Such a bias could lead to artificially high prediction rates if the number of *High* and *Low* trials was not equal for each subject (as was the case in this experiment). We performed 100 permutations per subject, each time randomly permuting the labels (*High* and *Low*) and then running the classification procedure as outlined above. The mean (SEM) prediction rate across all 1000 runs (100 permutations x 10 subjects) was  $49.3 \pm 4.0\%$ , where the standard errors are clustered by subject. This indicates that there was no significant bias in the classifier. Table 1 depicts a confusion matrix for the main experiment, showing the number of predicted *High* and *Low* types as a function of the true type.



**Table 1.** Confusion matrix for the main experiment, showing the number of predicted *Highs* and *Lows* as a function of the true type.

	Predicted type		
		High	Low
Actual type	High	152	112
	Low	109	127

### III. Calibration experiment: Methods

*Subjects.* 15 subjects (12 male, mean age 24) were recruited from the Caltech community. Caltech's Human Subjects Internal Review Board approved the experiment. One subject was excluded due to excessive head motion in the scanner.

*Economic setting.* The economic setting was identical to the one for the main experiment.

*Task.* The setup for the calibration experiment was the same as in the main experiment, with some small changes. The timeline was the same, except that instead of asking subjects to report their value (*High* or *Low*), we asked them to cast a vote (*Yes* or *No*) for the public good.

Outcomes were then determined by a simple voting mechanism. There was a threshold  $K$  such that the public good was produced if and only if  $K$  out of  $N$  subjects voted *Yes*. If the public good was produced everyone in the group received their private values (even if they voted against it) and the subjects who voted *Yes* split the cost  $C$  evenly among them. If the public good was not produced there were no costs or benefits to any of the subjects. The information screen showed the threshold  $K$  in addition to the group size  $N$  and the cost  $C$ .

As in the main experiment, there were 5 different values for  $N = \{5, 10, 15, 20, 25\}$  with corresponding thresholds  $K = (2N/5) = \{2, 4, 6, 8, 10\}$  and costs  $C = \$5N = \{\$25, \$50, \$75, \$100, \$125\}$ .

Subjects went through 50 trials with the same structure as before. Subjects' values in each individual trial were also drawn according to the rules explained above. A slight change from the main experiment was that the occurrence of Even/Odd screens was 5% and was random (not a fixed 4 trials as in the main experiment). Also, rather than paying subjects for all trials, subjects in the calibration experiment were paid for the sum of five randomly selected trials, one for each group size. The average total earnings were \$45.

The reason for the different task structure in the calibration experiment was because we wanted to use a calibration experiment that was as close as possible to traditional public goods cost-sharing experiments. We could not use the NIM task for calibration since that would have required telling the subjects what the predicted rate of the classifier would be, which at that point we did not know. Also, we wanted to estimate the classifier's baseline prediction rate in a setting where individuals did not know that we were trying to guess their values.

#### **IV. Calibration experiment: fMRI and decoding**

*fMRI scanning procedure.* The fMRI procedure was identical to the main experiment except that the timing was slightly different. The total time duration of the experiment was approximately 24 min, broken into two sessions of 12 min each. We did not align the onsets of value screens with the beginning of TRs.

*Data preprocessing.* It was identical to the main experiment.

*Classification analysis.* Classification was done exactly as in the main experiment.

#### **V. Rules of the neurometrically informed mechanism (NIM)**

This section provides a more formal and detailed description of the rules of the NIM used in the main experiment.

Before doing so we need to introduce some notation:

- $v_i \in \{[0, 2] \cup [8, 10]\}$  denotes the value of the public good to player  $i$ .

-  $\theta_i = \{1 \text{ if } v_i \in [0, 2]; 9 \text{ if } v_i \in [8, 10]\}$  denotes the value-type of a subject with value  $V_i$ , where 1=*Low* and 9=*High*.

-  $r_i \in \{1, 9\}$  denotes the type reported by subject  $i$ , where 1=*Low* and 9=*High*.

-  $s_i(\theta_i) \in \{1, 9\}$  denotes the type assigned by the classifier to subject  $i$ . Note that if the classifier is informative, the probability of being classified a *High* type depends on the individual's true type.

-  $\delta(r_1, \dots, r_N) \in \{0, 1\}$  denotes the level of public good as a function of the reports made by everyone in the group.

-  $\delta^*(\theta_1, \dots, \theta_N) = \{1 \text{ if } \sum_{i=1}^N \theta_i > C; 0 \text{ if } \sum_{i=1}^N \theta_i \leq C\}$  denotes the optimal level of public good given a distribution of types in the group. Note that the NIM assumes that *High* types have a value for the public good equal to \$9, and that *Low* types have a value equal to \$1.

The rules of the NIM are easily described using this notation. Each subject in the group reports simultaneously. The NIM then sets  $\delta(\mathbf{r}) = \delta^*(\mathbf{r})$ . If the public good is not

produced then there are no taxes and all subjects earn \$0 in that trial. In contrast, if the public good is produced then each player receives their value but has to also pay a tax given by:

$$t_i(\mathbf{r}, \mathbf{s}) = f_i(\mathbf{r}_i, \mathbf{s}_i(\theta_i)) + \mathbf{b}(\mathbf{r}, \mathbf{s}(\theta)),$$

where  $f_i(\mathbf{r}_i, \mathbf{s}_i(\theta_i))$  is a baseline tax that depends only on player  $i$ 's report and predicted type, and  $\mathbf{b}(\mathbf{r}, \mathbf{s}(\theta))$  is a rebalancing tax that depends on the behavior and predicted types of all the group members.

The baseline taxes  $f_i(\mathbf{r}_i, \mathbf{s}_i(\theta_i))$  used in the experiment are described in Figure 1B and in Table 2A. Note that since the classifier is noisy, there is uncertainty about the actual taxes that subjects need to pay. Table 2B describes the expected baseline taxes with the assumed classification rate of 60%.

The rebalancing transfer  $\mathbf{b}(\mathbf{r}, \mathbf{s}(\theta))$  is the same for all the subjects in the group and is added to the baseline tax to ensure that the NIM is budget-balanced. Therefore, it is given by

$$\mathbf{b}(\mathbf{r}, \mathbf{s}(\theta)) = \frac{C - \sum_{j=1}^N f_j(\mathbf{r}_j, \mathbf{s}_j(\theta_j))}{N}.$$

**Table 2.** (A Top) Baseline taxes used by the NIM. (B Bottom) Expected baseline taxes based on a classifier accuracy of 60%. Positive numbers denote taxes. Negative numbers denote transfers to the subjects.

		Classifier Prediction	
		Low	High
Reported Value	Low	-\$20	\$30
	High	\$16	\$1

		True Value	
		Low	High
Reported Value	Low	\$0	\$10
	High	\$10	\$7

## VI. Key properties of the NIM

This section provides a mathematical proof that the NIM satisfies social efficiency (SE), voluntary participation (VP), budget-balance (BB), and dominant strategy incentive compatibility (DSIC).

In order to do this we need to introduce an additional piece of notation:  $\pi_i(\theta_i, r_i | \theta_{-i}, r_{-i})$  denotes the expected payoff to subject  $i$  as a function of his type and report, conditional on the types and reports of the other group members. As is common in the mechanism design literature, the arguments here assume that the subjects are risk-neutral in monetary payoff (but see the Appendix for how the results extend to the case of risk-aversion).

In order to simplify the exposition, we assume that subjects' values for the public good are given by  $\theta_i$  not  $V_i$ . In other words, we assume that *High* types have a value equal to \$9 and *Low* types have a value equal to \$1. The arguments are easily extended to the more complex case, and all of the properties hold at the subject's stated belief about the NIM classification rate of 64%.



### *Incentive Compatibility*

DSIC requires that subjects earn higher expected payoffs by truthfully reporting their type rather than lying, regardless of the types and reports of the other subjects.

Mathematically, this requires that the following incentive constraint be satisfied:

$$\pi_i(\theta_i, \theta_i | \theta_{-i}, r_{-i}) \geq \pi_i(\theta_i, r_i | \theta_{-i}, r_{-i}) \text{ for all } \theta_i, r_i, r_{-i}, \text{ and } \theta_{-i}.$$

From the description of the NIM in the previous section we know that

$$\begin{aligned} \pi_i(\theta_i, r_i | \theta_{-i}, r_{-i}) &= \delta^*(r) [\theta_i - Et_i(r | \theta)] = \delta^*(r) [\theta_i - Ef_i(r_i | \theta_i) - Eb(r | \theta)] \\ &= \delta^*(r) \left[ \theta_i - Ef_i(r_i | \theta_i) - \frac{c}{N} + \frac{\sum_{j=1}^N Ef_j(r_j | \theta_j)}{N} \right] \\ &= \delta^*(r) \left[ \theta_i - \frac{c}{N} - \frac{(N-1) * Ef_i(r_i | \theta_i)}{N} + \frac{\sum_{j \neq i} Ef_j(r_j | \theta_j)}{N} \right], \end{aligned}$$

and that

$$\pi_i(\theta_i, \theta_i | \theta_{-i}, r_{-i}) = \delta^*(r) [\theta_i - Et_i(r | \theta)] = \delta^*(r) [\theta_i - Ef_i(\theta_i | \theta_i) - Eb(r | \theta)]$$

$$\begin{aligned}
&= \delta^*(\mathbf{x}) \left[ \theta_i - \mathbf{E}f_i(\theta_i | \theta_i) - \frac{c}{N} + \frac{\sum_{j=1}^N \mathbf{E}f_j(\mathbf{x}_j | \theta_j)}{N} \right] \\
&= \delta^*(\mathbf{x}) \left[ \theta_i - \frac{c}{N} - \frac{(N-1) * \mathbf{E}f_i(\theta_i | \theta_i)}{N} + \frac{\sum_{j \neq i} \mathbf{E}f_j(\mathbf{x}_j | \theta_j)}{N} \right],
\end{aligned}$$

where  $\mathbf{E}$  indicates the expectation operator over the classifier's signals.

It then follows that the DSIC condition can be rewritten as:

$$\delta^*(\mathbf{x}) \left[ \theta_i - \frac{c}{N} - \frac{(N-1) * \mathbf{E}f_i(\theta_i | \theta_i)}{N} + \frac{\sum_{j \neq i} \mathbf{E}f_j(\mathbf{x}_j | \theta_j)}{N} \right] \geq \delta^*(\mathbf{x}) \left[ \theta_i - \frac{c}{N} - \frac{(N-1) * \mathbf{E}f_i(\mathbf{x}_i | \theta_i)}{N} + \frac{\sum_{j \neq i} \mathbf{E}f_j(\mathbf{x}_j | \theta_j)}{N} \right]$$

To verify that the DSIC condition holds for all for all  $\theta_i, r_i, r_{-i}$ , and  $\theta_{-i}$  we must consider several cases:

- 1) The public good is not produced regardless of  $i$ 's report.
- 2) The public good is produced regardless of player  $i$ 's report.
- 3) The public good is produced if  $i$  reports *High* but is not produced if he reports *Low*.

Case 1: Since the public good is not produced regardless of  $i$ 's report, the subject gets a payoff of \$0 regardless of his actions, and thus the DSIC incentive constraint is trivially satisfied.

Case 2: Since the public good is produced regardless of  $i$ 's report, his incentive constraint becomes:

$$\theta_i - \frac{c}{N} - \frac{(N-1) * E f_i (\theta_i | \theta_i)}{N} + \frac{\sum_{j \neq i} E f_j (x_j | \theta_j)}{N} \geq \theta_i - \frac{c}{N} - \frac{(N-1) * E f_i (x_i | \theta_i)}{N} + \frac{\sum_{j \neq i} E f_j (x_j | \theta_j)}{N}$$

which is equivalent to

$$E f_i (\theta_i | \theta_i) \leq E f_i (x_i | \theta_i) .$$

The fact that this inequality is satisfied for both *High* and *Low* types can easily be verified by looking at Table 2B, which describes the expected baseline taxes for a classifier accuracy of 60%. In particular, a true *Low* type pays an expected baseline tax of \$0 if he reports *Low*, but pays an expected baseline tax of \$10 if he reports *High*. Similarly, a true *High* type pays an expected baseline tax of \$7 if he reports *High*, but pays an expected baseline tax of \$10 if he reports *Low*. As a result, if the conditions of case 2 hold, both types have an incentive to reveal their true type.

Case 3a: Subject  $i$  is a *High* type and can prevent the public good from being produced by reporting that he is a *Low* type. In this case the incentive constraint becomes:

$$9 - \frac{C}{N} - \frac{(N-1) * Ef_i(\theta_i | \theta_i)}{N} + \frac{\sum_{j \neq i} Ef_j(r_j | \theta_j)}{N} \geq 0$$

The fact that this inequality is satisfied for *High* types is shown in the VP section below.

Case 3b: Subject  $i$  is a *Low* type and can cause the public good to be produced by reporting that he is a *High* type. In this case the incentive constraint becomes:

$$0 \geq 1 - \frac{C}{N} - \frac{(N-1) * Ef_i(r_i \neq \theta_i | \theta_i)}{N} + \frac{\sum_{j \neq i} Ef_j(r_j | \theta_j)}{N}$$

Using the fact that  $\frac{C}{N} = 5$  and the baseline taxes for the NIM described in Table 2A we get that this condition reduces to:

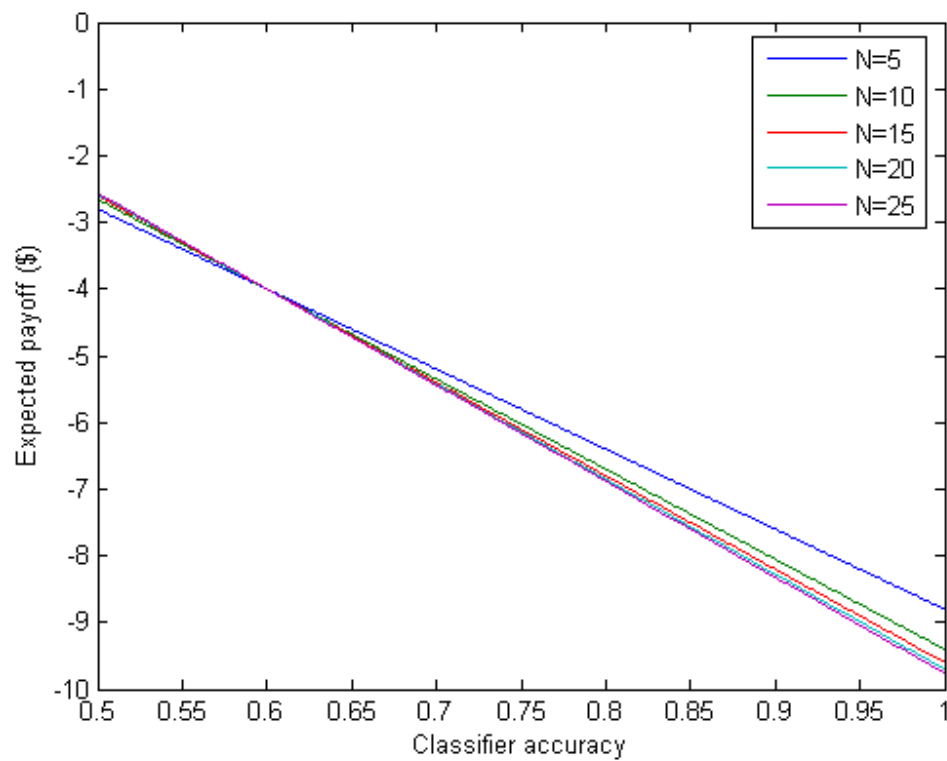
$$0 \geq -4 - \frac{(N-1) * [p * 16 + (1-p) * 1]}{N} + \frac{\sum_{j \neq i} Ef_j(r_j | \theta_j)}{N}$$

where  $p$  denotes the classification rate of the mechanism.

Note that in order to verify that this inequality is satisfied for all possible types and reports by the other subjects, it is sufficient to show that it holds for the case when the expected baseline taxes paid by the other subjects are maximized. If the inequality holds in that case, then it will hold in all cases. As shown in Table 2B, that happens when every other player is lying, in which case their expected baseline tax equals \$10.

Figure 4 plots the right hand side of the inequality assuming this “worst-case” scenario as a function of the subject’s own classifier accuracy ( $p$ ). Figure 4 shows that even in this scenario the subject’s expected payoff for lying is negative for all group sizes and classifier accuracies above  $p=0.5$ . Therefore, a decisive *Low* type will never want to lie and report *High*, since he would rather take a payoff of \$0 than a negative payoff.

**Figure 4.** Expected payoff for a decisive *Low* type who misreports being a *High* type when every other subject is also lying. The x-axis denotes the classifier's accuracy for the subject being studied. The classification accuracy for all other subjects is assumed to be 60%.



### *Voluntary Participation*

VP requires that all individuals earn positive expected payoffs from participating in the mechanism and adhering to the truth-telling strategy, regardless of the types and actions of the other subjects. Mathematically, VP leads to the following constraint:

$$\pi_i(\theta_i, \theta_{-i} | \theta_{-i}, r_{-i}) \geq 0 \text{ for all } \theta_i, r_i, r_{-i}, \text{ and } \theta_{-i}.$$

Following the derivation in the DSIC sub-section, this can be rewritten as:

$$\delta^*(r) \left[ \theta_i - \frac{c}{N} - \frac{(N-1) \pi E f_i(\theta_i | \theta_{-i})}{N} + \frac{\sum_{j \neq i} E f_j(r_j | \theta_j)}{N} \right] \geq 0 \text{ for all } \theta_i, r_i, r_{-i}, \text{ and } \theta_{-i}.$$

In order to verify that this constraint is satisfied we need to consider three different cases:

- 1) The public good is not produced.
- 2) The public good is produced and the subject is a *High* type.
- 3) The public good is produced and the subject is a *Low* type.

Case 1: Since the public good is not produced,  $i$ 's payoff is \$0 and the VP constraint is trivially satisfied.

Case 2: For *High* types the voluntary participation can be rewritten as follows:

$$9 - \frac{c}{N} - \frac{(N-1) * Ef_i(\theta_i | \theta_i)}{N} + \frac{\sum_{j \neq i} Ef_j(x_j | \theta_j)}{N} \geq 0$$

Recalling that  $\frac{c}{N} = 5$  and the baseline taxes from Table 2A, this condition reduces to:

$$4 - \frac{(N-1) * (p * 1 + (1-p) * 16)}{N} + \frac{\sum_{j \neq i} Ef_j(x_j | \theta_j)}{N} \geq 0$$

where  $p$  denotes the classifier's accuracy in predicting the subject's type.

As before, in order to verify that this inequality is satisfied for all possible types and reports by the other subjects, it suffices to show that it is satisfied for the case in which the expected baseline taxes by the other subjects are minimized. If the inequality holds in this worst-case scenario, then it will hold in all other scenarios as well. As can be seen from Table 2B, this occurs when there is only a minimum number of reported *High* types consistent with optimal production the public good, and when all players are reporting truthfully.

The minimum number of *High* types needed to produce the good is calculated using the definition of  $\phi^*(x)$  for the optimal level of the public good. Figure 5 plots the left hand side of the constraint under this worst-case scenario, as a function of the subject's own



classifier accuracy. The results demonstrate that the constraint is satisfied for all group sizes if the subject's classifier accuracy is above  $p=0.55$ .

Low types: For *Low* types the voluntary participation constraint can be rewritten as follows:

$$1 - \frac{c}{N} - \frac{(N-1) * E f_i (\theta_i | \theta_i)}{N} + \frac{\sum_{j \neq i} E f_j (x_j | \theta_j)}{N} \geq 0$$

Recalling that  $\frac{c}{N} = 5$  and the baseline taxes from Table 2A, this condition reduces to:

$$-4 - \frac{(N-1) * (p * (-20) + (1-p) * 30)}{N} + \frac{\sum_{j \neq i} E f_j (x_j | \theta_j)}{N} \geq 0$$

For the same reasons described above, to verify this constraint it suffices to show that it holds when there is only a minimum number of reported *High* types consistent with optimal production of the public good, and when all players are reporting truthfully. Figure 6 plots the left hand side of the inequality assuming this worst-case scenario, as a function of the subject's own classifier accuracy. The figure shows that it is satisfied at a  $p=0.6$  classifier accuracy for group sizes of  $N=5, 10, 15$  and  $20$ .

Due to a small computational mistake during the design of the experiment, VP is satisfied for the case of  $N=25$  only when the classification rate is 61%. Fixing this numerical

glitch is mathematically trivial and experimentally uninteresting, since the subjects were not given the option to opt out of individual trials. Also, note that the VP constraint was satisfied at the subjects' stated classification rate of 64%. Since this is the rate that determines subjects' behavior, we can conclude that VP was satisfied for the case of  $N=25$ .

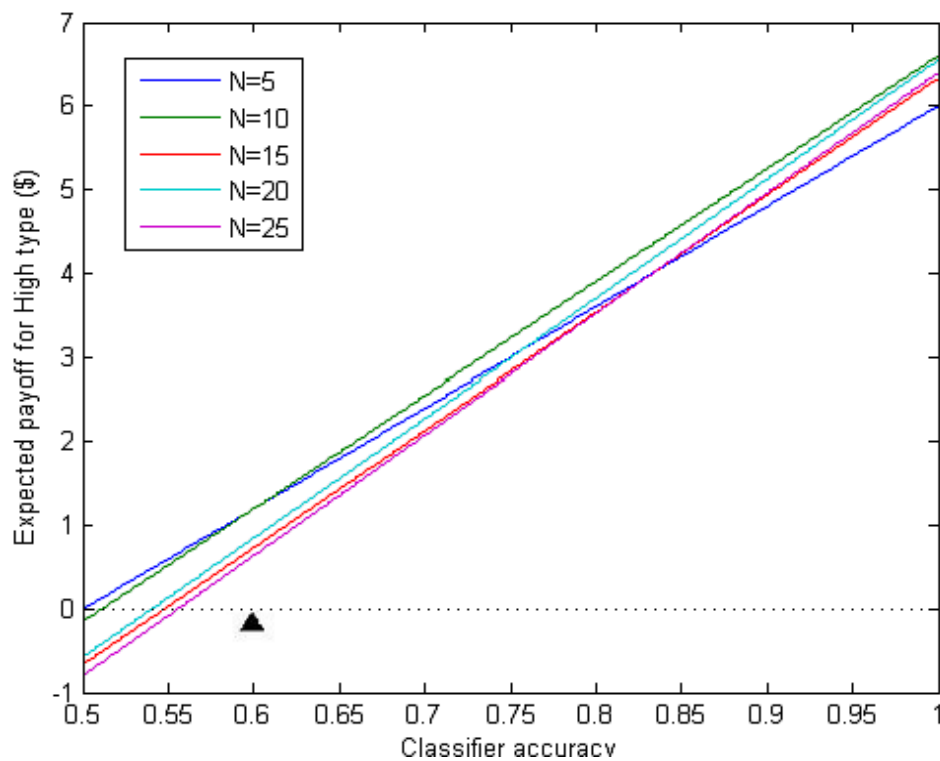
### *Budget-Balanced*

The NIM is budget-balanced by design. In particular, the rebalancing transfers  $\mathbf{b}(x, \mathbf{s}(\theta))$  ensure that the amount of taxes collected exactly equals the cost of the public good.

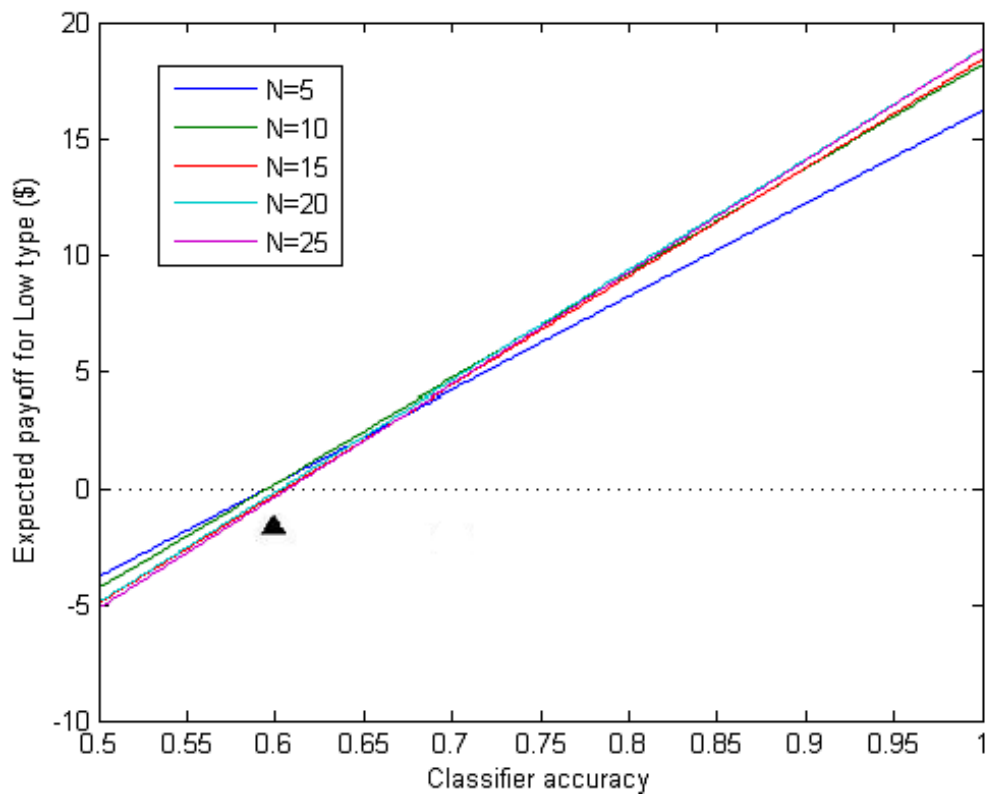
### *Social Efficiency*

Social efficiency requires that the optimal level of the public good be selected for any distribution of types in the group. But this follows directly from the rules of the NIM as long as subjects report optimally, which is guaranteed by DSIC.

**Figure 5.** Expected payoff for a truth-telling *High* type in the “worst-case” scenario. The x-axis denotes the classifier’s accuracy for the subject being studied. The classification accuracy for all other subjects is assumed to be at 60%.



**Figure 6.** Expected payoff for a truth-telling *Low* type in the “worst-case” scenario. The x-axis denotes the classifier’s accuracy for the subject being studied. The classification accuracy for all other subjects is assumed to be at 60%.



## VII. Additional properties of the NIM

Note that the NIM is characterized by its predictive accuracy and by the values of the four possible baseline taxes: one for each possible combination of reported value and classified value.

The goal of the mechanism designer is to pick these taxes so that the properties of DSIC, SE, VP, and BB are satisfied. This leads to a linear programming problem that has many possible solutions. Because there are many solutions, other criteria are imposed to choose a precise tax scheme. The tax values used in the experiment were selected with several additional criteria in mind: (1) to have a higher overall expected payoff for *High* types compared to *Low* types, (2) to provide a strictly positive incentive to report truthfully for both *High* and *Low* types, and (3) to minimize the effects of possible risk-aversion by limiting the variance of the potential payoffs. The fact that subjects behaved as predicted (typically revealing actual values in their reports), suggests that risk-aversion was not a problem, although more is discussed in the Appendix.

### VIII. Optimal mechanism in the absence of informative type signals

In this section we derive the optimal mechanism that does not have access to informative neural signals. Since it is well known that in this case there is no mechanism satisfying DSIC, VP, BB, and SE, we characterize the mechanism that generates the highest possible social efficiency while satisfying DSIC, VP and BB.

The measure of efficiency is given by average surplus:

$$\frac{1}{N} \left( \sum_{i=1}^N v_i - c \right)$$

By the revelation principle (36-38) subjects are simply asked to reveal their types: *High* or *Low*. In searching for the best mechanisms, we allowed the probability of producing the public good to vary from 0 to 1 as a function of the reported types, and we allowed any finite taxes (positive or negative) conditional on the subjects' reports.

The characterization of the optimal mechanism boils down to a linear programming problem. In order to define it we need to introduce the following notation:

-  $k \in [0, N]$  denotes the number of reported *High* types.

- $t_H(k, N)$  denotes the tax paid by a *High* type with  $k$  reported *High* types in a group of size  $N$ .
- $t_L(k, N)$  denotes the tax paid by a *Low* type with  $k$  reported *High* types in a group of size  $N$ .
- $\delta(k, N)$  denotes the probability that the public good is produced with  $k$  reported *High* types in a group of size  $N$ .
- $g(k, N)$  denotes the empirical probability of having  $k$  *High* types in a group of size  $N$ , given the distribution of values.

The mechanism needs to satisfy the following constraints:

- 1) Budget-balanced: For all  $k$  and  $N$ ,

$$k * t_H(k, N) + (N - k) * t_L(k, N) = C * \delta(k, N)$$

- 2) DSIC for the *High* types: For all  $k$  and  $N$ ,

$$g * \delta(k, N) - t_H(k, N) > g * \delta(k - 1, N) - t_L(k - 1, N)$$

- 3) DSIC for the *Low* types: For all  $k$  and  $N$ ,

$$1 * \delta(k - 1, N) - t_L(k - 1, N) > 1 * \delta(k, N) - t_H(k, N)$$

4) Voluntary participation for the *High* types: For all  $k$  and  $N$ ,

$$g + \delta(k, N) - t_H(k, N) > 0$$

5) Voluntary participation for the *Low* types: For all  $k$  and  $N$ ,

$$1 + \delta(k-1, N) - t_L(k-1, N) > 0$$

The optimal mechanism is given by the rules  $\delta(k, N)$ ,  $t_H(k, N)$ , and  $t_L(k, N)$  that satisfies these five constraints and maximizes the expected surplus function

$$\sum_{k=0}^N q(k, N) + \delta(k, N) + (k + g + (N - k) + 1 - C)$$

given the empirical distribution of *High* and *Low* types. The best mechanism satisfying these constraints is detailed in Table 3, and the average surplus that it generates is depicted in Figure 3C.

To generate the average surplus from each mechanism in Figure 3C, we took each of the 489 valid trials sequentially (11 were misses) from the main experiment, randomly completed the groups using information from other subjects' trials with the same  $N$ , and then determined the outcomes using the rules of the mechanism.



**Table 3.** Probabilities of providing the public good with the optimal non-NIM mechanism satisfying DSIC, VP, and BB. Each column is a different group size and each row is the number of reported *High* types.

# of <i>High</i> types	$N = 5$	$N = 10$	$N = 15$	$N = 20$	$N = 25$
1	0	0	0	0	0
2	0	0	0	0	0
3	0.1875	0	0	0	0
4	0.5	0	0	0	0
5	1	0.0031	0	0	0
6	X	0.0238	0	0	0
7	X	0.0833	0	0	0
8	X	0.2222	0.0005	0	0
9	X	0.5	0.0029	0	0
10	X	1	0.0117	0	0
11	X	X	0.0368	0.0001	0
12	X	X	0.0982	0.0004	0
13	X	X	0.2321	0.0015	0
14	X	X	0.5	0.0054	0
15	X	X	1	0.0162	0
16	X	X	X	0.0433	0.0002
17	X	X	X	0.1052	0.0007
18	X	X	X	0.2368	0.0025
19	X	X	X	0.5	0.0072
20	X	X	X	1	0.0191
21	X	X	X	X	0.0472
22	X	X	X	X	0.1094
23	X	X	X	X	0.2396
24	X	X	X	X	0.5
25	X	X	X	X	1

## References

1. P. Samuelson, *Rev. Econ. Stat.* **36**, 387 (1954).
2. E. Lindahl, in *Classics in the Theory of Public Finance*, R. Musgrave, A. Peacock, Eds. (Macmillan, London, 1919), pp. 168-176.
3. H. Moulin, *Rev. Econ. Stud.* **61**, 305 (1994).
4. Y. Chen, in *The Handbook of Experimental Economics*, C. Plott, V. Smith, Eds. (Elsevier Press, Amsterdam, 2004).
5. J. Ledyard, in *Handbook of Experimental Economics*, J. Kagel, A. Roth, Eds. (Princeton University Press, Princeton, NJ, 1995).
6. S. Gailmard, T. Palfrey, *J. Pub. Econ.* **89**, 1361 (2005).
7. E. Fehr, S. Gächter, *Nature* **415**, 137 (2002).
8. J. Falkinger, E. Ferhr, S. Gächter, R. Winter-Ebmer, *Am. Econ. Rev.* **90**, 247 (2000).
9. B. Hermann, S. Gächter, C. Thoni, *Science* **319**, 1362 (2008).
10. D. Diamantaras, E. Cardamone, K. Campbell, S. Deacle, L. Delgado, *A Toolbox for Economic Design*. (Palgrave Macmillan, New York, NY, 2009).
11. A. Gibbard, *Econometrica* **41**, 587 (1973).
12. R. B. Myerson, in *The New Palgrave Dictionary of Economics Online*, S. N. Durlauf, L. E. Blume, Eds. (Palgrave Macmillan, 2008).
13. J. Green, J. Laffont, *Econometrica* **45**, 427 (1977).
14. T. Groves, J. Ledyard, *Econometrica* **45**, 783 (1977).
15. E. Maskin, *Rev. Econ. Stud.* **66**, 23 (1999).
16. J. V. Haxby *et al.*, *Science* **293**, 2425 (2001).
17. J. D. Haynes, G. Rees, *Curr. Biol.* **15**, 1301 (2005).
18. Y. Kamitani, F. Tong, *Nat. Neurosci.* **8**, 679 (2005).
19. K. N. Kay, T. Naselaris, R. J. Prenger, J. L. Gallant, *Nature* **452**, 352 (2008).
20. A. J. O'Toole *et al.*, *J. Cogn. Neurosci.* **19**, 1735 (2007).
21. L. Pessoa, S. Padmala, *Cereb. Cortex.* **17**, 691 (2007).
22. S. M. Polyn, V. S. Natu, J. D. Cohen, K. A. Norman, *Science* **310**, 1963 (2005).

23. J. T. Serences, G. M. Boynton, *Neuron* **55**, 301 (2007).
24. D. Abreu, H. Matsushima, *Econometrica* **60**, 1439 (1992).
25. A. Smith, *An Inquiry into the Nature and Causes of the Wealth of Nations*. (Regnery Publishing, Washington DC, 1776).
26. T. A. Hare, J. O'Doherty, C. F. Camerer, W. Schultz, A. Rangel, *J. Neurosci.* **28**, 5623 (2008).
27. R. M. Isaac, D. James, *J. Risk Uncertainty* **20**, 177 (2000).
28. H. Plassmann, J. O'Doherty, A. Rangel, *J. Neurosci.* **27**, 9984 (2007).
29. C. R. Plott, K. Zeiler, *Am. Econ. Rev.* **95**, 530 (2005).
30. J. Roosen, J. A. Fox, D. A. Hennessy, A. Schreiber, *J. Agr. Resour. Econ.* **23**, 367 (1998).
31. B. E. Boser, I. Guyon, V. N. Vapnik, in *Annual Workshop on Computational Learning Theory*. (Pittsburgh, 1992).
32. C. C. Chang, C. J. Lin. (2001).
33. I. Guyon, S. Gunn, M. Nikravesh, L. Zadeh, *Feature Extraction: Foundations and Applications*. J. Kacprzyk, Ed., Studies in Fuzziness and Soft Computing (Springer, Berlin, 2006), vol. 207.
34. C. C. Chang, C. J. Lin, (2001).
35. B. E. Boser, I. M. Guyon, V. N. Vapnik, paper presented at the Annual Workshop on Computational Learning Theory, Pittsburgh, PA, 1992.
36. A. Gibbard, *Econometrica* **41**, 587 (1973).
37. J. Green, J. Laffont, *Econometrica* **45**, 427 (1977).
38. T. Groves, J. Ledyard, *Econometrica* **45**, 783 (1977).

## APPENDIX

### **Robustness of the NIM in the presence of risk/loss-aversion**

The theoretical results derived in Section VI of the Methods assume that individuals are risk- and loss-neutral (i.e., they make choices to maximize their expected financial payoffs). However, a wealth of experimental and field evidence suggest that subjects often exhibit substantial degrees of risk- and loss-aversion (*S1-S3*). Risk-aversion refers to the tendency to value a risk at a certainty-equivalent that is below its expected monetary value, and is usually parameterized by a concave utility function of money. Loss-aversion is the tendency to weight the disutility of losses more highly than the utility of equal-sized gains.

In this section we explore the role of both risk- and loss-aversion on our results. We do this in two parts. First, we explore what happens to the properties of the specific NIM actually used in the experiment when subjects are not risk- and loss-neutral. Second, we show that it is possible to construct alternative NIMs that are able to satisfy the four desired properties (DSIC, SE, BB, VP) even when there are large degrees of risk- and loss-aversion.

All of the arguments below assume that subjects make choices to maximize a Prospect Theoretic utility function (S3) of the form

$$u(x) = \begin{cases} x^\alpha & \text{if } x \geq 0 \\ -\lambda|x|^\alpha & \text{otherwise} \end{cases}$$

In this formulation  $\alpha$  measures the degree of risk-aversion; as  $\alpha$  falls below 1 the function for gains becomes more curved and risk-aversion is higher. The parameter  $\lambda$  measures the degree of loss-aversion. Note that this includes the case of risk-neutrality (expected-value maximization) as a special case when  $\alpha = \lambda = 1$ .

*A. The impact of introducing risk- and loss-aversion on the NIM used in the experiment.*

It is straightforward to see that this has no effect on the budget-balancing or social efficiency properties.

Now consider the case of VP under the assumption, used in the experiment, that the classifier accuracy rate is 60%. As in the proof above, to show that VP is satisfied for a type it suffices to show that it is satisfied in the worst-case scenario for the behavior and

types of the other players (see the voluntary participation section above for details). Figure S1 plots the expected utility of a truth-telling *High* type as a function of his loss-aversion and risk-aversion parameters for the case of  $N=5$ . (All figures in this section are done for groups of size  $N=5$ , but differences in group size have only very small effects). The regions in which expected utility is positive correspond to the sets of  $(\alpha$  and  $\lambda)$  values for which the VP constraint is satisfied. The figure shows that VP is satisfied at the special case of risk- ( $\alpha = 1$ ) and loss-neutrality ( $\lambda = 1$ ) and that satisfaction of the VP constraint is robust to the introduction of significant amounts of risk- and loss-aversion. Figure S2 plots the same thing for a truth-telling *Low* type. A comparison of the two figures shows that increases in the loss-aversion coefficient can cause the VP constraint to fail, unless the subject is also very risk-averse.

Finally consider the DSIC property. As in the proof above, we need to show that the incentive compatibility for each type is satisfied in each of three possible cases:

- 1) The public good is not produced regardless of  $i$ 's report.
- 2) The public good is produced regardless of  $i$ 's report.
- 3) The public good is produced if  $i$  reports *High*, but not if he reports *Low*.

The argument for the first case is not affected by the introduction of risk- or loss-aversion, and thus the incentive compatibility constraint still holds in that case.

Now consider the second case, in which the public good is produced regardless of  $i$ 's report. Figures S3 and S4 plot the change in expected utility between truth-telling and lying for a *High* and a *Low* type, respectively. The plots are generated assuming three other subjects reported *High* in the group (the minimum necessary for the good to be produced). The other omitted cases generate very similar conclusions. The two figures show that for the range of risk- and loss-aversion parameters considered, the incentive to report truthfully is positive for both types.

Now consider the third case. There are two sub-cases to consider. First is the sub-case of a *Low* type who could report *High*. Figure S5 plots the change in expected utility between lying and telling the truth in this case. We can see from this plot that the incentive to misreport is negative, so decisive *Low* types will want to report truthfully in the range of parameters described. Second is the sub-case of a *High* type who could report *Low*. As was the case in the risk/loss-neutrality case, the conditions for this incentive constraint are identical to the ones for the voluntary participation constraint described in Figure S1.

To summarize, these arguments show that the introduction of risk- and loss-aversion to the analysis of properties of the NIM used in the experiment does not reverse individuals' incentives to report truthfully in the NIM, but that the voluntary participation constraint can be violated when individuals have significant amounts of loss-aversion.

*B. Alternative NIM mechanisms.*

There are at least two ways to alleviate the problems introduced by risk- and loss-aversion described above.

The first solution is to improve the classification accuracy. As the classification rate improves, the variance of the baseline taxes can be reduced, thus reducing the effects of risk/loss-aversion. In fact, it is straightforward to see that in the extreme case of full predictability ( $p=1$ ), the government can impose the socially efficient allocation while assigning the costs of the public good to the different types in a way that does not involve any uncertainty.

The second solution is to modify the NIM by changing the baseline taxes. This is the recommended approach if a mechanism designer had reason to believe the target population of participants was averse to risk or loss and had some prior beliefs over the risk and loss parameters. Figure S6 describes an alternative set of potential baseline taxes for different prediction rates. Note that the magnitudes of the taxes decrease drastically as the prediction rate improves and that the baseline tax for a reported *High* type is always \$10. These taxes were chosen so that *High* types would be indifferent with respect to VP and IC in the worse-case scenario where the sum of the expected baseline taxes by the



other subjects is minimized (see Section VI for a detailed description of how this worst case is defined).

Note that since the only variables that have changed are the baseline taxes, budget-balance and social efficiency are still satisfied by the new NIM.

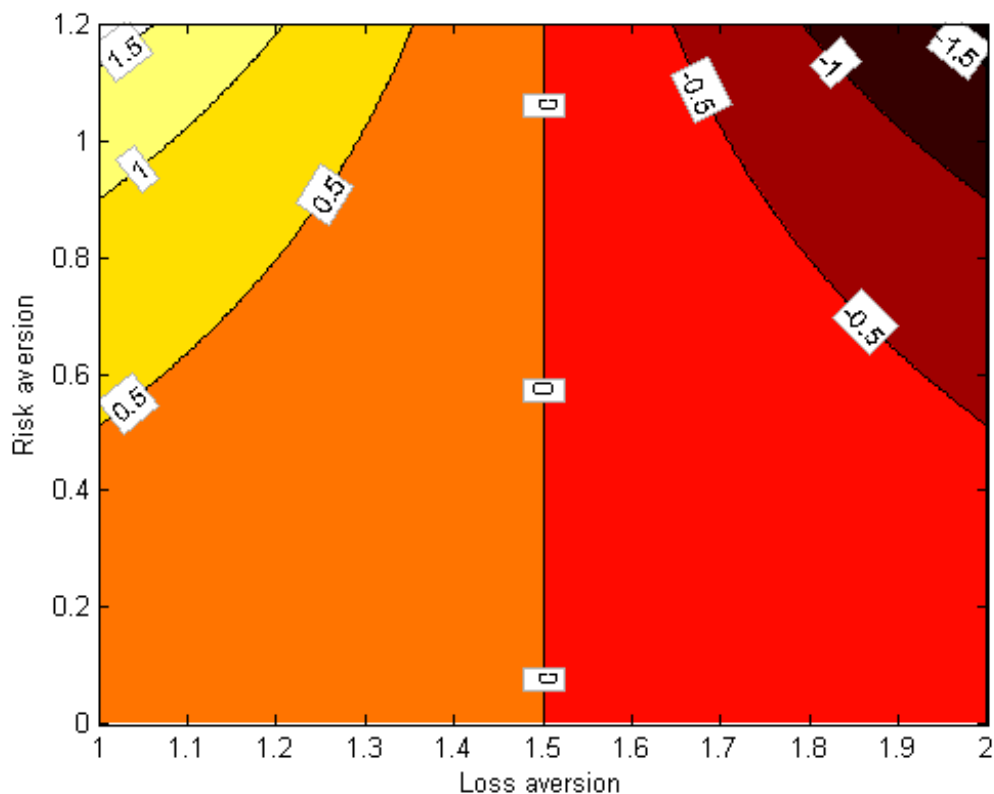
What about voluntary participation? The VP constraint of the *High* type is trivially satisfied since it defines the taxes used. Figure S7 shows the expected VP constraint for the *Low* types, assuming a classifier accuracy of 60%, and for the worst-case scenario in which the sum of the baseline taxes by the other subjects is minimized. A comparison with Figure S2 shows that the region of parameters in which the VP holds has significantly expanded, and in particular that the VP constraint is now satisfied for moderate risk/loss-aversion.

What about DSIC? As before, to prove that the DSIC property holds we need to consider several cases. Since the arguments are extremely similar, here we only describe the key steps. Figures S8 and S9 plot the expected utility change between truth-telling and lying, for *High* and *Low* types, respectively. These plots are generated assuming three other reported *High* types in the group (the minimum necessary for the good to be produced), which means that the subject is not decisive. For a decisive *Low* type there is no benefit to lying since he would always earn a negative payoff.

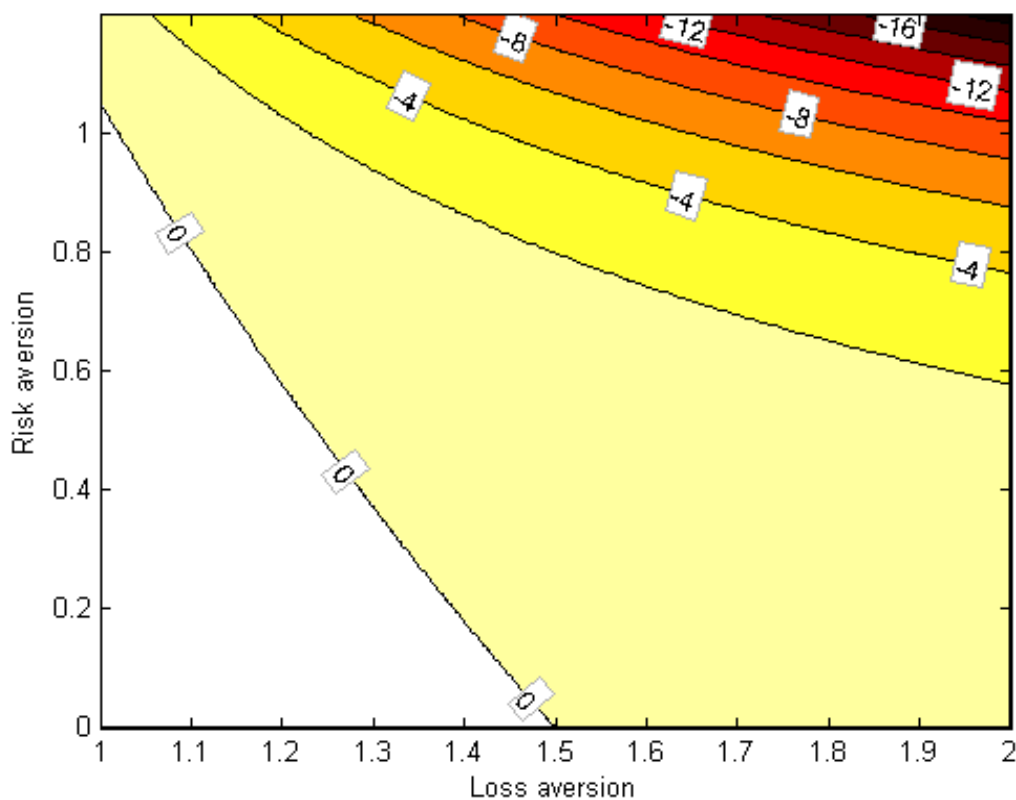
One natural question is how the VP constraint for the *Low* types changes in this class of NIMs as the prediction rate increases. Figures S10 and S11 plot the expected utility of *Low* types for classifier accuracies of 70% and 80%, respectively. We can see that as the classifier accuracy improves, the region of positive utilities greatly expands and VP will therefore be satisfied for any reasonable amount of risk/loss-aversion.

In conclusion, the analyses and arguments in this section show that it is possible to modify the NIM in ways that can accommodate moderate risk- and loss-aversion for the levels of classifier accuracy obtained in the experiment, and that the possible sensitivities of DSIC and VP satisfaction to risk- and loss-aversion are further reduced as the classifier accuracies improve.

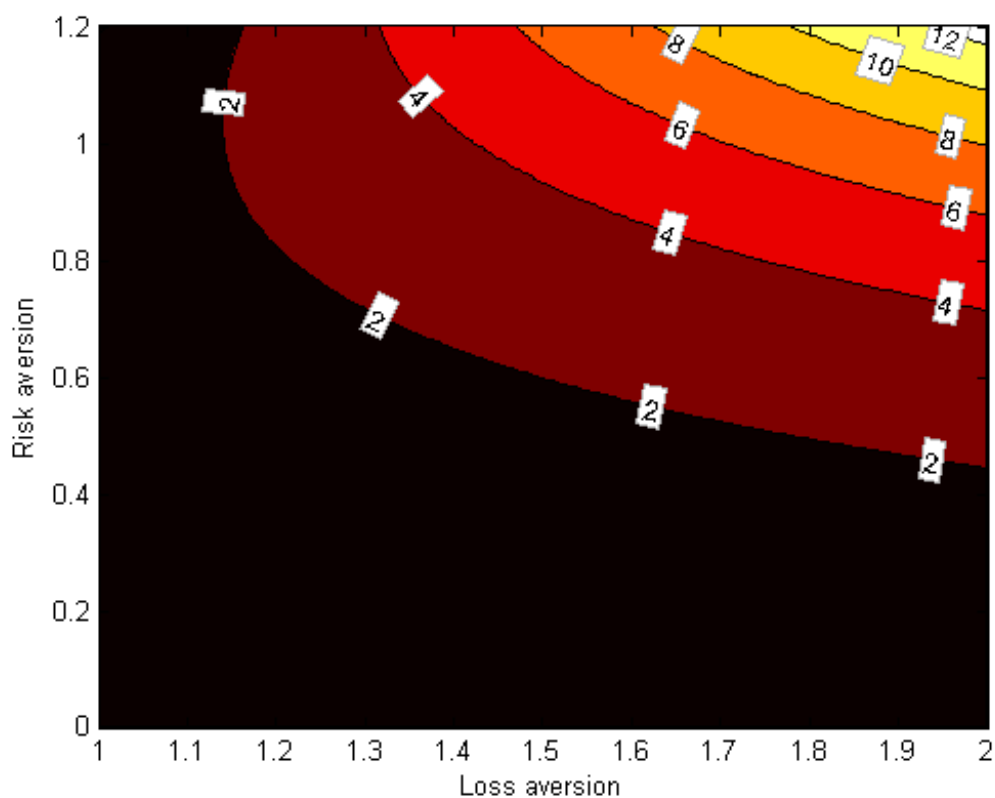
**Figure S1.** Contour plot of the expected utility for a truth-telling *High* type, as a function of the risk- and loss-aversion parameters, for a group size of  $N=5$ , classifier accuracy of 60%, and under the “worst-case scenario” in which the sum of the expected baseline taxes paid by the other players is minimized. For VP to be satisfied, this needs to be positive.



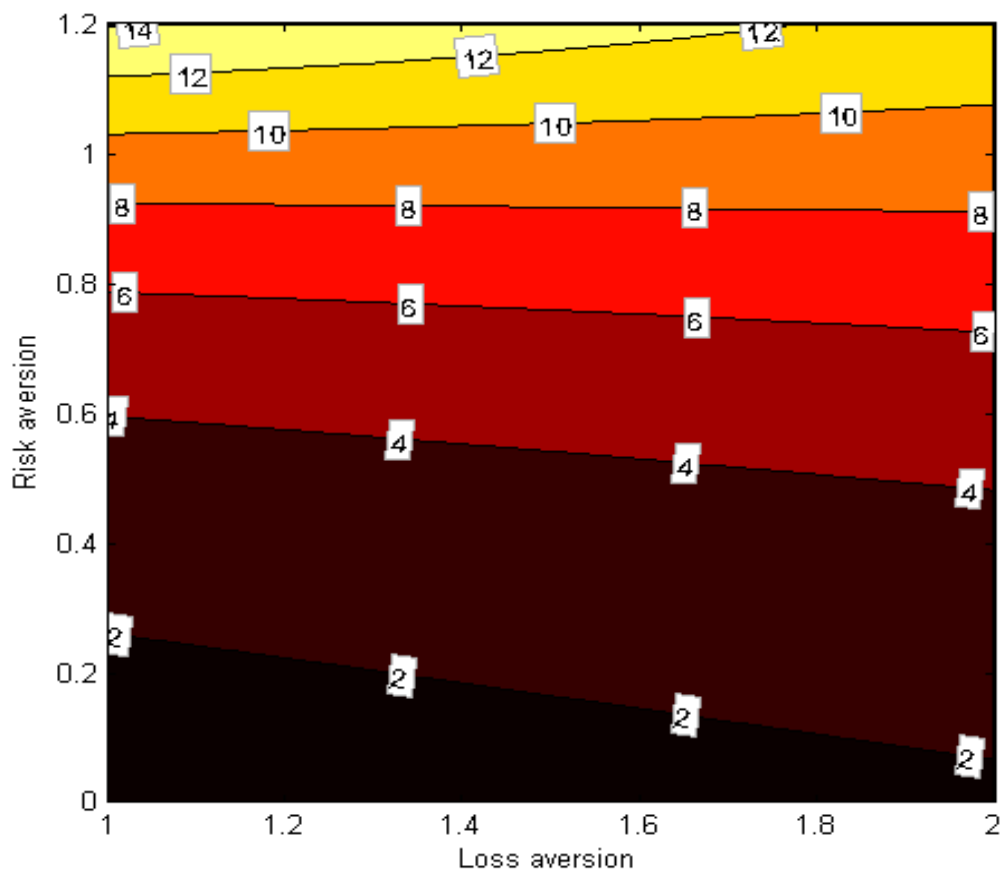
**Figure S2.** Contour plot of the expected utility for a truth-telling *Low* type, as a function of the risk- and loss-aversion parameters, for a group size of  $N=5$ , classifier accuracy of 60%, and under the “worst-case scenario” in which the sum of the expected baseline taxes paid by the other players is minimized. For VP to be satisfied, this needs to be positive.



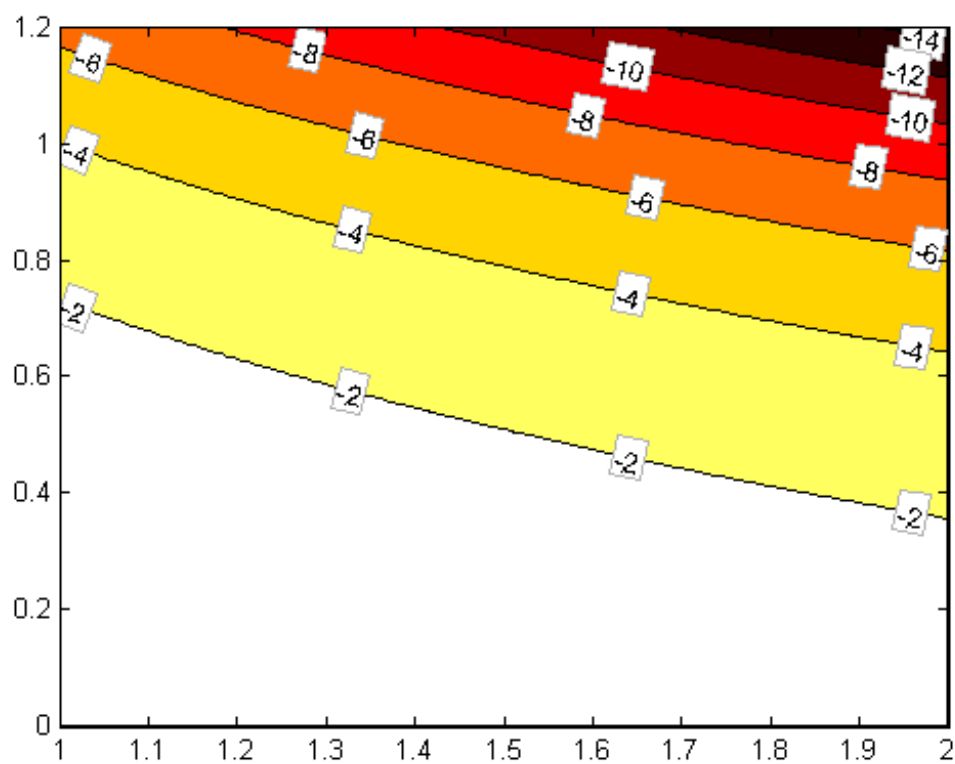
**Figure S3.** Contour plot of the expected utility improvement for truth-telling compared to lying for *High* types, as a function of the risk- and loss-aversion parameters, assuming a group size of  $N=5$ , classifier accuracy of 60%, and that three other truthful *High* types are in the group. For incentive compatibility to be satisfied, this needs to be positive.



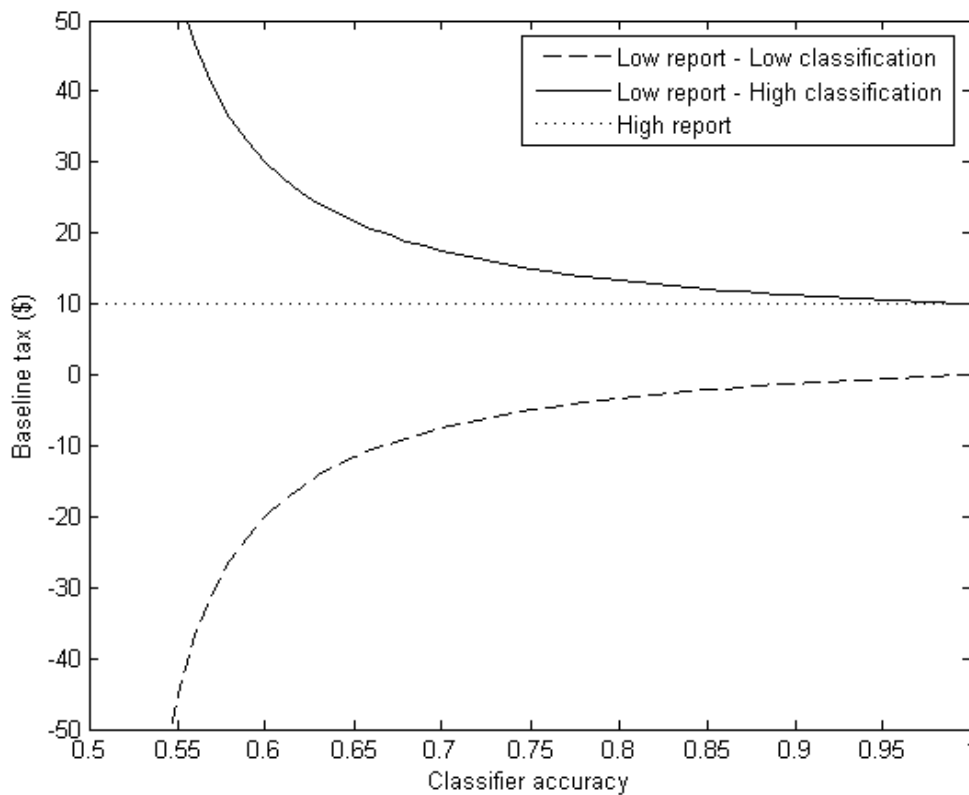
**Figure S4.** Contour plot of the expected utility improvement for truth-telling compared to lying for *Low* types, as a function of the risk- and loss-aversion parameters, assuming a group size of  $N=5$ , classifier accuracy of 60%, and that three other truthful *High* types are in the group. For incentive compatibility to be satisfied, this needs to be positive.



**Figure S5.** Contour plot of the expected utility change for a decisive *Low* type associated with misreporting being a *High* type, shown for group size  $N=5$ , and classifier accuracy of 60%. For incentive compatibility to be satisfied, this needs to be negative.

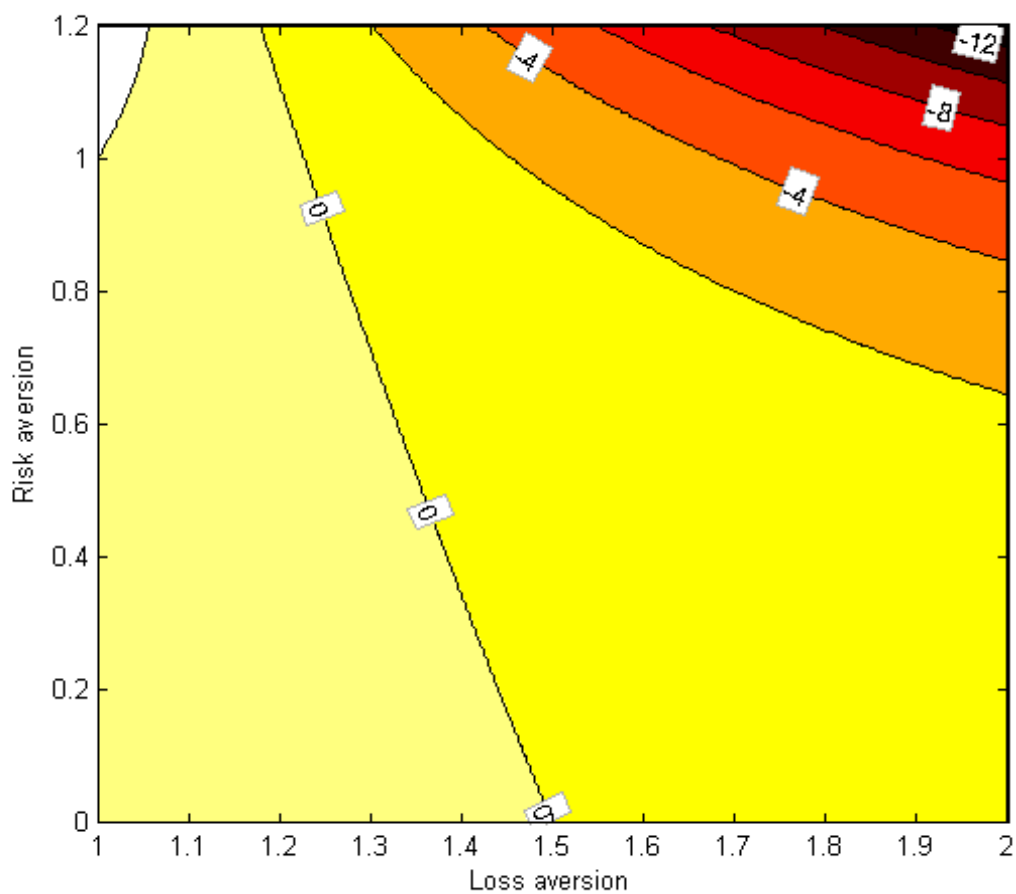


**Figure S6.** Baseline taxes as a function of the classifier accuracy. Note that the tax for a reported *High* type is the same regardless of the classifier prediction. Positive numbers indicate taxes, negative numbers indicate transfers to the subjects.

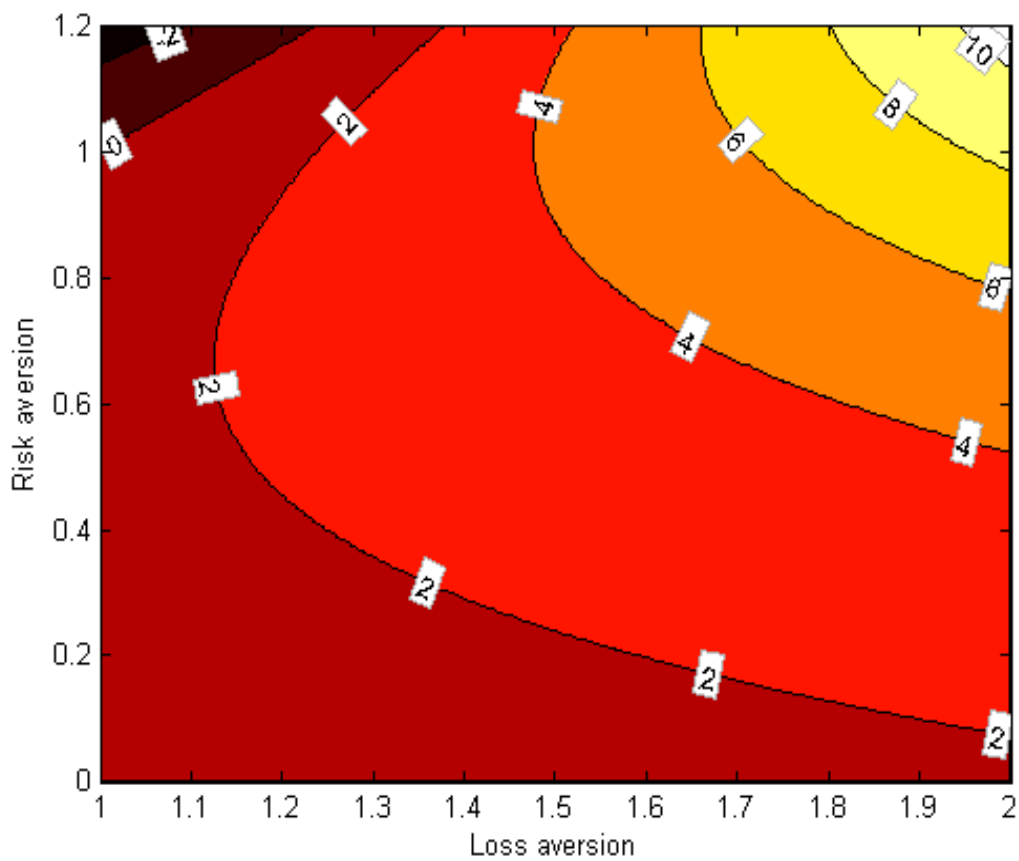




**Figure S7.** Contour plot of the expected utility for a truth-telling *Low* type in the alternative NIM, as a function of the risk- and loss-aversion parameters, for a group size of  $N=5$ , classifier accuracy of 60%, and under the “worst-case scenario” in which the sum of the expected baseline taxes paid by the other players is minimized. For VP to be satisfied, this needs to be positive.

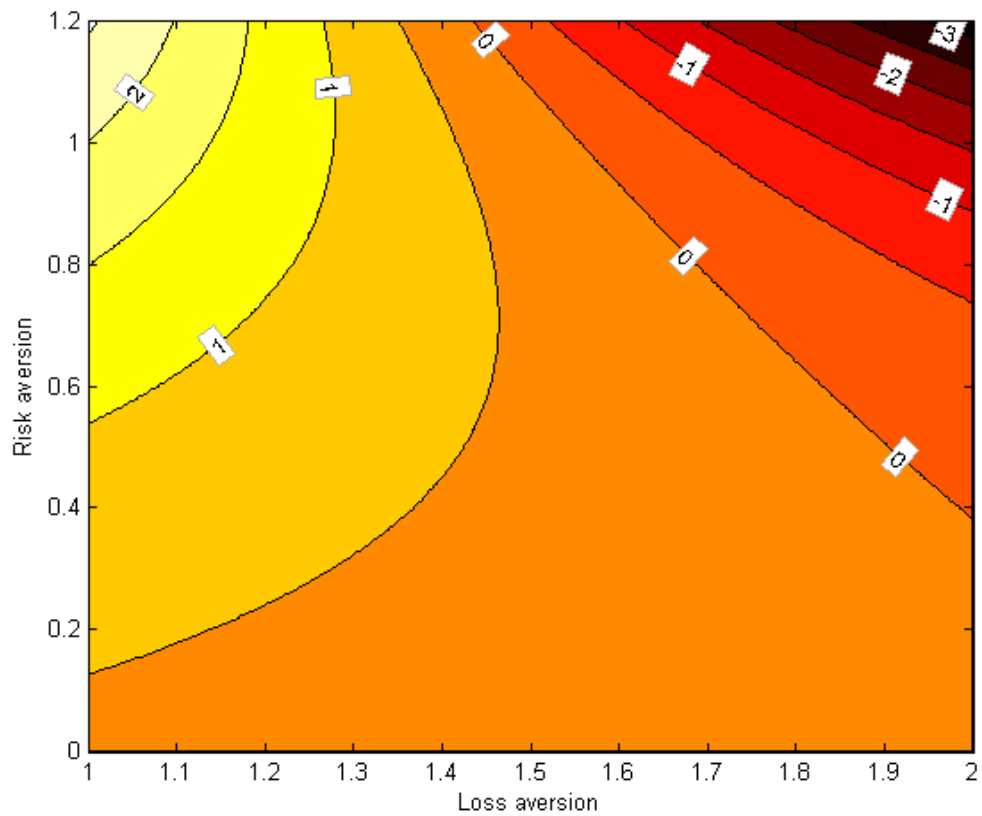


**Figure S8.** Contour plot of the expected utility improvement for truth-telling compared to lying for *High* types in the alternative NIM, as a function of the risk- and loss-aversion parameters, assuming a group size of  $N=5$ , classifier accuracy of 60%, and that three other truthful *High* types are in the group. For incentive compatibility to be satisfied, this needs to be positive

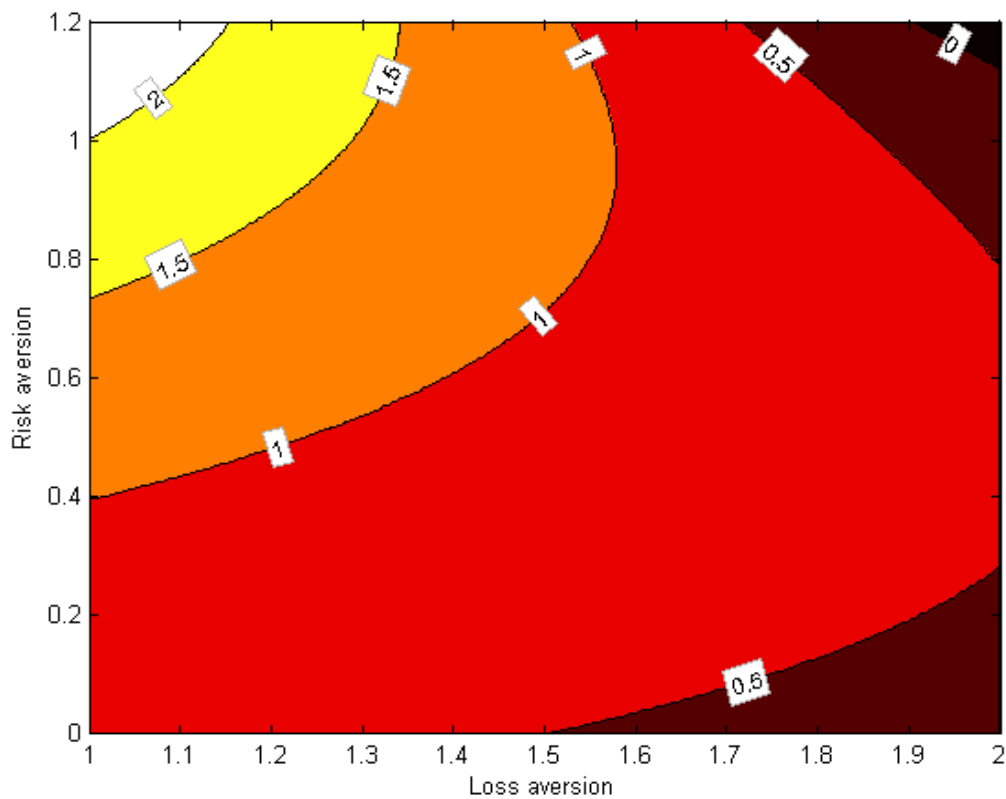




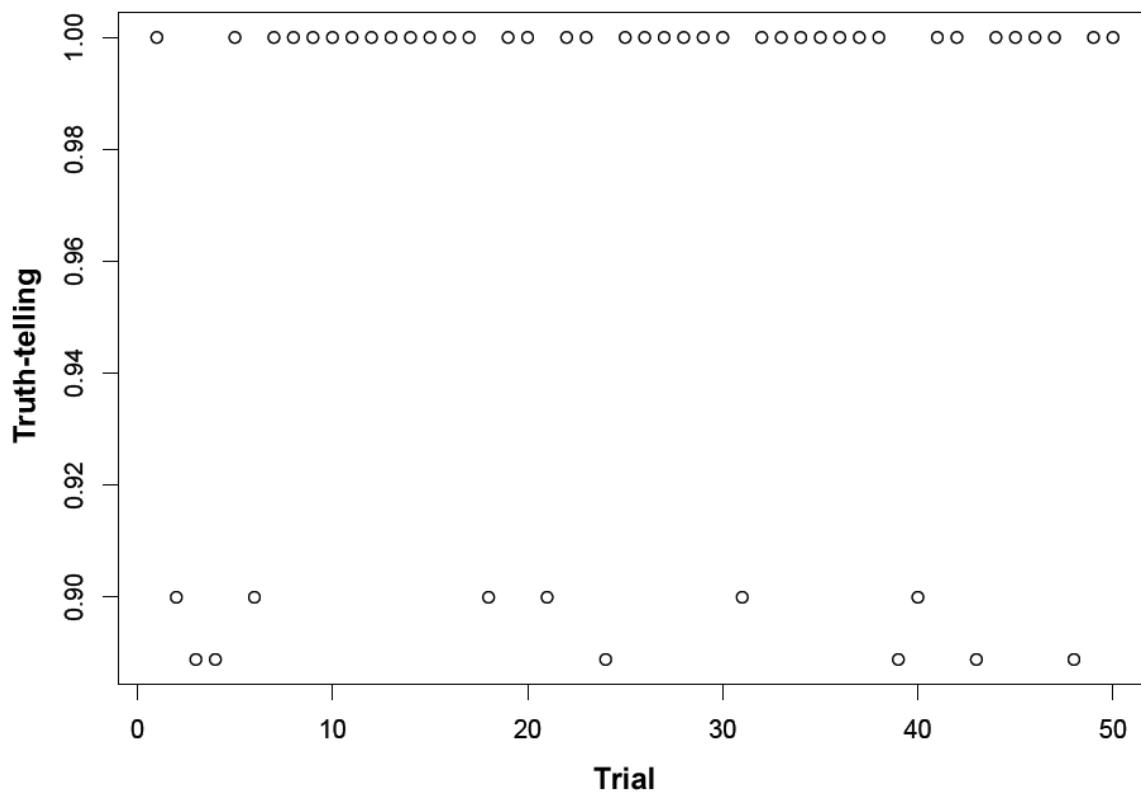
**Figure S10.** Contour plot of the expected utility for a truth-telling *Low* type in the alternative NIM, as a function of the risk- and loss-aversion parameters, for a group size of  $N=5$ , classifier accuracy of 70%, and under the “worst-case scenario” in which the sum of the expected baseline taxes paid by the other players is minimized. For VP to be satisfied, this needs to be positive



**Figure S11.** Contour plot of the expected utility for a truth-telling *Low* type in the alternative NIM, as a function of the risk- and loss-aversion parameters, for a group size of  $N=5$ , classifier accuracy of 80%, and under the “worst-case scenario” in which the sum of the expected baseline taxes paid by the other players is minimized. For VP to be satisfied, this needs to be positive



**Figure S12.** Frequency of truth-telling as a function of time (trials). Trials in which subjects did not respond in time were dropped from this analysis.



## Appendix References

- S1. C. A. Holt, S. K. Laury, *Am. Econ. Rev.* **92**, 1644 (2002).
- S2. A. Tversky, D. Kahneman, *J. Risk Uncertainty* **5**, 297 (1992).
- S3. P. Sokol-Hessner *et al.*, *P. Natl. Acad. Sci. USA* **106**, 5035 (2009).

### **Instructions**

This experiment is a study of group investment behavior.

There is NO deception in the experiment: if we tell you that we are going to do something, we will do it exactly as described.

By showing up to the instruction session you have already earned \$20 today. By showing up to the experiment you will earn an additional \$20, and then there will be more opportunities to earn money based on your performance in the experiment. You will receive some of that money at the end of the experiment, and the rest you will receive when we are done scanning the other subjects. (Instructions about how to collect the extra earnings will be given at the end of the experiment).

In each round of the experiment you will be pooled into a group consisting of either 5, 10, 15, 20, or 25 players (including yourself). The compositions of the group are random and change from trial to trial. The other players in your group will be actual people who have gone through this exact experiment as well. Each player in the group has a different value for the investment and these values change from round to round according to the rules described below. You can either have a HIGH value (\$8-10) or a LOW value (\$0-2) for the investment. For each group size there is also a different total cost to the group of the investment. Whether the investment is made will partially depend on this cost, as described below.



This experiment has one unusual feature. Each round, we will be applying a statistical algorithm to measures of your brain activity to guess whether your value for that round is HIGH (\$8-\$10) or LOW (\$0-\$2). The values of the other players in the groups will be predicted in the same way. Each round, we will also be asking you to report your value to us.

To predict your value in a given round, our statistical algorithm will look at the data from the other rounds and learn how your brain responds when you see a HIGH value compared to a LOW value. Then for the target round we are trying to predict, the algorithm will look at your brain activity at the time you see your value and try to guess whether that value is HIGH or LOW, by matching the pattern of activity in the target round to the other rounds. If the target-round activity is closer to activity in other HIGH-value rounds, the algorithm will guess that your target-round activity is HIGH. If the target-round activity is closer to activity in other LOW-value rounds, the algorithm will guess that your target-round activity is LOW.

You are free to try to control your brain activity in any way to affect how well the statistical algorithm can guess your value, and you are also free to misreport your value to us. However, you are required to:

- Keep your eyes open at all times (except for casual blinking)
- Look at the information on the screen

**VERY IMPORTANT: Failure to follow these instructions will compromise the integrity of the experiment and will be considered a violation of your subject responsibilities.**

The experiment will consist of 50 rounds, 10 each with group sizes of 5, 10, 15, 20, and 25. The order of the rounds is completely random and each round is independent of all the others. Your payoff for each round is determined by several things: your value for the investment, your reported value, your predicted value, and whether or not the investment is made:

- We will take the reported values from each player in the group. If the sum of those values is greater than the total cost of the investment, then the investment will be made (we assume HIGH value = 9 and LOW value = 1). If the sum of those values is less than the total cost of the investment, then the investment will not be made.
- If the investment is made, you will earn an amount of money equal to your investment value. In addition, you will also pay or receive one of the four following amounts:
  - If your reported value is HIGH and your predicted value is HIGH then you will pay \$1.
  - If your reported value is HIGH and your predicted value is LOW then you will pay \$16.
  - If your reported value is LOW and your predicted value is HIGH then you will pay \$30.
  - If your reported value is LOW and your predicted value is LOW then you will receive \$20.
- We will take these payments (which may be negative, indicating a payment to you) from you and every other player in the group and then use them to pay the cost of the investment. If there is money left over after paying the cost, that amount will be redistributed evenly between the players in the group. If on the other hand there is not enough money to pay the cost, the extra amount needed will be collected evenly from the players in the group.

If the investment is not made then you receive no payoff for that round.

The total costs for the 5 different group sizes are as follows (they are always the same):

Group size = 5, Total cost = \$25

Group size = 10, Total cost = \$50

Group size = 15, Total cost = \$75

Group size = 20, Total cost = \$100

Group size = 25, Total cost = \$125

You do not need to memorize these values since they will be displayed each round before you make a decision.

After we have data from all the players, we will take the 50 rounds that you have played, randomly select matching rounds from the other subjects to fill in the groups, and then calculate your average payoff. You will earn the average payoff from all 50 rounds, multiplied by a factor of 5.

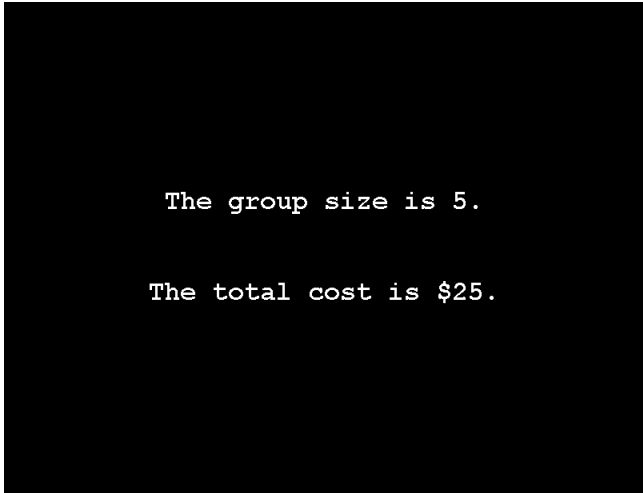
Procedure:

In each round you will first see a screen with your value for the investment. This value is different each round, and is randomly drawn from the intervals \$0-2 and \$8-10. Which interval the value is chosen from is also random. This means that in any given round, your value is equally likely to be anywhere between \$0-2 and \$8-10. The value screen will be displayed for 4 seconds and look as follows:



Your value is \$9.91.

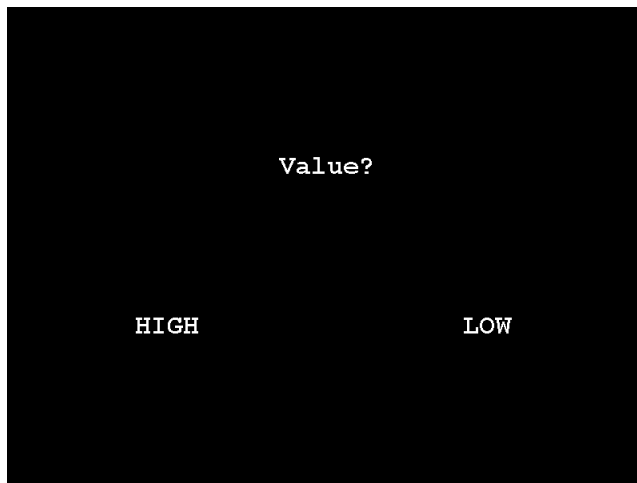
You will next see a screen with the group size and the total cost for that round. This information screen will be displayed for 10 seconds and look as follows:



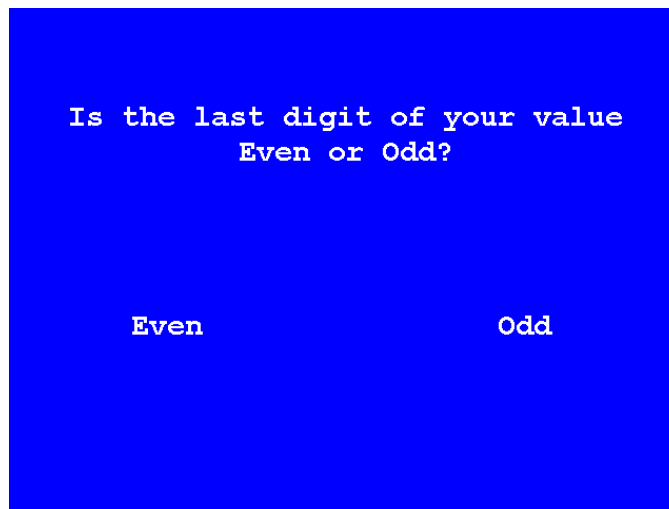
The group size is 5.  
The total cost is \$25.

Next, you will see a screen asking you to report your value, either HIGH or LOW. The position of HIGH and LOW on this screen is random (some rounds HIGH is on the left, some rounds HIGH is on the right). Press the “left” button if you would like to report the

LEFT value, or press the “right” button if you would like to report the RIGHT value. You only have 3 seconds on this screen to enter your decision, so move quickly. If you do not enter your choice in time then you will LOSE \$2 in that round, regardless of whether or not the investment is made. This screen will look as follows:



After you've made your choice on this screen the next round will begin, except under a special circumstance. In 8% of the trials (chosen randomly) we will be asking you at this point to tell us whether the last digit of your investment value is Even or Odd. Each time you get this question correct, you will earn an additional \$5 at the end of the experiment. For example, suppose that in a given round your value is \$9.36. In this case the correct answer would be 'Even' because the last digit of your value is 6, and 6 is an even number. If you chose 'Even' then you would receive a \$5 bonus; if you chose 'Odd' then you would not receive the bonus. You only have 4 seconds on this screen to enter your decision, so move quickly or you will not receive the bonus. The screen will look as follows:



After you've entered your choice on this screen the next round will begin.

Strategy:

In a previous version of this experiment our statistical algorithm was able to correctly predict players' values 60% of the time, on average. That is to say, if a player's true investment value was HIGH, our algorithm predicted a HIGH value with a probability of 0.6 and a LOW value with a probability of 0.4. Similarly, if a player's true investment value was LOW, our algorithm predicted a LOW value with a probability of 0.6 and a HIGH value with a probability of 0.4.

Given these probabilities, the payoffs are set up such that the best way to make money is to report your true investment value, regardless of what the other players are doing. If you misreport your value in a round, your expected payoff for that round will be worse than if you had reported your true value. For a mathematical explanation please see the appendix.

Again, you are free to misreport your value or try to manipulate your brain activity, but you are required to:

- Keep your eyes open at all times (except for casual blinking)
- Look at the information on the screen

You should however note the following:

(1) Making your brain activity more “noisy” could reduce the algorithm’s prediction accuracy downwards towards 50%. For any accuracy above 55%, reducing the accuracy will lower your expected payoffs, and below 55% your expected payoffs are constant. For details see the appendix.

(2) If, on the other hand, you can make your brain activity more consistent and drive the prediction accuracy up, this will actually increase your expected payoffs if you stick to a strategy of reporting your true investment value. As the algorithm’s prediction rate goes up, so does your expected payoff (if you report your true value). Again, feel free to work through the numerical calculations in the appendix to convince yourself that truth-telling is always your best strategy, and that if you stick to that strategy your expected payoff will go up as our ability to predict your value goes up.

#### **APPENDIX:**

All the following calculations leave out the extra payments for participation and the cost of the investment, but these do not change the results of any of these calculations.

**Best strategy is to tell the truth:**

- If your reported value is HIGH and your predicted value is HIGH then you will receive your investment value minus \$1.
- If your reported value is HIGH and your predicted value is LOW then you will receive your investment value minus \$16.
- If your reported value is LOW and your predicted value is HIGH then you will receive your investment value minus \$30.
- If your reported value is LOW and your predicted value is LOW then you will receive your investment value plus \$20.

Assuming the algorithm can indeed predict your value at a 60% rate, if you have a HIGH value then the probability of a HIGH signal is 0.6 and the probability of a LOW signal is 0.4. Therefore, if you report a HIGH value, then your expected payoff is:

$$\text{Value} + 0.6*(-\$1) + 0.4*(-\$16) = \text{Value} - \$7.$$

If you misreport a LOW value, then your expected payoff is:

$$\text{Value} + 0.6*(-\$30) + 0.4*(\$20) = \text{Value} - \$10.$$

This means that when you are a HIGH type, you will earn an average of \$3 per round more by reporting HIGH.

Similarly, if you have a LOW value then the probability of a LOW signal is 0.6 and the probability of a HIGH signal is 0.4.

Therefore, if you report a LOW value, then your expected payoff is:

$$\text{Value} + 0.6*(\$20) + 0.4*(-\$30) = \text{Value}.$$

If you have a LOW value and you misreport a HIGH value, then your expected payoff is:

$$\text{Value} + 0.6*(-\$16) + 0.4*(-\$1) = \text{Value} - \$10.$$



This means that when you are a LOW type, you will earn an average of \$10 per round more by reporting LOW.

**Driving down prediction accuracy:**

Consider the case where it is impossible for the algorithm to guess your investment value and so it predicts your correct value only 50% of the time:

If you have a HIGH value and you report a HIGH value, then your expected payoff is:

$$\text{Value} + 0.5*(-\$16) + 0.5*(-\$1) = \text{Value} - \$8.50.$$

If you have a HIGH value and you misreport a LOW value, then your expected payoff is:

$$\text{Value} + 0.5*(-\$30) + 0.5*(\$20) = \text{Value} - \$5.$$

If you have a LOW value and you report a LOW value, then your expected payoff is:

$$\text{Value} + 0.5*(-\$30) + 0.5*(\$20) = \text{Value} - \$5.$$

If you have a LOW value and you misreport a HIGH value, then your expected payoff is:

$$\text{Value} + 0.5*(-\$16) + 0.5*(-\$1) = \text{Value} - \$8.50.$$

This means that if you reported LOW in all HIGH rounds and LOW in all LOW rounds, which is the best you can do in this case, you would earn an average of (Value - \$5) every round. But in the case where we were predicting at a 60% rate, if you reported HIGH in all HIGH rounds and LOW in all LOW rounds, then you would earn an average of (Value - \$3.50). Feel free to work through the numbers for any other probability less than 0.6 to convince yourself that you will earn less money in that case.

Note that with this probability it is actually a better strategy to misreport your value when you are a HIGH type. However, this is only true up to a probability of 0.55. For any

higher prediction rate you will always earn more money by revealing your true investment value. And remember, at lower prediction rates you will be earning lower payoffs on average.

**Driving up prediction accuracy:**

If you have a HIGH value and you report a HIGH value, then your expected payoff is:

$$\text{Value} + 0.7*(-\$1) + 0.3*(-\$16) = \text{Value} - \$5.50.$$

If you have a HIGH value and you misreport a LOW value, then your expected payoff is:

$$\text{Value} + 0.7*(-\$30) + 0.3*(\$20) = \text{Value} - \$15.$$

If you have a LOW value and you report a LOW value, then your expected payoff is:

$$\text{Value} + 0.7*(\$20) + 0.3*(-\$30) = \text{Value} + \$5.$$

If you have a LOW value and you misreport a HIGH value, then your expected payoff is:

$$\text{Value} + 0.7*(-\$16) + 0.3*(-\$1) = \text{Value} - \$11.50.$$

As you can see, the best way to make money here is to report HIGH in HIGH rounds and LOW in LOW rounds. In this case you would earn an average of (Value - \$0.25), compared to the best you could do in the 60% case which averaged to (Value - \$3.50).

**QUIZ:**

Next you will answer a short quiz to ensure that you understand the instructions above.

If a statement is 'False', please rewrite the statement in a way that makes it 'True'.

Question 1: True or False: Your value for the investment is the same every round.

Question 2: True or False: To determine your investment value (and the investment values of other players) we randomly choose between the \$0-2 and the \$8-10 intervals, and then randomly choose an amount in that interval.

Question 3: How many rounds are there and how many of them will actually be used to determine your payment at the end of the experiment?

Question 4: True or False: Given that we can predict your value with a probability of 0.6, you will always make more money on average if you report your true investment value.

Question 5: True or False: If you were able to manipulate your brain activity to make it harder for the algorithm to predict your value, you would be able to earn more money.

Question 6: True or False: If you were able to manipulate your brain activity to make it easier for the algorithm to predict your value, you would be able to earn more money.

### Questionnaire

We would like to know how well you think we can predict your investment value in this experiment. The 60% rate we told you before the experiment was averaged across many different players, and the rate can vary from person to person. So, based on this information, please guess how well you think we'll be able to predict **your** values. To motivate you to think carefully, we will pay a bonus which is higher if you are closer to the actual percentage. Please give your answer in a 2-digit even integer, e.g. 56%, 60%, 72%, etc. (Note that since there are only 50 trials, and the algorithm predicts correctly or not on each trial, the percentage of correct guesses will be a multiple of 1/50 or 2%). If you guess the percentage exactly, we will pay you \$10. For each 2% step deviation from the true percentage (in either direction), \$1 will be subtracted. For example, if you guessed that we could predict your value 64% of the time and in fact we could predict it 70% of the time, you would receive \$7. You would receive \$7 because you started with \$10 but were off by 6%, which is three increments of 2% deviation. For each increment \$1 is subtracted, so \$3 is subtracted in total, which a net payment of \$7. You will receive this money along with the rest of your payoffs when you come back in a couple weeks. Now, please give us your guess:

The algorithm will be able to guess my value \_\_\_\_\_ % of the time.

## CHAPTER 2

### Neurometrically Informed Mechanism Design

#### I. Introduction

In the classical mechanism design problem, a social planner wants to implement an allocation that maximizes some notion of social welfare given the preferences of the group. The planner's problem is difficult, however, because he does not have direct knowledge about the individuals' preferences. Given this, the best he can do is to implement a mechanism that is characterized by a message space for each subject and a function mapping messages to outcomes. The mechanism solves the planner's problem if the outcomes induced in equilibrium, as a function of the underlying preferences, are the ones that the planner wants to implement.

A central preoccupation in the mechanism design literature has been to design mechanisms that satisfy three fundamental properties: efficiency, dominant strategy incentive compatibility, and voluntary participation. Efficiency requires that the outcome induced by the mechanism maximize the group's net expected utility (based on the underlying preferences) while also balancing the budget. This property is a basic requirement for a desirable mechanism. Voluntary participation requires that the expected utility from participating in the mechanism be positive for every subject regardless of her preferences and the actions of the other individuals. This property is desirable because it implies that all subjects benefit by participating. Finally, dominant strategy incentive

compatibility requires that every subject choose the message that generates the planner's desired outcome, regardless of the messages and preferences of the other subjects. This property is highly desirable because it implies that subjects have a strong incentive to comply with the mechanism, even if there is a lack of common knowledge of rationality or common knowledge of beliefs.

Early on, a series of classic impossibility results showed that in most circumstances it is impossible to design a mechanism that satisfies efficiency, voluntary participation, and dominant strategy incentive compatibility (Hurwicz, 1972; Gibbard, 1973; Satterthwaite, 1975). The literature reacted to these impossibility results by investigating what could be achieved if one or more of the properties were relaxed. For example, several articles have investigated what can be achieved if the requirement of voluntary participation is foregone and the dominant strategy requirement is relaxed to a Bayesian (d'Aspermont and Gerard-Varet, 1979) or a Nash Equilibrium (Groves and Ledyard, 1977; Maskin, 1999). Others have investigated relaxing the requirement of efficiency (Vickrey, 1961; Clarke, 1972; Groves, 1973). Although many of these results have generated profound insights into the nature of institutions and incentives, the solutions that they have generated fall short of the ideal criteria that originally motivated the literature.

One promising strand of this literature, on which our work draws, is found in Cremer and McLean, 1985 and 1988, and in McAfee and Reny, 1992. Using Bayesian incentive compatibility they show that "introducing arbitrarily small amounts of correlation into the

joint distribution of private information among the players is enough to render private information valueless” (McAfee and Reny 1992, p. 395) and to allow the mechanism designer to fully extract expected rents. This is equivalent to achieving expected efficiency and voluntary participation. Their examples and applications are focused on situations with correlated information and Bayes incentive compatibility under common knowledge of beliefs and rationality. But, as we will show below, with the right information structure, their approach allows us to dispense with common knowledge assumptions and actually recover dominant strategy implementation.

A fundamental assumption behind the impossibility results, which has been maintained in the mechanism design literature, is that the only way the planner can gain information about the individual preferences is by eliciting them behaviorally through a cleverly constructed mechanism. Although this has been a valid assumption for the last 30 years, modern neurometric technologies are now making it possible to obtain direct noisy signals of subjects’ preferences. For example, in a recent article (Krajbich et al., 2009) we showed that experimentally induced valuations for public goods could be predicted with 60% accuracy using a combination of functional magnetic resonance imaging (fMRI) and machine learning techniques. The field of “mind reading” (also called neural decoding) uses biological signals to classify mental states. Neurometric tools for mind reading are rapidly advancing and the accuracy of the measurements is steadily increasing (Haxby et al., 2001; Cox and Savoy, 2003; Kamitani and Yong, 2005; Polyn et al., 2005; Norman et al., 2006; Haynes et al., 2007; O’Toole et al., 2007; Pessoa and Padmala, 2007; Serences

and Boynton, 2007; Kay et al., 2008). Thus, there is reason to believe that in the future it will be possible to design practical mechanisms that make use of such signals.

The availability of direct signals about subjects' values raises a fundamental question in mechanism design, which is the topic of this chapter: Is it possible to use such noisy signals to construct mechanisms that satisfy efficiency, voluntary participation, and dominant strategy incentive compatibility? We refer to such mechanisms as neurometrically informed mechanisms (NIMs). In Krajbich et al. (2009) we provided an initial proof of concept by showing theoretically and experimentally that it is possible to create a NIM for a very simple public goods problem with two player types. In this chapter we address the larger question by developing a general theory of NIMs and testing the results experimentally.

Using an approach pioneered by Cremer and McLean (1985, 1988) and developed further by McAfee and Reny (1992), we are able to show that with an even mildly informative signaling technology *any* feasible allocation rule can be implemented using a mechanism that satisfies voluntary participation and dominant strategy incentive compatibility. This result shows that the availability of neurometrically informed signals has a profound impact on the mechanism design problem. Importantly, we also show that the “recipe” for constructing NIMs that satisfy the desired properties is relatively simple and transparent.



As in most of the mechanism design literature, the fundamental result is derived under the assumption of risk- and loss-neutral subjects. We test the robustness of our results to the introduction of risk- and loss-aversion in two public good experiments. The results show that the augmented mechanisms are robust to this complication for the degrees of loss- and risk-aversion observed in most of our sample.

The chapter is organized as follows. In Section II we review the basic impossibility results from classical mechanism design theory and develop a possibility result for neurometrically informed mechanisms. In Section III we present the results of our two experiments. In Section IV we discuss the scope and limitations of our results.

## **II. Theory**

This section begins with a review of some fundamental impossibility results from classical mechanism design theory. We then derive the basic theory of neurometrically informed mechanisms.

### **II.1. Review of classical mechanism design theory**

Consider environments with  $N$  individuals indexed by  $i=1, \dots, N$ . Individuals have quasi-linear preferences denoted  $u_i(x, v_i) - t_i$ , where  $v_i \in V^i$  denotes player  $i$ 's type,  $x \in X \subset \mathfrak{R}^L$  is the allocation of resources for the group, and  $t_i \in \mathfrak{R}$  denotes a payment from player  $i$ . We assume that the set of types for each individual,  $V^i$ , is a finite set.  $X$  denotes the set of feasible allocations.

The set of environments for the mechanism design problem is given by

$\{X, u_1, \dots, u_N, V^1, \dots, V^N\}$ . An allocation  $(x, t)$  is feasible if  $x \in X$  and  $\sum_{i=1}^N t_i \geq 0$ . An

allocation is efficient in a quasi-linear environment, given the individuals' types

$v = (v_1, \dots, v_N)$ , if and only if it maximizes  $\sum_{i=1}^N u_i(x, v_i)$  and  $\sum_{i=1}^N t_i = 0$ . Note that efficient

allocations satisfy the Pareto property, so there is no other feasible allocation that can make everyone better off.

A mechanism is given by a message space  $M = M_1 \times \dots \times M_N$  and an outcome function  $g$ .

Each individual  $i$  reports a message  $m_i \in M_i$ . The vector of messages then determines an

outcome according to the function  $g(m) = [x(m), t(m)] \in X \times \mathfrak{R}^N$ .

The mechanism design problem can now be described as follows. The social planner

would like to implement an allocation rule  $a(v) = (x(v), t^1(v), \dots, t^N(v))$  for every vector of

preferences  $v$ . However, he cannot do so directly because he does not know the individual types  $v_i$ . Instead he asks the subjects to play a mechanism  $(M, g)$ . Let  $m^*(v)$  denote the messages selected by the individuals in equilibrium. If the mechanism has the property that, for all  $v$ ,  $g(m^*(v)) = a(v)$  and  $m^*(v) = [m_1^*(v_1), \dots, m_N^*(v_N)]$  where  $m_i^*(v_i)$  is a dominant strategy for  $i$ , then we say that the mechanism  $(M, g)$  implements the allocation rule  $a$  in dominant strategies. Such a mechanism solves the social planner's problem because, regardless of the actual preferences or beliefs held by the individuals, the mechanism induces the desired allocation.

A class of mechanisms of particular interest are direct revelation mechanisms in which  $M^i = V^i$  for all  $i$  and for which  $m^*(v) = v$  is a dominant strategy equilibrium for all  $v$ . Such mechanisms are called incentive compatible direct revelation mechanisms. A fundamental result in mechanism design is the Revelation Principle (A. Gibbard, 1973, and R. Myerson, 1981), which states that if an allocation rule can be implemented in dominant strategies using a mechanism  $(M, g)$  then it can also be implemented using some incentive compatible direct revelation mechanism  $(V, g')$ . Because of the Revelation Principle we can, without loss of generality, focus only on incentive compatible direct revelation mechanisms in the rest of the chapter.

The mechanism design literature has focused on direct revelation mechanisms satisfying three basic properties: efficiency, incentive compatibility<sup>1</sup> and voluntary participation. A direct mechanism  $(V, g)$  is efficient if and only if the allocation  $(x(v), t(v))$  is efficient for all  $v$ . The mechanism is incentive compatible if and only if

$$u_i(x(v_i, v_{-i}), v_i) - t_i(v_i, v_{-i}) \geq u_i(x(v'_i, v_{-i}), v_i) - t_i(v'_i, v_{-i})$$

for all  $i$ ,  $v$ , and  $v'_i$ . This condition ensures that each individual's utility is highest by reporting  $m_i = v_i$  regardless of the types and reports of the other players. Finally, the mechanism satisfies voluntary participation if and only if

$$u_i(x(v_i, v_{-i}), v_i) - t_i(v_i, v_{-i}) \geq 0$$

for all  $i$  and  $v$ . This condition ensures that each individual receives a non-negative payoff from truthful reporting, regardless of the types and reports by the other players.

We are now in position to state the

---

<sup>1</sup> From this point on we will use the phrases "dominant strategy incentive compatibility" and "incentive compatibility" interchangeably.

**Fundamental Impossibility Theorem of Mechanism Design:** *If  $V$  is rich enough, then there is no mechanism that is Efficient and Incentive Compatible and satisfies Voluntary Participation.*

A number of variations of this theorem have been provided. Hurwicz (1972) provides one of the first results for exchange environments. Gibbard (1973) and Satterthwaite (1975) show that the result holds when  $V$  includes all preferences over at least three alternatives. Green and Laffont (1979) consider quasi-linear preferences but include all possible preferences for the non-linear part. Walker (1980) shows that that the result still holds for quasi-linear preferences for which the non-linear part is concave. Hurwicz and Walker (1990) provide the most general version of the result for quasi-linear environments.

With this impossibility theorem, the original search for mechanisms that satisfy dominant strategy incentive compatibility, efficiency, and voluntary participation seems doomed to failure. In the next section we show that the goal can be rescued using neurometrically informed mechanisms.

## II.2. Neurometrically Informed Mechanisms

In this section we consider what happens to the mechanism design problem when the social planner has access to a technology that provides noisy but informative signals about each individual's type. See the Introduction and Discussion sections for a discussion of why such signals are technologically feasible, and not mere theoretical curiosities.

The planner is able to observe a signal  $s \in S$  for each individual *after* they have announced their type.  $S$  is assumed to be a finite set.<sup>2</sup> The signals are distributed according to a density function conditional on the true types. A signal technology for an environment is thus given by a mapping  $T:V \rightarrow \Delta(S)$ , where  $\Delta(S)$  is the set of probability densities on  $S$ . For the signaling technology to be useful, the signals have to be sufficiently informative. We assume that  $T$  is *I-I* from  $V$  to  $T(V)$ . Note that the signal likelihood function does not depend on the subject's messages. Also, note that the subject neither knows what signal will be observed by the planner, nor can the subject manipulate the signal in any way. All he knows is his density function  $T(v_i)$ .

The availability of the signals allows us to augment the mechanism by introducing an additional tax function that depends on both the signals and the subjects' reports. In

---

<sup>2</sup> This is done for ease of exposition. The results in this paper also hold for larger signal spaces.

particular, for any direct mechanism  $(V, g)$  we can define an *augmented mechanism*  $(V, g, w)$ , in which the new outcome function is given by

$$g^a(s, m) = (x(m), t(m) + w(s, m))$$

where  $w(s, m)$  denotes the augmented tax. Efficiency requires that  $\sum_i w_i(s, m) = 0$ .

We can now define the key properties of interest for neurometrically informed mechanisms.

The mechanism  $(V, g, w)$  is dominant strategy incentive compatible if and only if for all  $i$ ,  $v_i$ , and  $m$ , we have that

$$u(x(v_i, m_{-i}), v_i) - t(v_i, m_{-i}) - E(w_i(s, v_i, m_{-i}) | v_i) \geq u(x(m_i, m_{-i}), v_i) - t(m_i, m_{-i}) - E(w_i(s, m_i, m_{-i}) | v_i)$$

where  $E$  denotes the expectation operator.

The mechanism  $(V, g, w)$  satisfies voluntary participation if and only if, for all  $i$ ,  $v_i$ , and  $m_i$ ,

$$u(x(v_i, m_{-i}), v_i) - t(v_i, m_{-i}) - E(w_i(s, v_i, m_{-i}) | v_i) \geq 0.$$

Note that for neurometrically informed mechanisms, incentive compatibility relies on the timing of the signal technology. At the time  $i$  is choosing her message, she does not know what signal  $s$  will be observed by the planner and thus must base her choice on the likelihood function of the signaling technology. Incentive compatibility states that all individuals are willing to report their true types regardless of the reports by the other subjects, given the expected payoffs induced by the augmentations. Under our incentive compatibility condition, truth-telling is the best response even if others misreport. We do not require common knowledge of either rationality or beliefs. We get dominant strategy behavior instead of ex-post incentive compatibility because the signals are independently distributed.

Voluntary participation also relies on the timing of the signal technology. At the time  $i$  is choosing her participation decision, she does not know what signals  $s$  will be observed by the planner and thus must base her choice on the likelihood function of the signaling technology. Voluntary participation states that all truth-telling individuals are willing to participate regardless of the reports by the other subjects, given the expected payoffs induced by the augmentations. We are normalizing  $u$  so that the value of the individual's outside option is  $0$ .

We now turn to describing how to find augmentations  $w$  to a mechanism  $(V, g)$  such that the neurometrically informed augmented mechanism  $(V, g, w)$  will be incentive compatible and satisfy voluntary participation. With the availability of a signaling technology,



having an agent reveal their type,  $v$ , is equivalent to having them reveal the probability distribution  $T(v)$ . There is a well-known way to induce revelation of a probability density using proper scoring rules (G.W. Brier, 1950). For a useful review see T. Gneiting and A.E. Raftery, 2007.

A scoring rule is a real valued function  $h : \Delta(S) \times S \rightarrow \Re$  that assigns a score, a real number, to an announced probability distribution,  $p \in \Delta(S)$  and a realization  $s$ . Let

$H(q,p) = \sum_{s \in S} h(q,s)p_s$  be the expected value of the score if  $q$  is announced when the true

density is  $p$ . If  $H(p,p) > H(q,p)$  for all  $q \neq p$  then we call  $h$  a proper scoring rule. There are many proper scoring rules. One well-known one is the logarithmic scoring rule

$$h(p,s) = \ln[p_s], \quad H(q,p) = \sum p_s \ln(q_s).$$

To use a scoring rule to induce agents to reveal their true types, we must make the incentives to reveal at least as great as the gains from non-revelation. In the original mechanism  $(V,g)$ , there is a maximum benefit to misreporting. This can be as large as

$$\alpha(v_i, m_i) = \max_{m_{-i}} [u(x(m_i, m_{-i}), v_i) - t_i(m_i, m_{-i})] - [u(x(v_i, m_{-i}), v_i) - t_i(v_i, m_{-i})]$$

A particularly useful class of neurometrically informed mechanisms are those constructed using person-by-person augmentation. In the first step we compute each individual's provisional tax  $r(s_i, m_i)$ , which only depends on her own message and signal. Second, to

ensure efficiency, the surplus raised  $\sum_{i=1}^N r(s_i, m_i)$  is redistributed back equally to all the individuals. This generates an augmented tax function

$$w_i(s, m) = \left( \frac{N-1}{N} \right) \left[ r(s_i, m_i) - \left( \frac{1}{N-1} \right) \left( \sum_{j \neq i} r(s_j, m_j) \right) \right]$$

Under person-by-person augmentation, incentive compatibility is equivalent to

$$\left( \frac{N-1}{N} \right) [R(v_i, m_i) - R(v_i, v_i)] \geq \alpha(v_i, m_i), \forall v_i, m_i$$

where  $R(v_i, m_i) = E(r(s, m_i) | v_i) = \sum_s r(s, m_i) T_s(v_i)$

We can begin the augmentation process with any proper scoring rule  $h(p, s)$  and consider  $h(T(m), s)$ . Then  $H(T(m), T(v)) < H(T(v), T(v))$  for all  $m, v$ . Now increase the gain on the proper scoring rule and let

$$\lambda^* = \max_{v^i \neq m^i} \left( \frac{N}{N-1} \right) \left[ \frac{\alpha(v_i, m_i)}{H(T(v_i), T(v_i)) - H(T(m_i), T(v_i))} \right].$$

The maximum exists because  $V$  is finite.

If we let  $r(m_i, s) = -\lambda^* h(T(m_i), s)$ , then it is easy to see that the mechanism  $(V, g, w)$  will be incentive compatible.

While any proper scoring rule can lead to an incentive compatible augmented mechanism, it is not necessarily true that the mechanism will then also satisfy voluntary participation. If  $(V, g)$  satisfies voluntary participation then for our person-by-person augmented mechanism to also satisfy voluntary participation, we need

$$\left[ R(v_i, v_i) - \left( \frac{1}{N} \right) \sum_{j=1}^N R(m_j, v_j) \right] \leq 0, \forall v_i \in V^i, m_{-j}.$$

A sufficient condition for this is  $H(v, v) = C$ , for all  $v$ , because from incentive compatibility  $R(m_j, v_j) = -\lambda^* H(m_j, v_j) > -\lambda^* H(v_j, v_j) = R(v_j, v_j), \forall m_j \neq v_j$ . To ensure that we can find a proper scoring rule  $h(p, s)$  such that  $H(p, p) = C$ , we need the signal technology to satisfy a well-known condition from Cramer and McLean (1985).

**CM Condition:** The signal technology satisfies  $\forall v' \in V, T(v) \notin \text{co}\{T(v)\}_{v \in V, v \neq v'}$  where  $\text{co}A$  is the convex hull of  $A$ .

As described in McAfee and Reny (1992), under the CM condition, a separating hyperplane argument gives us the result that

$$\forall v \in V, \exists z(v) \in \mathbb{R}^S \ni T(v)z(v) = 0 > T(v')z(v), \forall v' \neq v$$

Thus  $h(T(v),s) = z_s(v)$  is a proper scoring rule with the additional property that  $H(T(v),T(v)) = 0$  for all  $v$ .<sup>3</sup>

We summarize the possibility results for neurometrically informed mechanisms in

**A Possibility Theorem for Neurometrically Informed Mechanisms.** *Given a 1-1 signal technology  $T:V \rightarrow \Delta(S)$  and quasi-linear environments with finite type spaces:*

- a) *Given any direct revelation mechanism  $(V,g)$ , there is a person-by-person augmented mechanism  $(V,g,w)$  that is incentive compatible and yields the same expected outcome,*
- b) *Given any direct revelation mechanism  $(V,g)$  satisfying voluntary participation, if the signal technology  $T:V \rightarrow S$  satisfies the CM condition, there is a person-by-person augmented mechanism  $(V,g,w)$  that is incentive compatible, satisfies voluntary participation, and yields the same expected outcome.*

These results show that with access to a sufficiently informative signal technology any efficient allocation rule can be implemented in dominant strategies using a tax function that satisfies voluntary participation. This stands in sharp contrast with one of the most fundamental theorems of mechanism design that shows that, in the absence of such

---

<sup>3</sup> It is also true that there is a proper scoring rule,  $z(v)$ , on  $T(V)$  such that  $T(v)z(v) = C$  for all  $v$  iff the CM condition is true on  $T(V)$ .

signals, there is no standard mechanism that is efficient, incentive compatible, and satisfies voluntary participation.

### II.3 Some Additional Observations

Our results are even stronger than we have stated so far. The possibility theorem for neurometrically informed mechanisms also holds if preferences are interdependent, a situation in which standard dominant strategy mechanisms may not even exist. If utility functions depend on everyone's types, i.e.,  $u_i = u_i(x, v_1, \dots, v_N)$ , with neurometrically informed signals satisfying the Cremer-McLean condition, one can construct mechanisms that implement efficient allocations, satisfy voluntary participation, and are incentive compatible in dominant strategies. One does this in the choice of  $\lambda^*$  by also maximizing over all  $v_{-i}$ .

In addition to providing an existence proof, our results are constructive in the sense of providing a simple "recipe" for how to construct the necessary mechanisms. As the proof of the lemma emphasizes, all that is required to get incentive compatibility is the addition of an augmented tax function that sets taxes proportional to any pre-existing scoring rule from the literature. In fact, it is possible to show that any person-by-person augmentation tax function that provides dominant strategy implementation will be a proper scoring rule.

It is a little harder to get voluntary participation but there are some special cases where even that is simple. One example arises when the supports of  $T(v)$ ,  $\bar{S}(v) = \{s \in S \mid T_s(v) > 0\}$ , are different for all  $v$ . When  $v \neq v' \Rightarrow \bar{S}(v) \neq \bar{S}(v')$ , we can use the simplest possible scoring rule, which assigns a score of 0 to any signal in the support of  $T(v)$ , and a score of -1 to any other signal. This is a proper scoring rule and  $H(T(v), T(v)) = 0$  for all  $v$ . A second case occurs if the signal distribution maintains its shape and simply shifts around as  $v$  changes. If  $T_s(v) = T_{s+\mu}(v + \mu)$  for all  $\mu = v' - v, \forall v, v' \in V$ , then for scoring rules such that  $h(T(v), s) = h(T(v+\mu), s+\mu)$  it will be true that  $H(T(v), T(v)) = H(T(v+\mu), T(v+\mu))$ . One such scoring rule is the logarithmic rule,  $h(T(v), s) = \ln T_s(v)$ .

Notice several useful properties of the class of neurometrically informed mechanisms. First, all of the possibility results can be obtained from person-by-person augmentation, which makes the computation and description of the mechanisms relatively easy. Second, in the dominant strategy equilibrium where  $m = v$ , the expected utility to any individual is exactly the utility they get under honest reporting in the un-augmented mechanism. Thus, in the dominant strategy equilibrium the augmented taxes do not on average cause any wealth redistribution. Third, the mechanisms are balanced even *after* the signal. That is, they do not take any resources from the system, they only redistribute. However, although they satisfy voluntary participation *before* the signals are observed, they may not *after* the signals. We confront this experimentally next in Section III.

Finally, we have provided a complete proof and characterization for the case when the set of possible types  $V$  is finite. But it is easy to see that, given appropriate continuity in the allocation to be implemented and in the utility functions, we could generate an approximation result similar to that of McAfee and Reny 1992. Look at the definition of  $\lambda^*$  in the proof of the lemma. If  $V$  were a continuum, then the denominator would be arbitrarily close to zero for some values of  $v_i'$  near  $v_i$ , and the max would not exist. Scoring rules, necessary for dominant strategies, which have a max or min at zero, (necessary for voluntary participation) will flatten out as  $v_i'$  nears  $v_i$  but the gain from misrepresenting does not. Thus there is a fundamental impossibility when  $V$  is a continuum. But if we relax the dominant strategy condition to *epsilon-dominant strategy*, or voluntary participation to *epsilon-voluntary participation*, then we can get equivalent results for  $V$  that are not finite.

### **III. Experiments: Neurometrically Informed Public Goods.**

In this section we present the results of two experiments designed to test the results derived above. Testing the results with real subjects is important for two reasons.

First, the theory assumes that subjects are risk-neutral, whereas experimental tests show that most subjects exhibit risk- and/or loss-aversion (Kahneman and Tversky, 1979).

This is a potential problem for the theory because, since the signal technology is stochastic, subjects face a distribution of taxes, even if they report truthfully. As a result, voluntary participation and/or incentive compatibility in terms of expected utility might be violated for some individuals.

Second, although the neurometrically informed mechanisms that we have proposed are relatively simple, there is a concern about whether subjects will understand that it is in their best interest to report truthfully. This concern is justified, for example, by previous results showing that some subjects do not bid truthfully in second price auctions (Coppinger, Smith, and Titus, 1980).

### **III.1. Experiment 1**

The first experiment is designed to test whether the voluntary participation and incentive compatibility properties of a variant of the neurometrically informed mechanism proposed in Krajbich et al. (2009) are robust to the introduction of typical levels of risk- and loss-aversion.

***Environment.*** Groups of five subjects need to decide whether or not to produce a public good that has a cost of \$25. The preferences of subject  $i$  are given by  $v^i g - y^i$ , where  $g = 0, 1$



is the level of the public good,  $v^i = \{\$1, \$9\}$  is the value for the public good, and  $y^i$  is the net tax paid by the individual.

We assume that the social planner has access to a signal technology with the following properties:  $p(s = 1 | v = 1) = p(s = 9 | v = 9) = 0.8$  and  $p(s = 1 | v = 9) = p(s = 9 | v = 1) = 0.2$ .

In other words, the signal equals the true value with 80% probability and signals the other value with 20% probability.

**Neurometrically informed mechanism.** We consider a simple direct revelation mechanism in which subjects simultaneously report their values  $m^i = \{\$1, \$9\}$  and then the planner receives a signal for each of them  $s^i = \{\$1, \$9\}$ . The public good is ruled to induce efficiency over the reported values. Therefore, the public good is built whenever three or more subjects report a \$9 value. Each subject pays a gross tax  $t^i(m^i) + r^i(s^i, m^i)$  as given by the following table:

		message	
		\$1	\$9
signal	\$1	receive \$3.67	pay \$9
	\$9	pay \$14.67	pay \$9

If the amount raised from the gross taxes does not equal the cost of the public good, then the difference is evenly collected from (deficit) or returned to (surplus) the five subjects.

In Krajbich et al. (2009) we showed that the resulting mechanism satisfies efficiency,

voluntary participation, and dominant strategy incentive compatibility for risk-neutral subjects.

**Subjects.**  $N=50$  subjects participated in the study. They were recruited from Caltech's student population. In addition to their experimental earnings, subjects were paid \$20 for their participation.

**Experimental task.** In each experimental session, subjects played two rounds of the experimental task described below, once with a \$1 value for the public good, and once with a \$9 value (both presented simultaneously). In each round subjects had to make two sequential decisions after observing their private value. First, they had to cast a binary Yes-No vote for whether they wanted the neurometrically informed mechanism to be played. These votes mattered because with probability  $0.2 * I_{\#No \text{ votes}}$  the mechanism was not played. Second, they had to report a value  $m^i$ , which was binding only if the mechanism was played.

Note the logic of the experiment. In the voting decision, it is a dominant strategy for the subject to vote Yes if and only if her expected utility from the neurometrically informed mechanism is positive. As a result, the vote decision allows us to determine for each subject whether the mechanism satisfies voluntary participation. Furthermore, since the

vote does not affect the rules of the mechanism conditional on it being played, we can elicit this measure without affecting the subsequent incentives.

There was no feedback between rounds. The experiment was carried out with pen and paper. After the decisions were collected from all the subjects in the experimental session, each subject's decision was randomly paired with four other subjects from the group, and they were paid for one randomly chosen group outcome. This ensured that the distribution of values were independent across rounds and subjects.

The experimental instructions (included in Appendix I) provide an in-depth description of the rules of the mechanism and the payments that they induced under different contingencies.

***Measuring risk-aversion.*** Immediately after the main experimental task, subjects were asked to play a simple gambling task designed to estimate their risk- and loss-aversion parameters. In particular, subjects made 40 different binary choices between lotteries (Sokol-Hessner et al., 2009). Their responses were used to estimate the parameters of the following prospect theoretic model (Kahneman and Tversky, 1979) using Bayesian estimation (Wang, Camerer, Filliba, 2010):

$$U(x) = \begin{cases} x^\rho & \text{if } x \geq 0 \\ -\lambda |x|^\rho & \text{if } x < 0 \end{cases}$$

where  $\rho$  is a measure of the degree of risk-aversion, and  $\lambda$  is a measure of the degree of loss-aversion. Note that risk-aversion is decreasing in  $\rho$ , but loss-aversion is increasing in  $\lambda$ .

**Results.** The aggregate results suggest that for most subjects the mechanism was incentive compatible: 88% of subjects reported truthfully when they had a value of \$9, and 98% reported truthfully when they had a value of \$1.

We next tested if there was a systematic relationship between the coefficients of risk- and loss-aversion and subjects' choices to misrepresent their values. A logit regression of an indicator function for truth-telling on the coefficients of risk- and loss-aversion for each subject estimated no effect for loss-aversion ( $p=0.27$ ) or for risk-aversion ( $p=0.42$ ).

The aggregate results also suggest that most of the subjects believed that the mechanism satisfies voluntary participation: 100% of the subjects voted Yes when they had a value of \$9 and 78% voted Yes when they had a value of \$1.

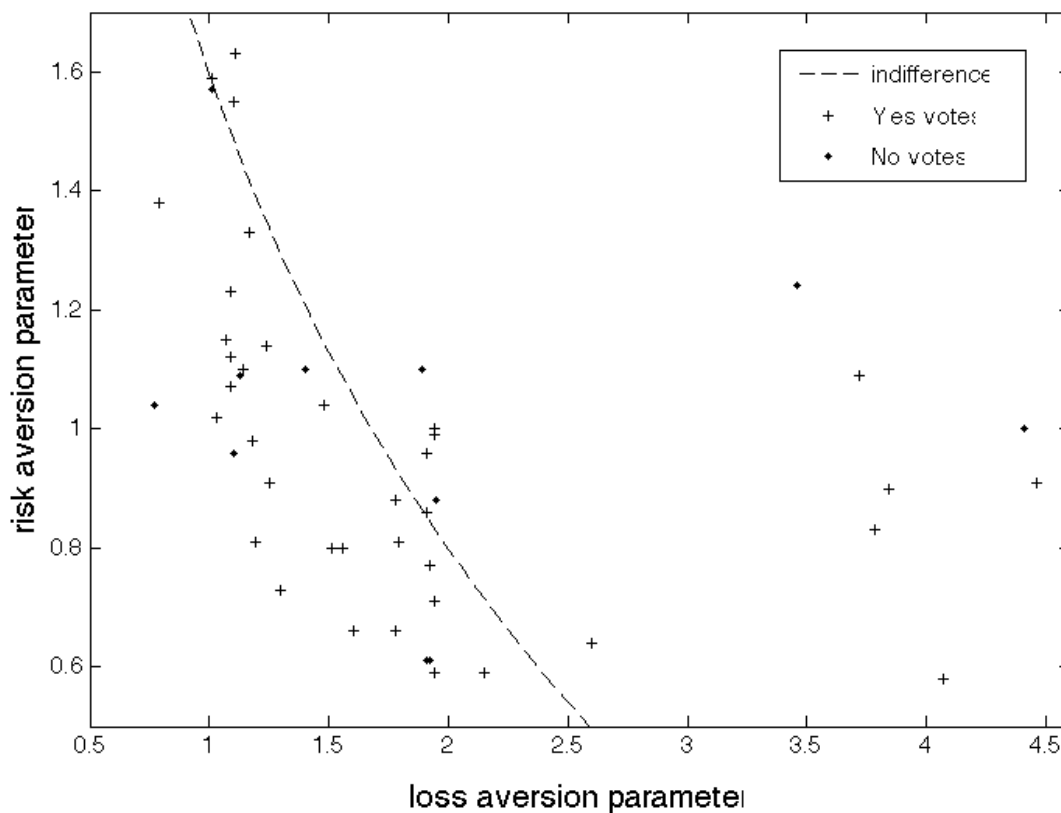
We next tested if there was a systematic relationship between the coefficients of risk- and loss-aversion and subjects' beliefs that the mechanism did not satisfy voluntary participation. A logit regression of an indicator function for No votes on the coefficients of risk- and loss-aversion for each subject estimated no effect for the coefficient of loss-aversion ( $p=0.5$ ) or for the coefficient of risk-aversion ( $p=0.55$ ).

Figure 1 provides an additional test of this relationship. The dashed line in the upper graph depicts the voluntary participation constraint, as a function of the coefficients of risk- and loss-aversion, for individuals with  $v_i=1$ . Note that the constraint is satisfied in the lower left and it is violated in the upper right. Each point in the graph represents a subject's parameters and her vote. As can be seen in the figure, a vast majority of the subjects fall to the left of or very near the indifference curve. Only 14% of subjects have sufficiently high loss-aversion where voluntary participation becomes an issue. Furthermore, only two out of those seven subjects actually vote No in the public goods game. Also, note that the voluntary participation calculation changes depending on the gross taxes paid by the other players. Therefore, here we have considered the "worst-case scenario" for the subject, where the expected sum of the other players' gross taxes is minimized, and therefore so is her budget-balancing refund. If voluntary participation is satisfied in this scenario, then it will be satisfied in all scenarios.

For the case where  $v_i=9$ , a truth-telling subject faces no uncertainty in his gross tax and earns a positive payoff in the worst-case scenario, and so voluntary participation will

always be satisfied. This is consistent with the experimental data. Together, these results allow us to conclude that violations of voluntary participation are associated with large degrees of loss-aversion, which are observed in a small fraction of the population.

**Figure 1.** Individual voluntary participation constraint as a function of the individuals risk- and loss-aversion parameters for the case of  $v_1 = 1$ . The dashed line depicts the voluntary participation constraint, as a function of the coefficients of risk- and loss-aversion. Note that the constraint is satisfied in the lower left and it is violated in the upper right. Each point in the graph represents a subject's parameters and her vote. Note that the constraint is calculated in the “worst-case scenario” for the subject, where the sum of the gross taxes paid by the other players is minimized. So if voluntary participation is satisfied in this case, it will be satisfied in every other scenario as well.



### III.2. Experiment 2

The second experiment is designed to test the performance of the neurometrically informed mechanism in more complex domains in which there are a large number of types.

**Environment.** Groups of five subjects need to decide how much of a public good to produce. The preferences of subject  $i$  are given by  $v_i \log(z) - y_i$  where  $v_i$  denotes the public good type,  $z$  is the total amount invested in the public good, and  $y_i$  is the net tax paid by the subject. We assume that every subject can have one of 20 possible types, so that  $V = \{1, 2, \dots, 20\}$ .

There is a signal technology with the following properties: if the subject's true type is  $v_i$  then the signal is uniformly distributed on  $[v_i - 10, v_i + 10]$ .

It is straightforward to show that efficiency requires producing a level of public goods

given by  $z(v) = \sum_{i=1}^n v_i$  and raising an equivalent amount of taxes given by  $\sum_{i=1}^n y_i(v) = z(v)$ .



An efficient allocation of particular interest is the one characterized by a Lindahl Equilibrium (Lindahl, 1958). In this allocation,  $i$ 's contribution is given by her marginal benefit for the public good times the level of the public good (both calculated at the efficient level  $z(v)$ ), which in this environment equals  $v_i$ .

**Standard Lindahl Mechanism (SLM).** A Lindahl mechanism is a direct revelation mechanism  $(V, g)$  that implements the optimal level of the public good given the reports (i.e.,  $z(m) = \sum_{i=1}^n m_i$ ) and funds it using the Lindahl taxes (i.e.,  $y_i(m) = m_i$ ). For these environments, the SLM is identical to the standard Voluntary Contributions Mechanism. (See Ledyard, 1995.) It is straightforward to show that the Standard Lindahl Mechanism is efficient and satisfies voluntary participation, but is not dominant strategy incentive compatible.

**Augmented Lindahl Mechanism (ALM).** We construct the Augmented Lindahl Mechanisms by applying the person-by-person augmentation methods described in Section III. In particular, we use the simple proper scoring rule  $h_i(s_i, m_i) = -(s_i - m_i)^2$  and  $\lambda^* = \left( \frac{N}{N-1} \right)$ . This implies that gross augmented taxes are given by

$r_i(s_i, m_i) = \frac{N}{N-1} (s_i - m_i)^2$  and the net augmented tax function is given by

$$w_i(s,m) = (s_i - m_i)^2 - \frac{1}{N-1} \sum_{j \neq i} (s_j - m_j)^2.$$

The net tax paid by the subjects is given by

$$t_i(s,m) = m_i + w_i(s,m).$$

For the signal technology here,  $H(v,m) = -\left(\frac{1}{20}\right) \int_{v-10}^{v+10} (s-m)^2 ds = -(v-m)^2 - \frac{100}{3}$ .

Thus  $H(v,m) < H(v,v)$  for all  $m \neq v$  and  $H(v,v) = -(100/3)$ . Therefore, we know that for risk- and loss-aversion neutral subjects it satisfies incentive compatibility and voluntary participation.

**Subjects.**  $N = 30$  Caltech undergraduates participated in the experiment. In addition to their earnings during the task, they paid a \$15 fee.

**Experimental task.** In each experimental session 10 subjects participated in 40 rounds of decision making, 20 with the SLM, and 20 with the ALM. 20 subjects played the rounds with the SLM before the ALM, the other 10 subjects played the opposite order.

In each round the private value was randomly and independently drawn for each subject from a uniform distribution on  $V$ . For each mechanism subjects remained in the same group of five, but between mechanisms subjects were randomly rematched. To allow for learning, at the end of each round subjects were told their earnings for that round as well as the group's total contribution to the public good (given by  $\sum_{i=1}^5 m_i$ ). In the ALM, subjects were also told the value of their signal for that round and the components of their total tax  $t_i$ .

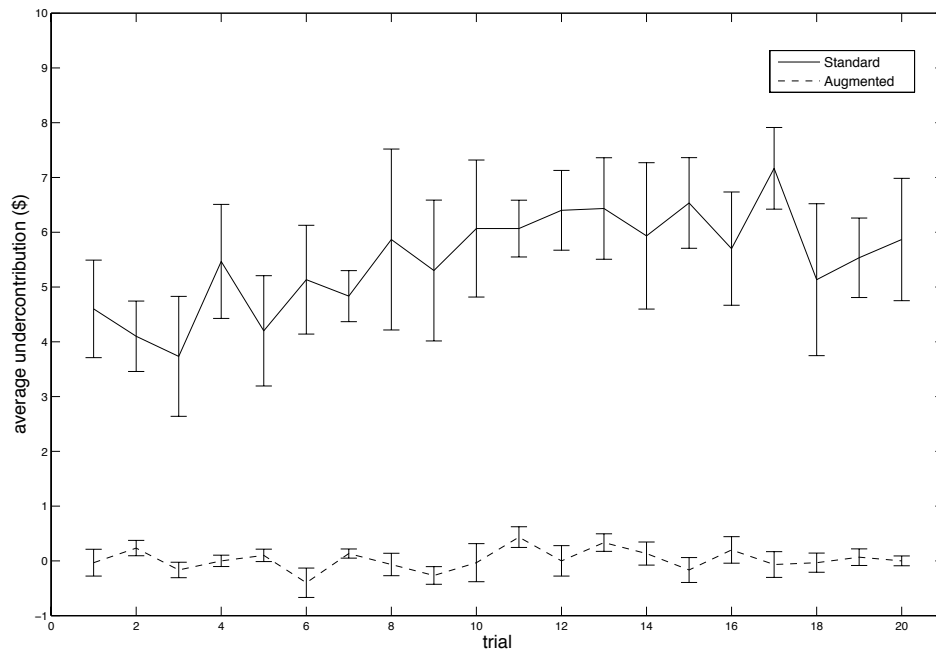
The experimental instructions (included in Appendix II) provide an in-depth description of the rules of both mechanisms as well as the payments that they induce under different contingencies. An important feature of the instructions is that we explicitly explain to the subjects that truth-telling maximizes their expected payoffs regardless of the decisions made by the other subjects. It is important to emphasize that while this aspect of the instructions is not an explicit requirement of the theory (which assumes that subjects know this), making sure that subjects fully understand key aspects of the distribution of payoffs is an integral part of applied mechanism design.

**Results.** In Figure 2 we report the deviations between truthful and actual reporting (given by  $v_i - m_i$ ), as a function of treatment and experimental round. Whereas there was no under-reporting in the ALM condition (mean=0.02, se=0.05, p=0.73), there was significant under-reporting in the SLM (mean=5.50, se=0.73, two-sided t-test p=0.0007).

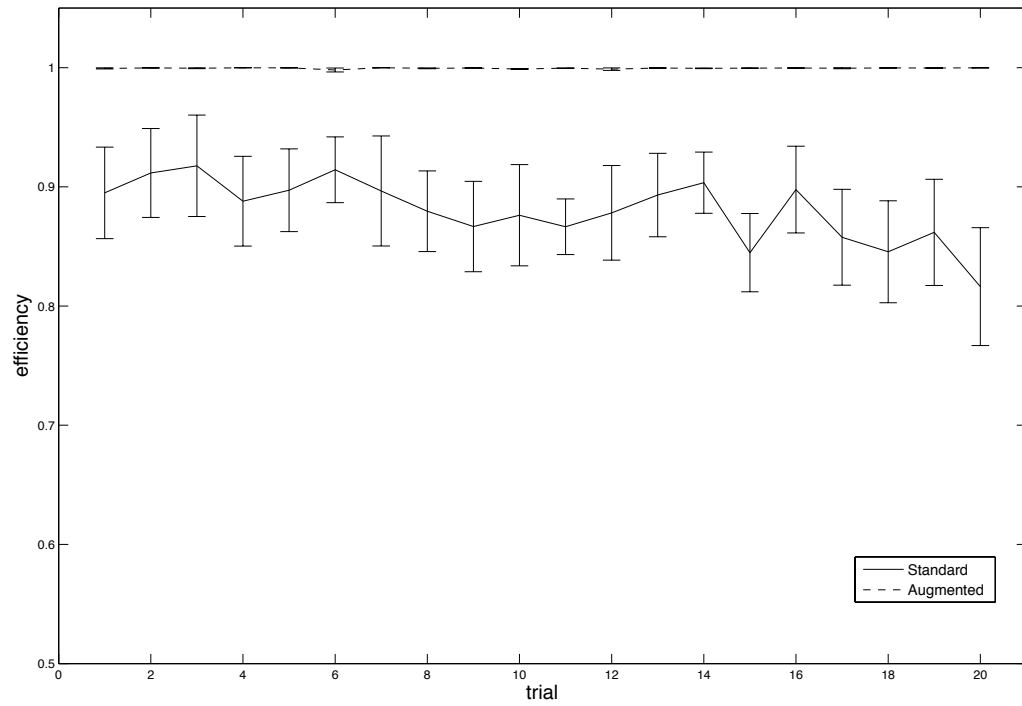
Furthermore, a mixed effects regression of the size of the deviations on round number for the ALM case reveals no learning effect ( $p=0.59$ ), but a similar regression for the SLM shows that under-reporting increases with round ( $\beta=0.10$ ,  $p=0.0005$ ). These results clearly illustrate the power of the neurometrically informed mechanisms: whereas the ALM elicits near-perfect truth-telling by the subjects, there is substantial free-riding in the SLM and it gets worse over time. The SLM results are consistent with standard linear Voluntary Contribution Mechanism results (See Ledyard 1995.)

In Figure 3 we display the efficiency of the allocations induced by both mechanisms. The figure plots efficiency as a function of round for each case, where efficiency is defined as

$$Efficiency = \frac{\text{actual total group payoff}}{\text{optimal total group payoff}}$$

**Figure 2.** Under-reporting in Experiment 2 by mechanism and trial.

**Figure 3.** Average efficiency in Experiment 2 by mechanism and trial.



Average efficiency was 99.95% (se=0.02%) on the ALM and 88% on the SLM (se=3%). Thus, the ALM generated significantly larger efficiencies (p=0.0097 two-sided t-test). Furthermore, a mixed effects regression of efficiency on round number for the ALM case revealed no time trends (p=0.32), whereas a similar regression for the SLM showed that efficiency decreased with time (beta= -0.0032 , p<0.0007), which translates to a drop of efficiency of 6.4% over 20 trials. These results show that whereas the ALM institution generated nearly perfect efficiency in every round, the SLM generates substantial inefficiencies that worsen over time.

Because we draw new values each round, it may not be clear from Figure 3 whether the efficiencies for the SLM are due to high contributions or high values. In Figure 4, we compare the efficiencies from the actual reports and to those that would occur in the Nash Equilibrium for the SLM and ALM. That is, we plot

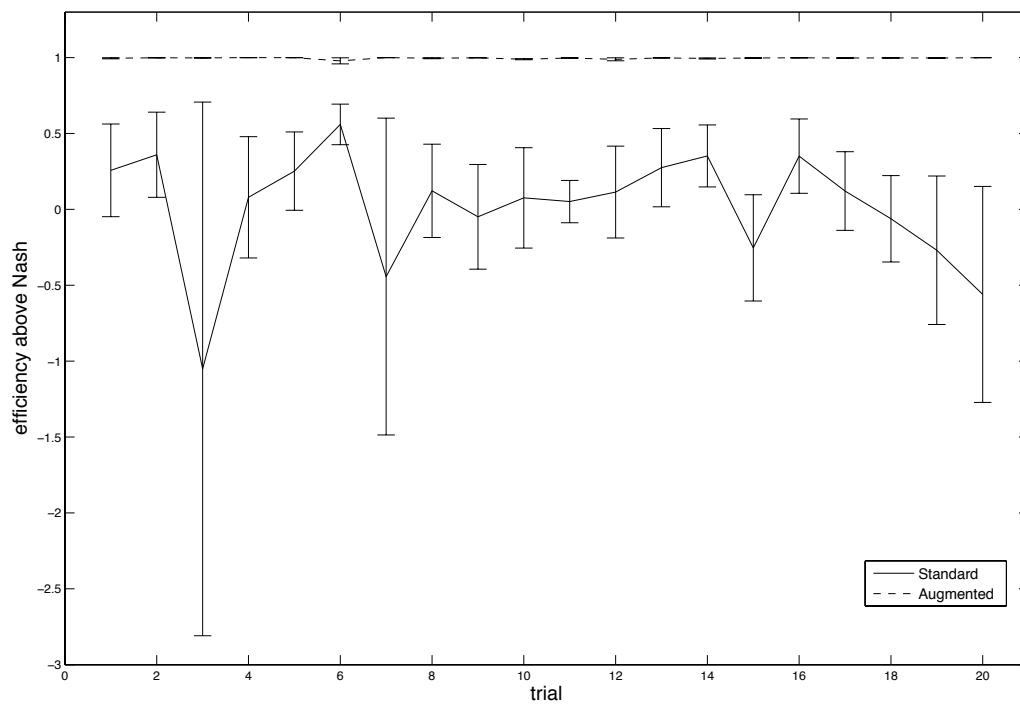
$$\frac{\text{actual group payoff} - \text{Nash equilibrium payoff}}{\text{optimal group payoff} - \text{Nash equilibrium payoff}}$$

It is important to realize that since we redraw values each round and since information about values is private, there is no game-theoretic reason why subjects should play the Nash Equilibrium, which usually has the highest valued type reporting their true value and all others reporting  $m = 1$ . Nevertheless, from the data displayed in Figure 2 it is not possible to reject the hypothesis that, on average, the outcomes are the Nash Equilibrium

outcomes ( $p < 0.9$ ). Further, unlike standard linear Voluntary Contribution Mechanism results, we do not see initial contributions that are higher than the Nash Equilibrium that then decline over time ( $p < 0.7$ ). Here, contributions are near to the Nash Equilibrium from beginning to end. Consistent with the observations from Figure 2, the efficiencies of the ALM are at the optimum. The Augmented Lindahl Mechanism provides a significant improvement over the Standard Lindahl Mechanism.



**Figure 4:** Comparing Nash Equilibrium to actual choices in Experiment 2 by mechanism and trial.



#### **IV. Final Remarks**

The theoretical results in this chapter show that the availability of even mildly neurometrically informative direct signals about the individuals' preferences has a profound impact on the mechanism design problem: they make it possible to design mechanisms that implement any desired allocation and satisfy voluntary participation and dominant strategy incentive compatibility. This stands in sharp contrast with the classic impossibility results showing that without such signals it is generally impossible to design such mechanisms.

As in most of the mechanism design literature, the theoretical results were derived under the assumption that individuals are risk- and loss-averse. To address the robustness of the results to relaxations of this assumption we investigated the properties of neurometrically informed mechanisms in two public goods domains. The results show that voluntary participation and incentive compatibility are easily accommodated for the range of loss- and risk-aversion parameters observed in most of our subject population.

Several aspects of our results deserve more discussion. First, how informative does the signal technology have to be for our general possibility result to hold? The key assumption, namely that the signal technology satisfy the CM condition, is quite weak. For example, in the two-type case it is satisfied whenever the likelihood function over

signals is different for different values of the true preferences, which can be satisfied even if the probability of the high signal given that the type is high is barely above chance. One caveat, though, is that the neurometrically informed mechanisms generate payoff variability that depends on the realization of the signal. Furthermore, the lower the quality of the signal, the larger the amount of variability that will be present in the payoffs, which could be a problem if subjects exhibit sufficient amounts of loss-aversion. Fortunately, however, current trends in neurometric technologies suggest that high-accuracy signals might be available in the near future (Haxby et al., 2001; Cox and Savoy, 2003; Kamitani and Yong, 2005; Polyn et al., 2005; Norman et al., 2006; Haynes et al., 2007; O'Toole et al., 2007; Pessoa and Padmala, 2007; Serences and Boynton, 2007; Kay et al., 2008).

Second, what are likely sources of signals in future applications? The findings in Krajbich et al. (2009), as well as those in the references in the previous paragraph, suggest that fMRI could be a good source of high-quality signals. However, such measurement technologies remain expensive, and thus it will be important to explore other less expensive sources of signals, such as electroencephalography (EEG), pupil-dilation, skin-conductance, and facial electromyography (EMG) (Tassinari and Cacioppo, 1992; Lang et al., 1993; Aboyoun and Dabbs, 1998; Bradley, 2000; Dimberg et al., 2000; Partala et al., 2000; Bradley et al., 2008). Preliminary work in our laboratory suggests that these signals might be able to provide sufficiently informative signals, at least in simple contexts. It is also important to emphasize that our results also apply to non-neurometric signals. For example, the results also apply to a situation in which the

planner has sufficiently informative priors about individual subjects' preferences based on previous behavioral or demographic data. However, the caveat with such signals is that individuals cannot know the values of the signals or be able to manipulate them, so non-neurometric signals would have to be collected covertly.

Third, what if it is only possible (perhaps due to costs) to obtain signals from a subset of the members in the group? This problem can be addressed using "random augmentation". This works as follows. Subjects make their choices for the neurometrically informed mechanism without knowing if a signal will be available for them. Afterwards, the planner randomly selects a subset of the group and obtains signals only for them. The augmented taxes can be redefined so that everybody's incentives at the initial phase are the same as in the case of full signal monitoring. For example, suppose that the planner selects each individual for signal extraction with a constant probability  $q$ . It is easy to see that an augmented tax given by

$$\widehat{w}^i(s,m) = \frac{1}{q} w^i(s,m)$$

induces the same incentives as in the case of full monitoring (where  $w^i(s,m)$  are the augmented taxes for the  $q=1$  case). This idea is reminiscent of the literature on auditing, where a principal induces an agent to truthfully report by probabilistically auditing them and charging a high fee if misrepresentation is found. See for example, Border and Sobel (1987) and Baron and Besanko (1984).

We conclude on an optimistic note. The rapid and likely rise of neurometric technologies has the potential to make it feasible to design much better mechanisms and institutions in a large number of applications. Importantly, as demonstrated here and in Krajbich et al. (2009), the development of such institutions will require a careful combination of methods from neuroscience and computer science and ideas and models from economics. In particular, the insights of mechanism design theory developed over the last 30 years are likely to be critical to make progress in designing new classes of neurometrically informed mechanisms.

## References

Aboyoun, D.C. & Dabbs, J.M. (1998) "The Hess pupil dilation findings: sex or novelty?" *Social Behavior and Personality: An international journal* 26, 415-419

Baron, D.P. & Besanko, D. (1984) "Regulation, asymmetric information, and auditing" *Rand Journal of Economics* 15, 447-470.

Bergstrom, T., Blume, L., & Varian, H. (1986) "On the private provision of public goods" *Journal of Public Economics* 29, 25-49

Border, K.C. & Sobel, J. (1987) "Samurai accountant: a theory of auditing and plunder" *The Review of Economic Studies* 54, 525-540

Bradley, M.M. (2000) "Emotion and motivation" in *Handbook of Psychophysiology*, eds. J.T. Cacioppo, L.G. Tassinary & G.G. Berntson, Cambridge University Press

Bradley, M.M., Miccoli, L., Escrig, M.A. & Lang, P.J. (2008) "The pupil as a measure of emotional arousal and autonomic activation" *Psychophysiology* 45, 602-607

Brier, G.W. (1950) "Probabilistic forecasts of precipitation in terms of quantiles using NWP model output", *Monthly Weather Review* 132, 338-347

Clarke, E. (1972) "Multi-part pricing of public goods" *Public Choice* 11, 17-33

Coppinger, V., Smith, V.L. & Titus, J. (1980) "Incentives and behavior in English, Dutch, and sealed-bid auctions" *Economic Inquiry* 18, 1-22

Cox, D.D. & Savoy, R.L. (2003) "Functional magnetic resonance imaging (fMRI) "brain reading": detecting and classifying distributed patterns of fMRI activity in human visual cortex", *Neuroimage* 19, 261-270

Cremer, J. & McLean, R.P. (1985) "Optimal selling strategies under uncertainty for a discriminating monopolist when demands are interdependent" *Econometrica* 53, 345-361

Cremer, J. & McLean, R.P. (1988) "Full extraction of the surplus in bayesian and dominant strategy auctions" *Econometrica* 56, 1247-1257

D'Aspremont, C., & Gerard-Varet, L. (1979) "Incentives and incomplete information" *Journal of Public Economics*, 11, 25-45.

Dimberg, U., Thunberg, M. & Elmehed, K. (2000) "Unconscious facial reactions to emotional facial expressions" *Psychological Science* 11, 86-89

Fehr, E. & Gächter, S. (2002) "Altruistic punishment in humans" *Nature* 415, 137-140

Gibbard, A. (1973) "Manipulation of voting schemes: a general result" *Econometrica* 41, 587-601

Green, J. & Laffont, J.J. (1977) "Characterization of satisfactory mechanisms for the revelation of preferences for public goods" *Econometrica* 45, 427-438

Gneiting, T. & Raftery, A.E. (2007) "Strictly proper scoring rules, prediction, and estimation" *Journal of the American Statistical Association* 102, 359-378

Groves, T. (1973) "Incentives in teams" *Econometrica* 41, 617-631

Groves, T. & Ledyard, J. (1977) "Optimal allocation of public goods: a solution to the 'free-rider' problem" *Econometrica* 45, 783-809

Harsanyi, J.C. (1967) "Games with incomplete information played by 'Bayesian' players" *Management Science* 14, 159-182

Haxby, J.V., Gobbini, M.I., Furey, M.L., Ishai, A., Schouten, J.L. & Pietrini, P. (2001) "Distributed and overlapping representations of faces and objects in ventral temporal cortex" *Science* 293, 2425-2430

Haynes, J., Sakai, K., Rees, G., Gilbert, S., Frith, C. & Passingham, R. (2007) "Reading hidden intention in the human brain" *Current Biology* 17, 323-328

Herrmann, B., Thoni, C. & Gächter, S. (2008) "Antisocial punishment across societies" *Science* 319, 1362-1367

Hurwicz, L. (1972) "On informationally decentralized systems" in *Decision and Organization*, eds. B. McGuire and R. Radner, Amsterdam: North Holland 297-336

Hurwicz, L. & Walker, M (1990) "On the generic nonoptimality of dominant-strategy allocation mechanisms: a general theorem that includes pure exchange economies" *Econometrica* 58, 683-704

Kahneman D. & Tversky, A. (1979) "Prospect theory: an analysis of decision under risk" *Econometrica* 47, 263-291

Kamitani, Y. & Tong, F. (2005) "Decoding the visual and subjective contents of the human brain" *Nature Neuroscience* 8, 679-685

Kay, K.N., Naselaris, T., Prenger, R.J. & Gallant, J.L. (2008) "Identifying natural images from human brain activity" *Nature* 452, 352-355

Krajbich, I., Camerer, C.F., Ledyard, J. & Rangel, A. (2009) "Using neural measures of economic value to solve the public goods free-rider problem" *Science* 326, 596-599



Lang, P.J., Greenwald, M.K., Bradley, M.M. & Hamm, A.O. (1993) "Looking at pictures: affective, facial, visceral, and behavioral reactions" *Psychophysiology* 30, 261-273

Ledyard, J.O. (1995) "Public goods: a survey of experimental research" *Handbook of Experimental Economics*, Princeton University Press

Lindahl, E. (1958) "Just taxation-a positive solution" in the *Theory of Public Finance*, eds. R. Musgrave and A. Peacock, Classics Macmillan, London 98-123

Maskin, E.S. (1999) "Nash Equilibrium and welfare optimality" *The Review of Economic Studies*, 66, 23-38

McAfee, R.P., & Reny, P.J. (1992) "Correlated information and mechanism design" *Econometrica* 60, 395-421

Myerson, R. (1981) "Optimal auction design" *Mathematics of Operations Research*, 6, 58-73

Nikiforakis, N. (2008) "Punishment and counter-punishment in public good games: can we really govern ourselves?" *Journal of Public Economics* 92, 91-112

Norman, K.A., Polyn, S.M., Detre, G.J. & Haxby, J.V. (2006) "Beyond mind-reading: multi-voxel pattern analysis of fMRI data" *Trends in Cognitive Sciences* 10, 424-430

O'Toole, A.J., Jiang, F., Abdi, H., Penard, N., Dunlop, J.P. & Parent, M.A. (2007) "Theoretical, statistical, and practical perspectives on pattern-based classification approaches to the analysis of functional neuroimaging data" *Journal of Cognitive Neuroscience* 19, 1735-1752

Partala, T., Jokiniemi, M. & Surakka, V. (2000) "Pupillary responses to emotionally provocative stimuli" in *The 2000 symposium on eye tracking research & applications* 123-129 (ACM Palm Beach Gardens, Florida, United States)

Pessoa, L., Padmala, S. (2007) "Decoding near-threshold perception of fear from distributed single-trial brain activation" *Cerebral Cortex* 17, 691-701

Polyn, S.M., Natu, V.S., Cohen, J.D. & Norman, K.A. (2005) "Category-specific cortical activity precedes retrieval during memory search" *Science* 310, 1963-1966

Satterthwaite, M. (1975) "Strategy-proofness and Arrow's conditions: existence and correspondence theorems for voting procedures and social welfare functions" *Journal of Economic Theory* 10, 187-217

Serences, J.T. & Boynton, G.M. (2007) "Feature-based attentional modulations in the absence of direct visual stimulation" *Neuron* 55, 301-312

Sokol-Hessner, P., Hsu, M., Curley, N.G., Delgado, M.R., Camerer, C.F., Phelps, E.A. (2009) "Thinking like a trader selectively reduces individuals' loss aversion" *Proceedings of the National Academy of Sciences of the United States of America* 106, 5035-5040

Tassinary, L.G. & Cacioppo, J.T. (1992) "Unobservable facial actions and emotions" *Psychological Science* 3, 28-33

Vickrey, W. (1961) "Counterspeculation, auctions, and competitive sealed tenders" *Journal of Finance* 16, 8-37

Walker, M. (1980) "On the nonexistence of a dominant-strategy mechanism for making optimal public decisions" *Econometrica* 48, 1521-15

Wang, S., Camerer, C. & Filliba, M. (2010). "Bayesian adaptive design for optimal elicitation of risk preferences", Caltech working paper.

## APPENDIX

### **I: Instructions from risk/loss-aversion experiment**

This experiment is a study of group decision making.

There is NO deception in the experiment: if we tell you that we are going to do something, we will do it exactly as described.

There will be two rounds in this experiment. In both rounds of the experiment you will be in a group of 5 people (four others and yourself) who have to decide whether to make a group investment. Each person in the group receives a different value from the investment if it is made. In one round your value for the investment will be \$1 and in the other your value will be \$9. The values of the other people in your group are also either \$1 or \$9. Their values are independent of each other— that is, each person's value is equally likely to be \$1 or \$9 regardless of how many other people have \$9 or \$1 values. To fill your group, we will randomly select rounds from the other people in the experiment and use the decisions they made in those rounds.

#### **The two decisions you make**

In both rounds you will make two decisions *after* you find out your value. Your decisions, and the decisions of the others in your group, will determine whether the group makes the investment, and will determine how much money you might earn.

Next we will describe the decisions you make. Then we will explain how the decisions of the people in your group lead to the money you can earn.

The first decision is to report whether your value is \$1 or \$9. You can report your actual value accurately, or report the opposite value from the one you have.

Your second decision is whether to vote YES or NO to make the investment.

You will *not* be told any information about the choices of others in your group between decisions or rounds.

### **How decisions determine your earnings**

Your earnings are determined by two steps:

#### **When the investment is *not* made**

The first step is whether the group decides to make the investment. If the investment is *not* made, everyone in your group earns nothing in that round.

The investment will *not* be made if either of two situations occurs.

(1) The investment has a cost to the group of \$25. The investment will *not* be made if the total of the values that are reported by all the group members (including you) is less than \$25. Since the possible values are either \$1 or \$9 for everyone, this means

mathematically that if the number of people reporting \$9 values is zero, one, or two, then the investment *will not* be made. If the number of people reporting \$9 is three, four or five, then the investment *might* be made, if the second condition (below) is also satisfied.

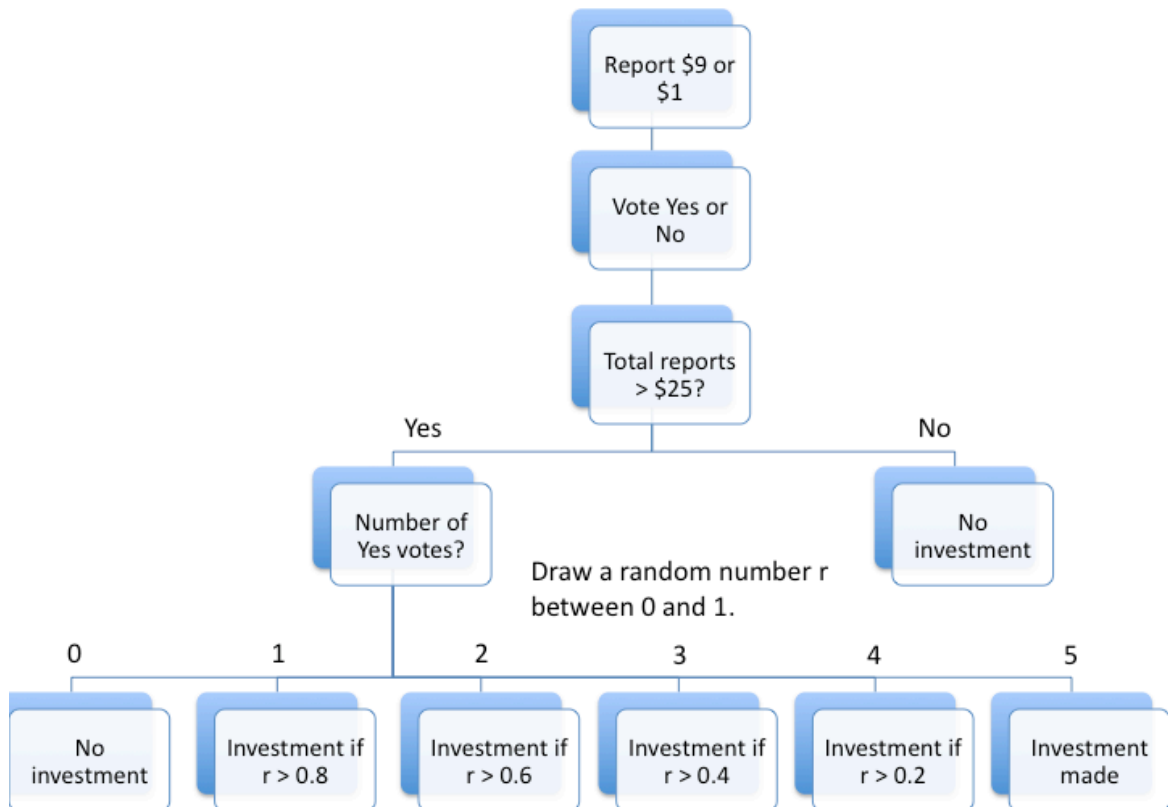
(2) Even if three or more people report values of \$9, whether the investment is made also depends on the number of YES and NO votes, and on chance. If everyone in your group votes YES the investment will certainly be made. If everyone in your group votes NO, the investment will certainly *not* be made. In general, the percentage chance that the investment will be made is 20% times the number of YES votes. For example, if three people vote YES (and the other two people vote NO) then the chance that the investment will be made is 60%.

The table below reminds you of the percentage chances that the investment takes place for all the possible numbers of votes in your group:

No. of YES votes	0	1	2	3	4	5
No. of NO votes	5	4	3	2	1	0
Chance the investment is made	0%	20%	40%	60%	80%	100%

After the percentage chance of investment is determined by the votes, a random number will be drawn to determine whether the investment is made.

To summarize, the investment is *not* made if two or fewer people report values of \$9, OR if there are NO votes and the chance outcome determines that no investment is made. Remember that if the investment is *not* made, everyone in your group earns nothing. Below is a timeline explaining this procedure graphically (with time flowing down):



### Your earnings if the investment *is* made

If the investment *is* made, your earnings will depend on three numbers (and on the numbers and decisions of the others in your group):

1. Your actual value (\$1 or \$9)
2. What you reported about your value (in your first decision)
3. A random guess generated by the computer that is related to your actual value

The random computer guess will be equal to your actual value with an 80% probability, and will be equal to the opposite value with a 20% probability. For example, if your value is \$9, then the computer guess will be \$9 with 80% probability and the computer guess will be \$1 with 20% probability. That means that if the computer did this 100 times, the most likely outcome would be 80 correct guesses and 20 incorrect ones.

Guesses of the values of the other four people in your group will be made independently in the same way.

If the investment is made, the first part of your earnings is just your actual value. If your actual value is \$1 you get \$1. If your actual value is \$9 you get \$9.

The second part of your earnings is an extra amount you pay or receive based on your *reported* value and the computer's guess of your *actual* value. The possible amounts are shown in the table below. If you report \$9 you pay \$9 no matter what the computer guesses. If you report \$1, however, you receive \$3.67 if the computer guesses \$1, or you pay \$14.67 if the computer guesses \$9.

		Your report	
		\$1	\$9
Computer's guess	Guess \$1	receive \$3.67	pay \$9
	Guess \$9	pay \$14.67	pay \$9

The third part of your earnings is an extra amount you either get or pay in order to pay the \$25 cost of the investment. The amounts paid by everyone in the group from the computer guess payments just described (minus the amounts received) will first be totaled up. If the total amount paid is greater than \$25, then any extra money paid (above \$25) will be distributed evenly back to everyone in the group. However, if the amount paid is less than \$25, the extra amount needed to create a total payment of \$25 will be collected evenly from the people in the group.

To summarize: If the investment is made, your earnings are the sum of three components:

1. Your actual value (either \$1 or \$9)
2. The additional amount you pay, or receive, which depends on your *reported* value and on the computer's guess
3. The amount redistributed to everyone (if the payments from part (2) are more than \$25), or collected from everyone (if the payments from part (2) are less than \$25).

Your dollar earnings from the experiment will be equal to the earnings from one of the two, randomly selected rounds.

### **The expected dollar earnings from voting YES and reporting your actual value**

Your choices in this experiment are whether to vote YES or NO, and whether to report your actual value (\$1 or \$9) or to report the opposite. The financial consequences of these decisions are rather complicated (and also depend on what others in your group do).

Therefore, it may be helpful to you to know *expected* cash earnings from one set of decisions you can make (also called “average” earnings). {If you are unfamiliar with the idea of an “expected” payoff please read the footnote on the next page.<sup>4</sup>} You can make

---

<sup>4</sup> The expected monetary payoff is the amount you would be likely to earn if you made the same decisions over and over, so that the relative frequency of chance events comes to be very close to the stated probability. For example, suppose you flip a fair coin and earn \$3 if heads comes up, and lose \$1 if tails comes up. If you flip many times, the percentage of times you earn \$3 and the percentage of times you lose \$1 would start to even out, to half and half. Then your average earnings would be  $.5(\$3) + .5(-\$1)$ , which is \$1. On any one coin flip you would not earn \$1—you



these decisions if you want to, or any other decisions you like. This section is simply designed to help you see the implications of a particular kind of strategy.

The rules and payments in this experiment are set up so that, the best way to maximize your average or expected earnings is to vote YES, and report your actual value. Making these decisions gives higher expected earnings regardless of the values, reports or computer guesses of the other people. If you report the opposite of your actual value in a round, your average or expected earnings for that round will be worse than if you had reported your actual value.

The payments are also designed so that your average or expected payoff is always positive if you report your actual value, regardless of what the other people are doing. Therefore, if you vote YES and report your actual value, your average or expected earnings will be positive. If you vote NO and report your actual value, your average or expected earnings will be positive, but closer to \$0 than if you had voted YES.

**The next two tables** below describe your actual earnings for all the possible reports and computer guesses for you and the other four people in your group.

Each column represents a different set of reports and guesses for the other people in your group. Since people who report \$9 pay \$9 regardless of the computer's guess, there is only one row when your report is \$9. Each column heading indicates the number of other people in the group who reported \$9, the number who reported \$1 and the computer guessed \$1, and the number who reported \$1 and the computer guessed \$9. These tables

---

would earn either \$3 or -\$1. But if you kept a running total of your average earnings it would start to get very close to \$1.

omit the possibility that there are zero or one \$9 reports because then the investment will not be made and you earn \$0 no matter what you do.

The first table shows your earnings if your actual value is \$9. The first row (in grey) shows your earnings if you report \$9. The second and third rows (in white) show your earnings if you report the opposite (reporting \$1 when your actual value is \$9) and the computer guesses \$9 (the second row) or \$1 (the third row). Keep in mind that the computer would guess \$9 with a probability of 80% and \$1 with a probability of 20%, as indicated in the “chance” column.

The second table shows your earnings when your actual value is \$1. The first and second rows (in grey) show your earnings if you report \$1 and the computer guesses \$1 (the fourth row) or \$9 (the fifth row). The third row (in white) shows your earnings if you report the opposite (reporting \$9 when your actual value is \$1). Keep in mind that the computer would guess \$1 with a probability of 80% and \$9 with a probability of 20%, as indicated in the “chance” column.

Actual Value	Value Report	Computer Guess	Chance	2 \$9 2 \$1 0 \$9 Guess	2 \$9 1 \$1 1 \$9 Guess	2 \$9 0 \$1 2 \$9 Guess	3 \$9 1 \$1 Guess 0 \$9 Guess	3 \$9 0 \$1 Guess 1 \$9 Guess	4 \$9
\$9	\$9	\$9/\$1	100%	-\$1.07	\$2.60	\$6.27	\$1.47	\$5.13	\$4
\$9	\$1	\$9	80%	\$0	\$0	\$0	-\$3.07	-\$0.60	-\$0.54
\$9	\$1	\$1	20%	\$0	\$0	\$0	\$11.60	\$15.27	\$14.14

Actual Value	Value Report	Computer Guess	Chance	2 \$9 2 \$1 Guess 0 \$9 Guess	2 \$9 1 \$1 Guess 1 \$9 Guess	2 \$9 0 \$1 Guess 2 \$9 Guess	3 \$9 1 \$1 Guess 0 \$9 Guess	3 \$9 0 \$1 Guess 1 \$9 Guess	4 \$9
\$1	\$1	\$1	80%	\$0	\$0	\$0	\$3.60	\$7.27	\$6.14
\$1	\$1	\$9	20%	\$0	\$0	\$0	-\$11.07	-\$7.40	-\$8.54
\$1	\$9	\$9/\$1	100%	-\$9.07	-\$5.40	-\$1.73	-\$6.53	-\$2.87	-\$4

**The table** below describes your average or expected earnings for reporting your actual value or reporting the opposite (called “misreporting”) for all the possible values and reports of the other people.

Each column represents a different set of reports by the other people in your group. Since people who report \$9 pay \$9 regardless of the computer’s guess, misreports by those people with actual \$1 values do not affect your payoffs, so there are no separate columns for those events. So in the table, the term “misreports” always refers to people with \$9 values who report \$1. Also, as in the previous tables, this table also omits the possibility that there are zero or one \$9 reports because then the investment will not be made and you earn \$0 no matter what you do.

The first two rows of the table show the average or expected earnings if your actual value is \$9. The first row shows your average earnings if you report \$9. The second row shows your average earnings if you report the opposite (reporting \$1 when your actual value is \$9). Notice that in *every* column describing what other people might do, the first row average earnings are higher than the second row average earnings.

The third and fourth rows show the average earnings when your actual value is \$1 and you report \$1 (the third row) or report the opposite, \$9 (the fourth row). In *every* column the third row average earnings are higher than the fourth row average earnings. In fact, the third row is always zero or positive, and the fourth row average earnings are always negative.

Actual Value	Value Report	2 \$9 0 Misreports	2 \$9 1 Misreport	2 \$9 2 Misreports	3 \$9 0 Misreports	3 \$9 1 Misreport	4 \$9
\$9	\$9	\$0.40	\$2.60	\$4.80	\$2.20	\$4.40	\$4.00
\$9	\$1	\$0	\$0	\$0	\$0.60	\$2.80	\$2.40
\$1	\$1	\$0	\$0	\$0	\$1.40	\$3.60	\$3.20
\$1	\$9	-\$7.60	-\$5.40	-\$3.20	-\$5.80	-\$3.60	-\$4.00

**QUIZ:**

Next you will answer a short quiz to ensure that you understand the instructions above.

If a statement is 'False', please rewrite the statement in a way that makes it 'True'.

Question 1: True or False: Your value for the investment is the same every round.

Question 2: True or False: To determine the investment values of the other group members, for each one we randomly choose a value of either \$9 or \$1.

Question 3: True or False: Given that the computer correctly guesses your value with a probability of 80%, you will always make more money on average if you report your actual investment value.

Question 4: True or False: If you vote NO for the investment, the investment will never be made.

Question 5: True or False: Even if you report your actual investment value, voting NO for the investment could help you avoid situations where your expected earnings are negative.

Question 6: True or False: Your actual earnings are always higher when you report your actual investment value.

**Response Sheet**

Please circle your decisions below for both rounds.

Your investment value is **\$9**.

Do you **report** **\$9** or **\$1** ?

Do you **vote** **YES** or **NO** ?

Your investment value is **\$1**.

Do you **report** **\$9** or **\$1** ?

Do you **vote** **YES** or **NO** ?

## IIA: Instructions from VCM experiment

This experiment is a study of group decision making.

There is NO deception in the experiment: if we tell you that we are going to do something, we will do it exactly as described.

The experiment will consist of **20 rounds**. In each round of the experiment you will be placed into a group of 5 players. In each round, you and the other members of your group will make a decision on the size of a group investment. Your decision and the decisions of the others will determine how much money you earn.

**Your decision** is how much to contribute to the group investment. You can contribute \$1 to \$20. We label this contribution  $m_i$ . The size of the investment,  $x$ , is determined by taking the sum of the contributions by the members of your group. That is,  $x = \sum_i m_i$ .

**Your payoff from an investment  $x$**  is  $v_i \cdot \ln(x)$ . Your value multiplier  $v_i$  is different in each round and is different for each person in your group. Your multiplier  $v_i$  can range from \$1 to \$20. This is true for every person. Later in the instructions we will provide earnings graphs so you don't need to do any calculation of your own.

**To summarize:** Your earnings in one round will be computed as

$$\text{Earnings} = \text{investment payoff} - \text{your contribution}$$

$$= v_i * \ln (\sum_k m_k) - m_i$$

At the end of the instructions we provide some graphs showing your average or expected earnings for different multipliers  $v_i$ , investments  $x$ , and contributions  $m_i$ .

You will be paid your average total earnings from all 20 rounds. Your average earnings will be displayed at the end of every round.

Procedure:

In each round you will first see a decision screen. On this screen you will see your value multiplier for that round. This multiplier changes from round to round, and is randomly drawn from \$1 to \$20.

Below your value multiplier, we ask you to report your contribution. Use the number keys to type in a number between 1 and 20 and click the red “Make Offer” button to submit your contribution. If you enter a value outside this range then the screen will refresh and you will be asked again to enter your contribution. You can use the “Backspace” key if you make a mistake while typing. This screen will look as follows:



The screenshot shows a game interface with a grey background. At the top left, it says "Period" and "1 of 10". At the top right, it says "Remaining time(sec): 572". In the center, it displays "Your multiplier for this round" as 5 and "Make your contribution for this round" as 3. A red button labeled "Make offer" is located at the bottom right.

After you've made your choice on this screen you will be told the total group contribution and your earnings for that round, as well as your average earnings from all the rounds so far. This screen will look as follows:



After you hit continue on this screen the next round will begin.

On the next pages are graphs showing your earnings as a function of your contribution and the contributions of the other players. Each graph is for a different value of your multiplier. On the x-axis is your contribution, and each line represents a different sized contribution from the other four players:

**QUIZ:**

Next you will answer a short quiz to ensure that you understand the instructions above.

If a statement is 'False', please rewrite the statement in a way that makes it 'True'.

Question 1: True or False: Your multiplier value for the investment is the same every round.

Question 2: True or False: To determine your multiplier value (and the multiplier values of other players) we randomly choose an amount from \$1 to \$20.

Question 3: How many rounds are there? To determine your final earnings do we take the sum of the earnings from those rounds, or the average earnings from those rounds?

Question 4: True or False: There are some situations where you can get higher earnings by contributing something other than your multiplier value.

**IIB: Instructions from augmented mechanism experiment**

This experiment is a study of group decision making.

There is NO deception in the experiment: if we tell you that we are going to do something, we will do it exactly as described.

The experiment will consist of **20 rounds**. In each round of the experiment you will be placed into a group of 5 players. In each round, you and the other members of your group will make a decision on the size of a group investment. Your decision and the decisions of the others will determine how much money you earn.

**Your decision** is how much to contribute to the group investment. You can contribute \$1 to \$20. We label this contribution  $m_i$ . The size of the investment,  $x$ , is determined by taking the sum of the contributions by the members of your group. That is,  $x = \sum_i m_i$ .

**Your payoff from an investment  $x$**  is  $v_i \cdot \ln(x)$ . Your value multiplier  $v_i$  is different in each round and is different for each person in your group. Your multiplier  $v_i$  can range from \$1 to \$20. This is true for every person. Later in the instructions we will provide earnings graphs so you don't need to do any calculation of your own.

In each round, a computer will randomly guess your multiplier,  $v_i$ , in a way that is related to the actual value for  $v_i$ . The guess will be drawn from a uniform distribution that ranges from +/- \$10 of your multiplier. For example, if your multiplier was \$12, then the computer would be equally likely to guess a number anywhere from \$2 to \$22. You will be charged a tax based on the computer's guess,  $g_i$ , and your contribution,  $m_i$ . Your tax

will be  $t(g_i, m_i) = (g_i - m_i)^2$ . Note that you pay the lowest tax when the computer's guess ( $g_i$ ) and your contribution ( $m_i$ ) are the same.

We will take these taxes from you and every other player in the group. We will then take the average tax paid by the *other* 4 players and pay that amount back to you as a refund. This refund process will happen for every player in the group.

**To summarize:** Your earnings in one round will be computed as

$$\begin{aligned} \text{Earnings} &= \text{investment payoff} - \text{your contribution} - \text{tax} + \text{refund} \\ &= v_i * \ln\left(\sum_k m_k\right) - m_i - [(g_i - m_i)^2] + \frac{1}{4}\left(\sum_{k \neq i} t_k\right) \end{aligned}$$

At the end of the instructions we provide some graphs showing your average or expected earnings for different multipliers  $v_i$ , investments  $x$ , and contributions  $m_i$ . Note that you don't have any influence on your refund, it is entirely dependent on the contributions of the other people in the group.

You will be paid your average total earnings from all 20 rounds. If your total earnings are negative then that amount will be deducted from your show-up fee. At the end of every round you will be told your computer guess, tax, refund, and earnings from that round, as well as the contributions of the other four people in your group.

Procedure:

In each round you will first see a decision screen. On this screen you will see your value multiplier for that round. This multiplier changes from round to round, and is randomly drawn from \$1 to \$20.

Below your value multiplier, we ask you to report your contribution. Use the number keys to type in a number between 1 and 20 and click the red “Make Offer” button to submit your contribution. If you enter a value outside this range then the screen will refresh and you will be asked again to enter your contribution. You can use the “Backspace” key if you make a mistake while typing. This screen will look as follows:

The screenshot shows a decision screen with a light gray background. At the top left, it says "Period" followed by "1 of 10". At the top right, it says "Remaining time[sec]: 572". In the center, there are two lines of text: "Your multiplier for this round" with the number "5" to its right, and "Make your contribution for this round" with a text input field containing the number "3". At the bottom right of the input field area, there is a red button labeled "Make offer".

After you’ve made your choice on this screen you will be told the total group contribution, your tax, your refund, and your earnings for that round, as well as your average earnings from all the rounds so far. This screen will look as follows:

Period		Remaining time[sec]: 0	
1 of 10		Please reach a decision	
Your contribution	3		
Group contribution	21		
Computer's guess of your multiplier	-2		
Tax	25		
Refund	2		
Your total profit last period		-10.5	
Your average profit in the experiment		-10.5	
<input type="button" value="Continue"/>			

After you hit continue on this screen the next round will begin.

### Strategy:

The investment payoff, contribution, tax and refund are all parts of your total earnings for a given round. To help you understand why the payoffs and taxes are the way they are, we will now highlight a few key points. When we talk about your average or expected earnings, we mean averaging across the different amounts you might earn depending on the computer's guess about your multiplier.

- (1) Given the distribution of the computer's guess, the taxes are chosen so that on average you make the most money by contributing an amount equal to your value multiplier, regardless of the multipliers, contributions, or computer guesses of the other players. If you make some other contribution in a round, your average or expected earnings for that round will be lower than if you had contributed your multiplier value.

- (2) If you contribute your multiplier value, your average earnings will always be positive. However, if you contribute a different amount, your average or expected earnings can become negative.

On the next pages are graphs showing your average earnings as a function of your contribution and the contributions of the other players. Each graph is for a different value of your multiplier. On the x-axis is your contribution, and each line represents a different sized contribution from the other four players. As you can see, your highest average earnings occur when your contribution is equal to your multiplier value:

### **QUIZ:**

Next you will answer a short quiz to ensure that you understand the instructions above. If a statement is 'False', please rewrite the statement in a way that makes it 'True'.

Question 1: True or False: Your multiplier value for the investment is the same every round.

Question 2: True or False: To determine your multiplier value (and the multiplier values of other players) we randomly choose an amount from \$1 to \$20.

Question 3: How many rounds are there? To determine your final earnings do we take the sum of the earnings from those rounds, or the average earnings from those rounds?

Question 4: True or False: There are some situations where you can get higher average or expected earnings by contributing something other than your actual multiplier value.



## CHAPTER 3

### **Visual Fixations Guide the Computation and Comparison of Value in Simple Choice**

There is a growing consensus in behavioral neuroscience that the brain makes simple choices by first assigning a value to all of the options under consideration and then comparing them.<sup>1-3</sup> This has motivated a growing interest in characterizing the exact computational properties of the processes responsible for the value comparison, and in understanding the extent to which they are able to generate reward-maximizing choices.

Although many popular models of value-based choice implicitly assume that the comparison process involves a trivial instantaneous maximization problem,<sup>4,5</sup> casual observation suggests that the underlying processes at work are more sophisticated and that visual fixations are likely to play a role. Consider, for example, a typical buyer at the grocery store choosing between two snacks: a bag of chips and a candy bar. Instead of approaching the counter and immediately selecting his preferred option, the individual's gaze shifts repeatedly between the items until one of them is eventually selected.

We propose a model of how simple value-based binary choices are made and of the role of visual fixations in the comparison of values. The model describes how the brain makes decisions in the experimental situation described in Fig. 1A, which simulates in the

laboratory a typical choice situation. Subjects are shown high-resolution pictures of two food items and are free to look at them as much as they want before indicating their choice with a button press. The model makes stark qualitative and quantitative predictions about the relationship between fixation patterns and choices, which we test using eye-tracking.

The theory developed here builds on the framework of drift-diffusion models of binary response selection,<sup>6-19</sup> and especially on applications of these models to the realm of perceptual decision making,<sup>18, 20-31</sup> where they have been shown to provide accurate descriptions of the psychometric data as well as provide important insights into the activity of the lateral intraparietal area (LIP). These models assume that stochastic evidence for one response (compared to the other) is accumulated over time until the integrated evidence passes a decision-threshold and a choice is made. The level of the threshold is set to balance the benefit of accumulating more information with the cost of taking more time to reach a decision.

There are two important differences between our work and the related studies in the realm of perceptual discrimination that are important to emphasize from the outset. First, in the standard Newsome-Shadlen random dot motion task, the subject is exposed to a single stochastic stimulus that provides signals about the value of two potential responses. In contrast, in our task subjects are exposed to two non-stochastic pictures of food items and have to estimate their value in order to select the most rewarding one. Second, fixations do not play a role in the standard perceptual discrimination task

because subjects maintain central fixation at all times. In contrast, in our task subjects fixate back and forth between the two stimuli. The key idea of the model that we propose is that fixations affect the drift-diffusion value comparison process by introducing a temporary drift bias towards the fixated item. This drift bias in turn leads to a choice bias to items that are fixated on more often. Our experimental results show that, as predicted by the model, there are strong correlations between the fixations and choices.

## Results

**Computational model.** Following the literature on drift-diffusion, our model assumes that the brain computes a relative decision value (RDV) that evolves over time as a Markov Gaussian process until a choice is made (Fig. 1B). The RDV starts each trial at 0, continually evolves over time at one of two possible rates (depending on the fixated item), and a choice is made when it reaches a barrier at either +1 or -1. If the RDV reaches the +1 threshold then the left item is chosen, if it reaches the -1 threshold then the right item is selected.

The key difference with the standard drift-diffusion model is that the slope with which the RDV signal evolves at any particular instant depends on the fixation location. In particular, the average rate at which the RDV changes over time is proportional to the weighted difference between the values of the fixated and non-fixated items. The weight discounts the value of the unfixated item relative to the fixated one.

Thus, when the subject is looking at the left item the evolution of the RDV is given by

$$V_t = V_{t-1} + d(r_{left} - \theta r_{right}) + \varepsilon_t$$

and when the subject is looking at the right item, the evolution of the RDV is given by

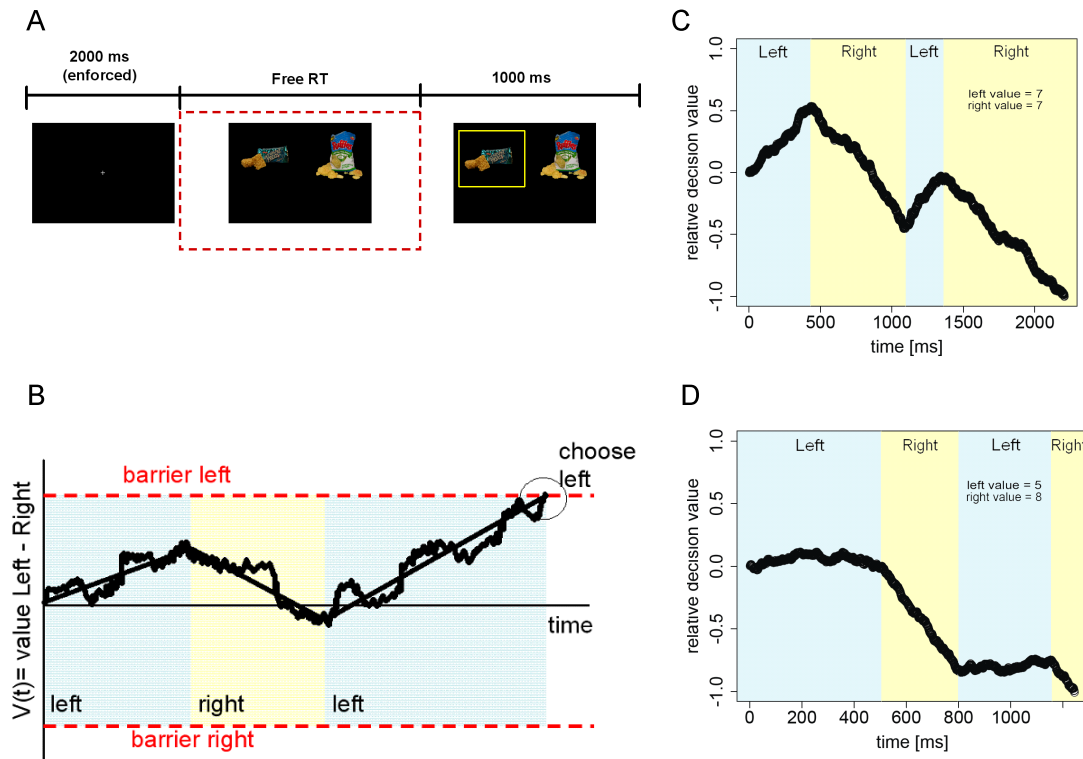
$$V_t = V_{t-1} - d(r_{right} - \theta r_{left}) + \varepsilon_t$$

where  $V_t$  is the value of the RDV at time  $t$ ,  $r_{left}$  and  $r_{right}$  denote the values of the two options,  $d$  is a constant controlling the speed of integration (in units of  $\text{ms}^{-1}$ ),  $\theta$  between 0 and 1 is a parameter reflecting the bias towards the fixated option, and  $\varepsilon_t$  is white Gaussian noise with variance  $\sigma^2$ , (randomly sampled once every ms).

With respect to the fixation process, the model assumes that the first fixation goes to the left item with probability  $p$ , independent of the values of the options, that fixations alternate between the two items until a barrier is crossed, and that fixations have a maximum duration given by a fixed distribution that depends on the difficulty of the choice, as measured by  $r_{best} - r_{worst}$ . Note that a fixation terminates if either its maximum duration is reached, or the RDV terminates the choice process by crossing a barrier.

Fig. 1C and 1D describe two simulated runs of the model and provide some intuition about the forces at work. Note that the evolution of the RDV is generally towards the

**Figure 1.** **A)** Choice trial. Subjects are forced to fixate at the center of the screen for 2 seconds. They are then presented with images of two items and given as much time as they want to make their choice. After a selection is made a yellow box highlights the chosen item for 1 second. **B)** Model. A relative decision value (RDV) evolves over time with a slope that is biased towards the item that is being looked at. The slope dictates the average rate of change of the RDV, but there is also an error term drawn from a Gaussian distribution. When the RDV hits the barrier a choice is made for the corresponding item. The shaded vertical regions represent the item being looked at. **C & D)** Simulated runs of the model using  $d=0.005$ ,  $\sigma = 0.05$ , and  $\theta = 0.6$ , in order to give a better intuition for the decision process.



fixated item, but that the rate of the evolution depends on the values of the two items. For example, in panel 1D, the RDV signal integrates towards the left barrier when the subject fixates on the left item, even though it has a lower value than the right item. This introduces a critical role for visual fixations in the integration process, a role that will be important in explaining the results described below.

**Hypotheses and Model Fitting.** We carry out a simple eye-tracking experiment to investigate the extent to which the drift-diffusion model outlined here is able to capture key patterns of the relationship between the fixation and choice data. We were particularly interested in testing between three alternative models: **M1)** The regular drift-diffusion model, given by the case of  $\theta = 1$ ; **M2)** A drift-diffusion model with full fixation bias, given by the case of  $\theta = 0$ ; and **M3)** A drift-diffusion model with partial fixation bias, given by the case  $0 < \theta < 1$ . The experiment consists of two stages. In the first stage subjects are presented with images of 70 different snack food items and are asked to rate how much they would like to eat each item at the end of the experiment, using a scale from -10 to 10. The liking ratings provide an independent measure of the value of individual items. In the second stage subjects are asked to make choices between pairs of neutral or appetitive foods. Subjects complete 100 such choice trials and afterwards eat the item they chose in a randomly selected trial. During the choice stage we measured eye-movements at a rate of 50 Hz. In the analyses below we use the liking ratings from stage 1 as measures of the value of the two items.

We fitted the model to the even numbered trials from the group data using maximum likelihood estimation (see Methods for details). The model has three free parameters: the constant determining the speed of integration  $d$ , the discount parameter  $\theta$ , and the noise parameter  $\sigma$ . The model was fit under the assumption that time evolves in 1 ms discrete steps. We selected the parameters that maximized the probability of the observed choices and reaction times, conditional on the values of the items. The best-fitting model had parameters  $d = 0.0002ms^{-1}$ ,  $\theta = 0.3$ ,  $\sigma = 0.02$ , with a log-likelihood value of -3704.

We also used the same procedure to fit models with  $\theta = 1$  and  $\theta = 0$ . In both cases the best-fitting models had parameters  $d = 0.0002ms^{-1}$  and  $\sigma = 0.02$ . We then used the likelihood ratio statistic to test that  $\theta$  was significantly less than 1 (log-likelihood = -3708,  $p < 0.008$ ) and significantly larger than 0 (log-likelihood = -3710,  $p < 0.0005$ ). This provides support for M3 over the standard and full fixation bias drift-diffusion models. (See also Table 1 and Fig. S1-S10, which compare the fits of all the relevant figures presented below for the best-fitted values of the three models). We also carried out a restricted fit of the model to individual subject data. The mean (standard deviation) estimated  $\theta$  value from individual model fits was 0.52 (0.3), and 35/39 subjects have an estimate of this parameter less than 1 (see Methods and Fig. S11-S12 of the Appendix).

In order to investigate the ability of the model to predict the data quantitatively, we then simulated the model 1000 times per pair of liking ratings, using the estimated maximum likelihood parameters, and by sampling fixation lengths from the actual empirical fixation

data (taking into account that fixation durations are related to decision difficulty, as described below). Throughout, we assume that fixations always alternate between the two items, and that the location of the first fixation is chosen probabilistically to match the empirical data (look left first with probability 74%). The results of the simulations are described below. Note that in all comparisons of the model to the data, we present only the odd numbered trials, since the model was fitted to the even numbered trials.

**Basic psychometrics.** Figure 2 describes the match between the simulated and the actual data. Not surprisingly, given that the parameters were chosen to fit these two variables, the model predicts the choice and reaction time curves quite well in the odd trials. Fig. 2A provides the results for the choice data ( $\chi^2$  goodness-of-fit statistic = 4.47,  $p=0.92$ ), and shows that choices are a logistic function of the value differences, which means that the best option is only selected 78 % of the time. Note that the amount of noise in the choice process is controlled by the Gaussian noise in the integration process, and by the random fixation durations. This figure also shows the comparably poor-fitting  $\theta = 0$  model.

Fig. 2B provides analogous results for reaction times (goodness-of-fit:  $p=0.10$ , see Methods for details), and shows that reaction times decrease with difficulty (mixed effects regression estimate -211 ms/rating,  $p<10^{-11}$ ), a property of drift-diffusion models that also extends to our model.



**Table 1.** Summary of the goodness-of-fit statistics for all the figures in the main text. Each number is the p-value from the goodness-of-fit test of that particular model to the data. Note that the intermediate model ( $\theta = 0.3$ ) fits better than the other two models in most cases except Fig. 4C and 5D/E (indicated by the gray shading).

	2A	2B	2C	4B	4C	5A left	5A right	5B	5C	5D	5E
$\theta = 0.3$	0.92	0.1	0.39	0.997	0.824	0.96	0.96	0.75	0.0062	0.21	0.0016
$\theta = 0$	$10^{-5}$	0.01	$10^{-5}$	0.01	0.96	0.83	0.19	0.0002	$10^{-13}$	0.76	0.1
$\theta = 1$	$10^{-16}$	0.0007	$10^{-15}$	$10^{-16}$	0.04	$10^{-16}$	0.0001	$10^{-13}$	$10^{-11}$	0.0009	$10^{-9}$

**Figure 2.** **A)** Psychometric choice curve. **B)** Reaction times as a function of the difference in liking ratings between the best and worst items, which is a measure of difficulty. **C)** Number of fixations in a trial as a function of choice difficulty. In all figures the red dashed line indicates the simulated data using the MLE parameters and the subject data only includes the odd numbered trials. In A, the blue dash dotted line indicates the simulated data for the  $\theta = 0$  model. Bars denote standard errors, which are clustered by subject. Tests are based on a paired two-sided t-test.

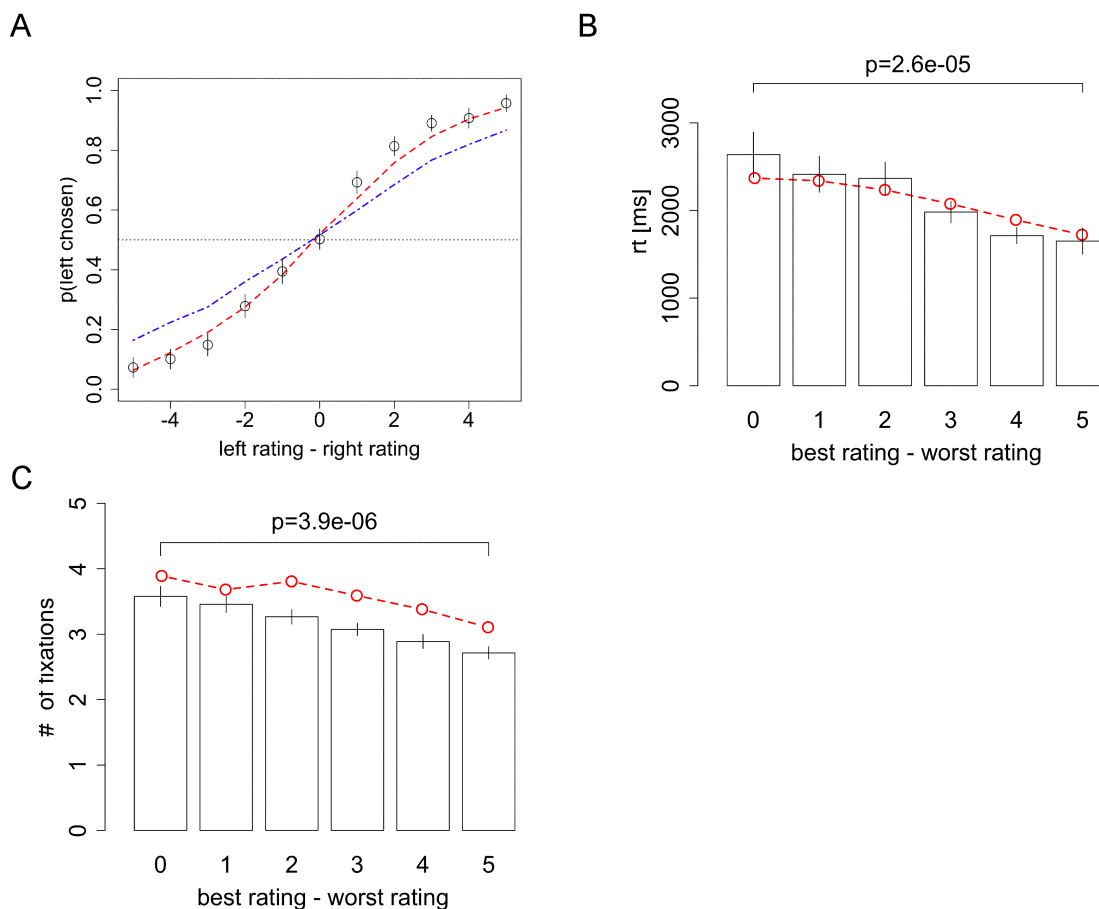
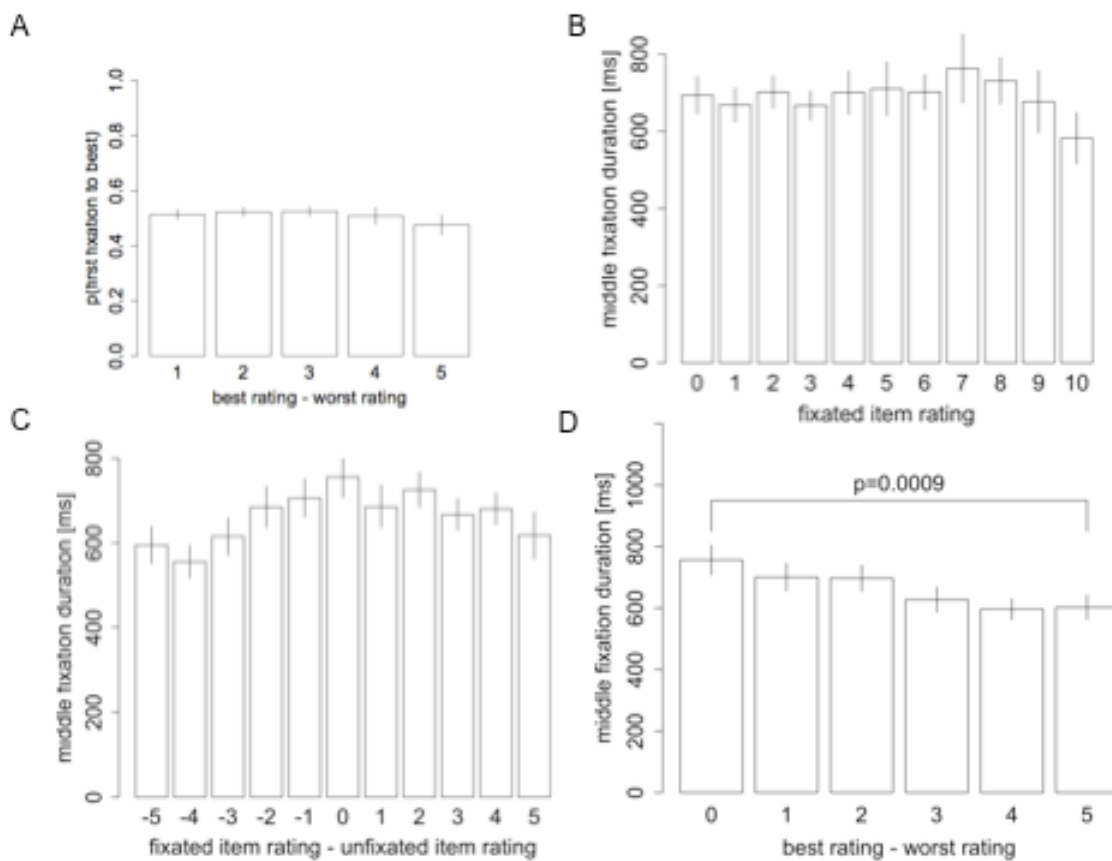


Fig. 2C shows that the model can also account quantitatively for the number of fixations (goodness-of-fit:  $p=0.39$ ), even though this psychometric match was not targeted by the model fitting procedure, and that the number of fixations decrease with difficulty (mixed effects regression estimates  $-0.171$  fixations/rating,  $p<10^{-15}$ ).

**Properties of the visual search process.** The model makes strong assumptions about the nature of the fixation process, which are tested in Figure 3. First, Fig. 3A shows the probability that the first fixation is to the best item, which is insignificantly different from 0.5 and is unaffected by the difference in ratings (mixed effects regression estimates: intercept = 0.518,  $p<0.31$ ; slope =  $-0.0009$ /rating,  $p<0.88$ ). Second, Fig. 3B shows that the middle fixation duration is independent of the value of the fixated item (mixed effects regression estimate: 6.4 ms/rating,  $p<0.21$ ). Third, Fig. 3C and Fig. 3D show that there is a slight dependence of middle fixation duration on the difference in value between the fixated and non-fixated items (mixed effects regression estimate: 11.4 ms/rating,  $p<0.0052$ ), but that there is a larger dependency of middle fixation duration on the difficulty of the decision (mixed effects regression estimate  $-33.8$  ms/rating,  $p<10^{-5}$ ). Note that the dependency of middle fixation durations on value depicted in Fig. 3D is taken into account in the estimation and simulation procedures since we assume that fixation durations are drawn from the observed empirical distribution for trials with the same level of difficulty. (See also Fig. S13-S15 for analogous figures of the first fixation properties).

**Figure 3.** **A)** Probability that the first fixation is to the best item. In all cases they are not significantly different from 50%. **B)** Middle fixation duration as a function of the liking rating of the fixated item. **C)** Middle fixation duration as a function of the difference in liking ratings of the fixated and unfixated items. **D)** Middle fixation duration as a function of the difference in liking ratings of the best and worst rated items. Bars denote standard error bars, which are clustered by subject. Tests are based on a paired two-sided t-test.



We also fitted the empirical distribution of middle and first fixations to several alternative statistical distributions and found the best fits with log-normal distributions (log-likelihood = -24740 for the first fixations and = -34802 for the middle fixations; see Table S1 and Fig. S16-S22 and Methods).

**Core model predictions.** Albeit extremely simple, the model makes several strong predictions about the relationship between visual attention, choices, and reaction times that we also tested using the eye-tracking data.

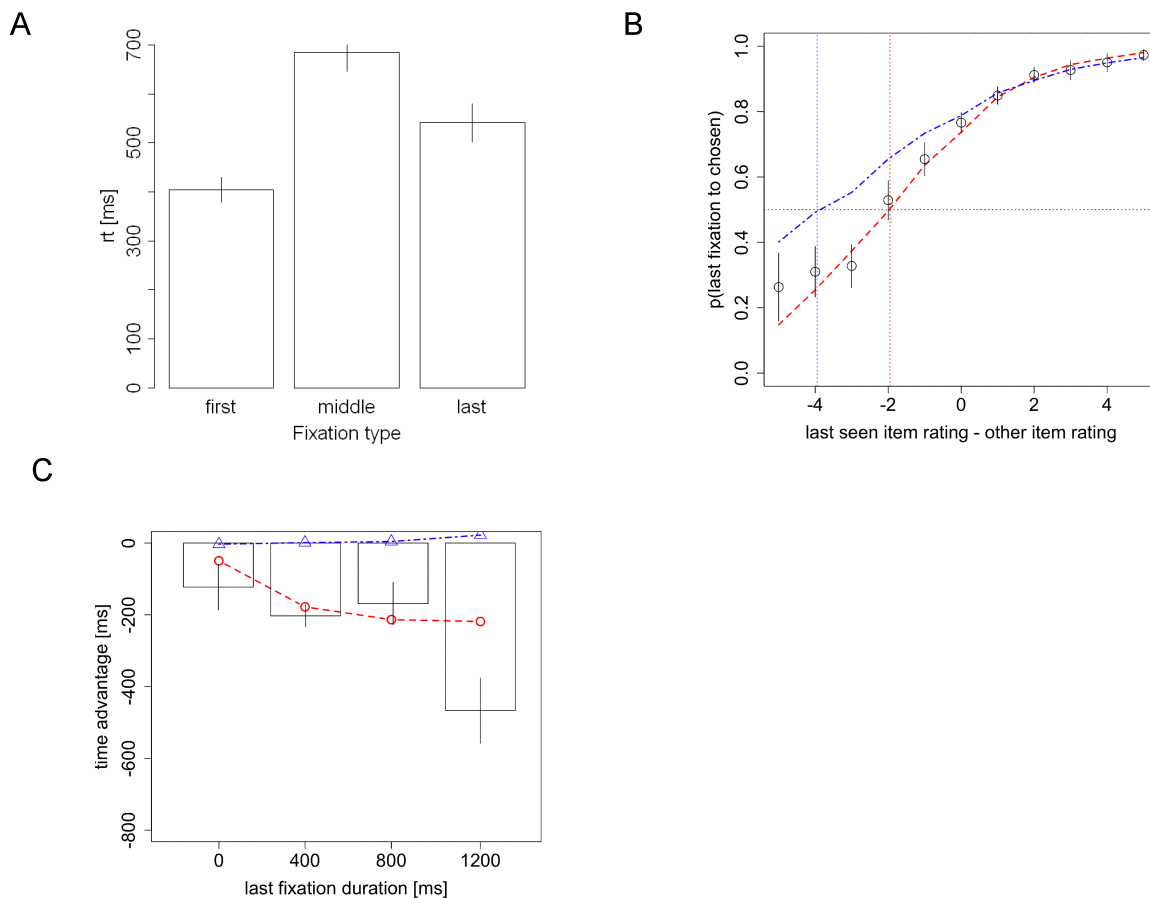
First, consistent with the data, the model predicts that final fixations should be shorter than middle fixations, since fixations are interrupted when a barrier is crossed (Fig. 4A,  $p < 0.0002$ ). The figure also shows that first fixations were shorter than middle ones ( $p < 10^{-14}$ ), which was not predicted ex-ante by the theory, but was incorporated ex-post into the computational model's estimation and simulation procedures.

Second, the model predicts that subjects will generally choose the item they looked at last, unless that item is much worse than the other one. Fig. 4B shows that this was indeed the case ( $\chi^2$  goodness-of-fit statistic = 1.96,  $p = 0.997$ ). To see why this must be the case, recall that the RDV climbs towards the barrier of the fixated item unless the fixated item is sufficiently worse, so that the drift rate becomes negative. This figure also shows the comparably poor-fitting  $\theta = 0$  model.

Third, the model predicts that the longer you have looked at item A during a trial, the longer you will have to look at item B before choosing it over item A. The intuition for this is simple: on average, the longer one looks at item A the farther the RDV gets from item B's barrier, and thus the farther it will have to travel back in order to hit that threshold. As shown in Fig. 4C, the effect is marginal, but the model prediction is consistent with the data (goodness-of-fit:  $p = 0.82$ ; mixed effects regression coefficient =  $-0.08$ ,  $p < 0.11$ ). This figure also shows how poorly the  $\theta = 1$  model fits in this analysis.

**Choice biases.** The model also predicts that when  $\theta < 1$  the decision processes should exhibit several choice biases. First, it predicts a last fixation bias: subjects should be more likely to choose an item (for a given rating difference) if their last fixation is to that item as opposed to the other item. Furthermore, this difference should become more pronounced as the decision becomes more difficult. This is a prediction of the model because the value of the unfixated item is always discounted (by  $\theta$ ) relative to the fixated item. Fig. 5A shows that there is a sizable bias in both the simulated and the subject data (logit mixed effect regression:  $p < 10^{-16}$ ,  $\chi^2$  goodness-of-fit statistic = 3.64,  $p = 0.96$  for last fixation left and = 3.58,  $p = 0.96$  for last fixation right). The  $\theta = 1$  model predicts no effect of the last fixation, and this is clearly rejected by the data.

**Figure 4.** **A)** Fixation duration by type. Middle fixations indicate any fixations that were not the first or last fixations of the trial. **B)** Probability that the last fixation is to the chosen item as a function of the difference in liking ratings between the fixated and unfixated items in that last fixation. **C)** Amount of time spent looking more at Item B prior to the last fixation (to Item A), as a function of the duration of that last fixation. In all figures with simulation data, the red dashed line indicates the simulated data using the MLE parameters, and the subject data only includes the odd numbered trials. In B, the blue dash dotted line indicates the simulated data for the  $\theta = 0$  model, and the vertical dotted lines indicate the points at which the simulation curves cross the horizontal line at chance. In C, the blue dash dotted line indicates the simulated data for the  $\theta = 1$  model. Bars denote standard error bars, which are clustered by subject.

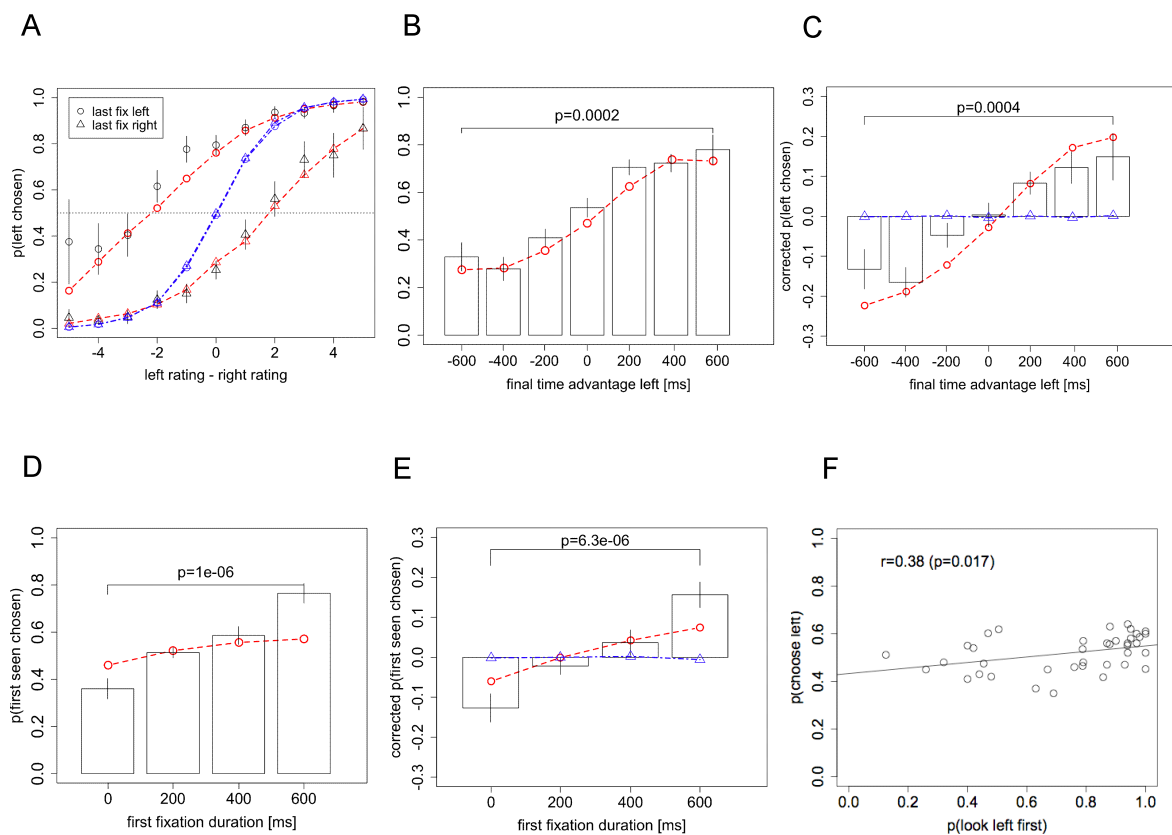


Second, the model predicts that there should be a choice bias that depends on the total amount of time spent looking at one item versus the other. Controlling for value differences, the probability of choosing an item should increase with the excess time that it is looked at. A mixed effects logit regression shows that this is indeed the case ( $p < 10^{-8}$ ). This prediction follows from the fact that the RDV always evolves more towards an item's barrier when it is being looked at than when it is not. Fig. 5B, and the fact that fixation duration and order are independent of an item's value, show this bias in the data and the simulations ( $\chi^2$  goodness-of-fit statistic = 3.43,  $p = 0.75$ ). Fig. 5C further tests this hypothesis by correcting for the difference in liking ratings. For each trial we take the actual choice (1 or 0) and subtract the average probability that left was chosen in all trials with that difference in liking ratings. These "corrected" choice probabilities are plotted in Fig. 5C as a function of the fixation time advantage for the left item (goodness-of-fit:  $p = 0.0062$ ). This eliminates any possible influence of the measured liking ratings on the fixation durations and shows that there is a substantial effect of total fixation time on choice. The  $\theta = 1$  model predicts no effect of exposure time on choices, and again this is clearly rejected by the data.

In a related result, Fig. 5D shows that the duration of the first fixation is predictive of whether that first-seen item will be chosen ( $\chi^2$  goodness-of-fit statistic = 4.55,  $p = 0.21$ , mixed effects regression:  $p < 0.028$ ). Analogous to Fig. 5C, Fig. 5E corrects for the difference in liking ratings between the first-seen item and the other item (goodness-of-fit:  $p = 0.0016$ ). Again, this eliminates any possible influence of the measured liking ratings on the first fixation durations and shows that there is a substantial effect of first



**Figure 5.** **A)** Psychometric choice curve conditional on the location of the last fixation. **B)** Probability that left is chosen as a function of the excess amount of time that the left item was fixated on during the trial. **C)** Analogous to B, except subtracting the probability of choosing left for each difference in liking ratings. **D)** Probability that the first-seen item is chosen as a function of the duration of that first fixation. **E)** Analogous to D, except subtracting the probability of choosing the first-seen item for each difference in liking ratings. **F)** Probability of choosing left as a function of the probability of looking left first. Each circle represents a different subject. In all figures with simulation data, the red dashed line indicates the simulated data using the MLE parameters, the blue dash dotted line indicates the simulated data for the  $\theta = 1$  model, and the subject data only includes the odd numbered trials. Bars denote standard error bars, which are clustered by subject. Tests are based on a paired two-sided t-test, except in F, where we use standard two-sided t-tests.



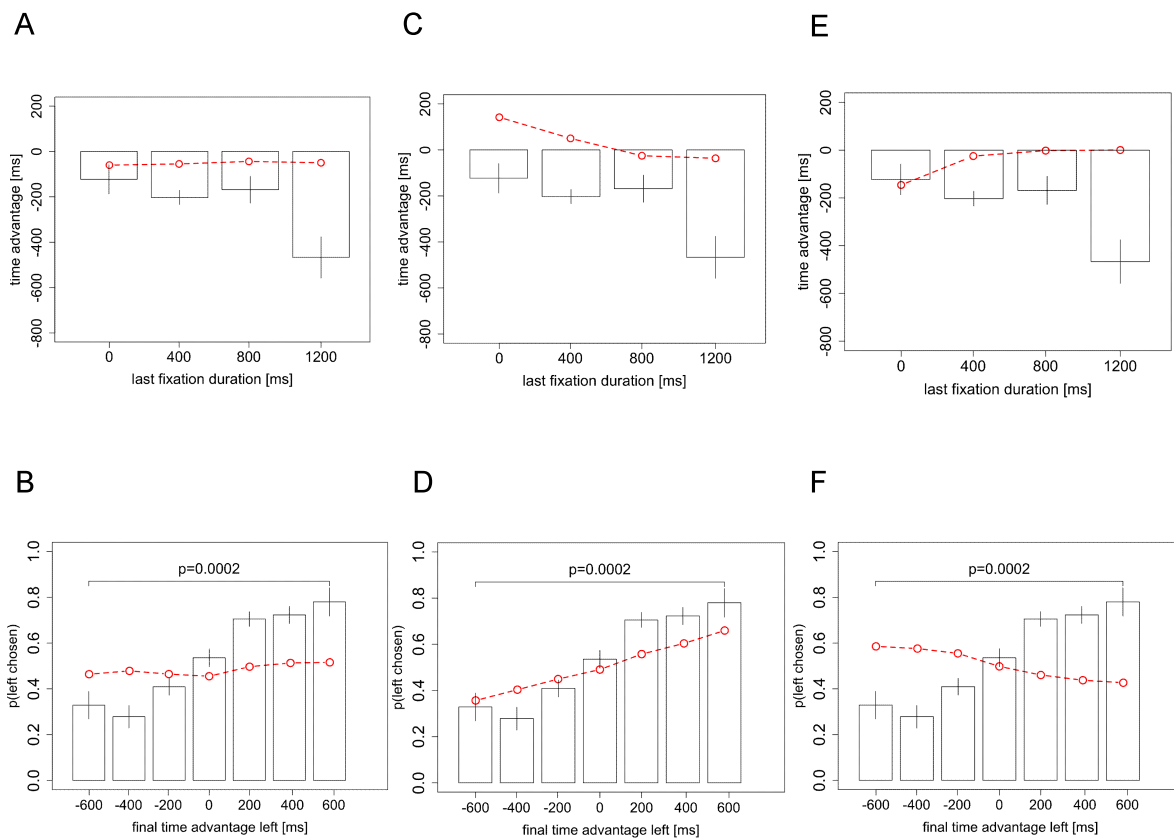
fixation duration on choice. Just like before, the  $\theta = 1$  model fails to explain the relationship between the first fixation duration and choice.

Third, the model predicts that any left-looking biases should translate into left choice biases. Fig. 5F shows that this was also the case: the more likely a subject is to look left first, the more likely he is to choose items on the left, with a correlation of 0.38 ( $p = 0.017$ ) and a Spearman's non-parametric correlation of 0.49 ( $p = 0.0017$ ). Again, this is consistent with the model since a fixation bias towards one item increases the likelihood of hitting that item's choice threshold.

**Alternative models.** There are several other ways in which fixations could interact with the decision process without using the exact model developed above. In order to investigate the robustness of our model here we investigate the ability of three natural alternatives to account for the data (see Methods for details).

The first alternative model also uses a drift-diffusion model framework, but assumes that the drift rate with which the RDV signal is integrated is independent of the fixations. Instead, the model assumes that fixation lengths are affected by the current value of the RDV signal. In particular, we assume that at each time point the probability that the current fixation will be terminated is inversely proportional to the distance between the RDV and the choice barrier for the unfixated item. We investigated the qualitative

**Figure 6.** Replications of Fig 4C and 5B with comparisons to alternative model 1 (panels **A-B**), alternative model 2 (panels **C-D**), and alternative model 3 (panels **E-F**). In all figures the red dashed line indicates the simulated data using the alternative model, and the subject data only includes the odd numbered trials. Bars denote standard error bars, which are clustered by subject. Tests are based on a paired two-sided t-tests.



properties of this model by simulating it using the best-fitting parameters from the  $\theta = 1$  version of our model. The results in Fig 6A-B and Fig. S23 show that while this model approximates the basic psychometric data in Fig. 2A-C well, it cannot account for the choice biases seen in Fig. 5B and 4C.

The second alternative assumes that a RDV signal is computed and affected by fixations as in our main model, but differs on how a decision is triggered. In particular, we assume that the decision time is determined exogenously by a separate unmodeled process, and that the subject chooses the option with the best RDV at that time. We investigated the qualitative properties of this model by simulating it using the best-fitting parameters from the  $\theta = 0.3$  version of our model, and randomly sampling decision times from the actual empirical distribution. The results in Fig. 6C-D and Fig. S24 show that, while the model accounts for the psychometric data reasonably well, it is unable to account for some of the choice biases discussed above.

The third alternative model is similar to our model, except that the fixations now change the locations of the choice barriers rather than the drift rate. We simulated this model using the best-fitting parameters from the  $\theta = 1$  model, and assuming that the magnitude of the barrier for the fixated item is lowered from 1 to 0.8. The value of 0.8 was chosen so that the model would fit the trends in Fig. 2A-C (see Fig. S25). However, as before, Fig. 6E-F show that although the model approximates the basic psychometric data reasonably well, it cannot account for some of the critical choice biases.

## Discussion

The results presented here provide insight into the nature of the computational and psychological processes guiding simple choices. In particular, we found that a simple extension of the drift-diffusion model in which fixations play a casual role in the value integration process is able to provide a remarkable quantitative account of various relationships between the fixation and choice data, as well as of several sizable choice biases.

A robustness analysis was also able to rule out several natural alternative models for how fixations might interact with the choice process. The first alternative model addressed whether fixations might simply reflect the evolution of a standard drift-diffusion model by dwelling on items that have a high RDV. A second model addressed whether the choice process is bounded, by allowing the RDV to evolve over the entire predetermined reaction time. Finally, a third alternative model addressed whether fixations affect the drift rate or the location of the choice barriers. Each of these models fits the basic psychometrics in Fig. 2 quite well, but was unable to explain some critical choice biases. Of course, we cannot rule out the existence of other models that could account for the patterns in this dataset better than the model presented here. However, given the results presented here, it would be surprising if one could find a model that could do so without the fixations modulating the value computation or comparison process.

A critical question raised by our results is whether the visual fixation process has a *causal* effect on the value comparison process. Several pieces of evidence suggest that this might be the case. First, our model assumes that this causal effect is present within a drift-diffusion model framework and is able to provide a remarkable quantitative fit to many moments of the data. Second, we have shown that a simple alternative drift-diffusion model in which values affect fixations, but in which the opposite is not true, cannot account for key aspects of the data. Third, consistent with the findings reported here, a couple of related studies have shown that it is possible to bias choices by manipulating relative fixation durations<sup>32,33</sup>. For example, Armel, Beaumel, and Rangel<sup>32</sup> manipulated fixation durations while subjects made binary choices using stimuli identical to the ones used in this chapter and were able to bias choices for sufficiently close items by about 6-15%. However, it is important to emphasize that the evidence provided here is not sufficient to establish a causal effect of fixations on choices.

Our model does not rule out the possibility that values might have some effect on the pattern of fixations. In fact, as shown in Fig.3D, and assumed in the model's estimation and simulation procedures, fixation durations in our dataset increase with the difficulty of the choice. However, the key aspect of the model and results is that even if these feedback results are present, random variation in fixation durations might affect the choice process itself. In this study we have treated these feedback effects from values to fixations as exogenous. Understanding the computational properties of these effects is an important open question for future research in this area.

The theory developed here builds on the framework of drift-diffusion models of binary response selection<sup>6-19</sup>, and especially on applications of these models to the realm of perceptual decision making<sup>18, 20-31</sup>. Several differences between the two approaches are worth discussing. First are differences in the nature of the computational problem. In the standard Newsome-Shadlen random dot motion task, subjects are exposed to a stochastic stimulus that is assumed to generate perceptual noise signals in area MT. Under appropriate assumptions, it can be shown that the drift-diffusion model implements an optimal decision making process that amounts to a sequential-likelihood ratio test<sup>18, 23, 24, 34, 35</sup>. In our model the stimuli are non-stochastic, in the sense that the image is non-changing. However, we hypothesize that in order to construct value the brain needs to integrate a series of noisy signals about the value of the stimuli, in this case generated internally. In particular, we hypothesize that the brain assigns value to the stimuli by sequentially and stochastically extracting features of the stimuli, retrieving the learned values for such features, and then integrating those values. Although the objective nature of the noise is quite different in both cases, in the absence of fixations, the computational problem has similar properties. The second difference has to do with the role of the fixations. In particular, fixations do not play a role in the standard perceptual discrimination task because subjects maintain central fixation at all times. In our task subjects fixate back and forth between the two stimuli. The key idea of the model that we propose is that fixations affect the way the choice is made by introducing a temporary bias in the drift-diffusion process towards items that are looked at more. The results in the

chapter and the Appendix are well explained by the model, which predicts the various correlations between the eye-tracking and the psychometric data.

Our model is also related to the models of decision field theory developed by Busemeyer and collaborators<sup>13, 37-40</sup>, which also consider sequential integration models in the spirit of the drift-diffusion model in which fixations matter. There are several differences between this literature and our study. First, decision field theory assumes that items are multidimensional and that fixations matter only to the extent that they focus the integration of value in one dimension or another. In contrast, we focus on choices among unidimensional stimuli and fixations matter because they bias the integration of value in favor of one of the items. Second, the predictions of decision field theory regarding the impact of fixations on choice have not been tested directly using eye-tracking.

Note that while the drift-diffusion model implements an optimal statistical decision process in the case of perceptual decision making, the model that we investigated here does not seem to have that property. In particular, it is difficult to reconcile the presence of the integration bias with an optimal decision process. An important question for future research is to determine the extent to which the model approximates an optimal Bayesian decision making problem in which fixations are determined endogenously, perhaps with a switching cost.



One critical question is how the brain implements this model of decision making. One brain region that is likely critical is the medial orbital frontal cortex (mOFC). A number of studies have shown that the mOFC encodes value signals at the time of choice<sup>41-47</sup>, which are the likely inputs to the comparator process studied here. We conjecture that fixations affect this process by amplifying the relative value signal for the fixated item in the mOFC.

The results have several important implications for the quality of choice processes and decision making in general. First, since fixations might be attracted by other visual features of the items that are uncorrelated with value, such as their visual saliency<sup>48</sup> or their location, the model predicts that such irrelevant factors could affect choices. A couple of studies have shown such effects by exogenously manipulating relative exposure times<sup>32,33</sup>. Second, the model more generally predicts that systematic biases in fixations could lead to deficits in decision making. Extensions of this framework might help to understand why individuals with autism who generally avoid eye contact exhibit deficits in social decision making<sup>49</sup>. Finally, the model explains how cultural norms (e.g., reading left to right) can interact with basic computational processes to produce cultural choice biases which help to explain, for example, why shelf and computer screen space on the top-left is more valuable.

## Methods

*Subjects.* 39 Caltech students participated in the experiment. Only subjects who self-reported regularly eating the snacks foods (e.g., potato chips and candy bars) used in the experiment and not being on a diet were allowed to participate. These steps were taken to ensure that the food items we used would be motivationally relevant. This would not have been the case if the subjects did not like junk food. Subjects were paid a \$20 show-up fee, in addition to receiving one food item (as described below). Caltech's Human Subjects Internal Review Board approved the experiment.

*Task.* Subjects were asked to refrain from eating for 3 hours prior to the start of the experiment. After the experiment subjects were required to stay in the room with the experimenter for 30 minutes while eating the food item that they chose in a randomly selected trial (see below). Subjects were not allowed to eat anything else during this time. The experiment was programmed using E-Prime software.

The experiment had two phases. In the first phase subjects provided liking ratings for 70 different food items. Every trial subjects were presented with a high-resolution picture of a different food item for 3 seconds. Subjects were instructed to fixate on the item for the full 3 seconds. On the next screen, subjects were asked to rate the item on a scale from -10 to 10, indicating how much they would like to eat the item at the end of the experiment. They did so using an on-screen slider bar, manipulated using the left and right arrow keys on the keyboard. The initial location of the slider was randomized to

reduce anchoring effects. This rating screen had a free response time. The food was kept in the room with the subjects during the experimental session to assure them that all the items were available. Furthermore, subjects briefly saw all the items at this point so that they could effectively use the rating scale.

In the second phase subjects made choices between pairs of food items (Fig. 1A). Every trial subjects were presented with two food items and asked to choose which one they would rather eat at the end of the experiment. They were told that one trial would be randomly chosen at the end of the experiment, and that they would receive the food item they chose in that trial. Subjects made their choice on this screen by pressing the left or right arrow keys on the keyboard to indicate a choice of the left or right item, respectively. This choice screen had a free response time. There were 100 trials in this phase of the experiment.

Food items that received a negative rating in the *rating* phase of the experiment were excluded from the *choice* phase. We did not tell subjects about this feature of the experiment because doing so could have changed their incentives during the *rating* phase.

The items shown in each trial were chosen pseudo-randomly according to the following rules: (1) No item was used in more than 6 trials; (2) The difference in liking ratings between the two items was constrained to be 5 or less; (3) If at some point in the

experiment (1) and (2) could no longer both be satisfied, then the difference in allowable liking ratings was expanded to 7, but these trials occurred for only five subjects and so were discarded from the analyses. The location of the items (right vs. left) was completely randomized.

After subjects indicated their choice, a yellow box was drawn around the chosen item (with the other item still on-screen) and displayed for 1 second. This feedback screen was followed by a fixation screen before the beginning of the next trial.

*Eye-tracking.* Subjects' fixation patterns were recorded using a Tobii (Sweden) desktop-mounted eye-tracker. This eye-tracker recorded fixation location at a rate of 50 Hz. We recorded macrofixations using an ROI method (i.e., we recorded whether subjects were looking at the left item, right item, or elsewhere) but we did not record microfixations.

Before each choice trial, subjects were required to maintain a fixation at the center of the screen for 2 seconds before the items would appear. This was done to ensure that subjects started every choice screen fixating on the same location.

*Data analysis.* Choice trials with gaps in the eye-tracking data at the beginning or end of the trial were excluded from analysis. The mean (SEM) number of trials dropped per subject was  $2.8 \pm 1.5$ . We define a gap as a period of time greater than 40 ms when the

eye-tracker did not record a fixation on either item. For all measurements following the first item fixation and preceding the last item fixation of the trial, blank fixations were dealt with according to the following rules:

(1) If the blank fixations were recorded between fixations on the same item, then those blank fixations were changed to that item. So for example a fixation pattern of “Left Item”, “Blank,” “Left Item” would become “Left Item”, “Left Item”, “Left Item.” The assumption here is that the eye-tracker simply lost the subject’s eyes during this time. The alternate hypothesis is that the subject looked away from the item without looking at the other item, but we consider this to be an unlikely scenario.

(2) If the blank fixations were recorded between fixations on different items, then those blank fixations were recorded as non-item fixations and discarded from further analysis. The assumption here is that the subject took time to shift his gaze from one item to the other, and during that time was not fixating on either item.

*Group model fitting.* The computational model was fit to the choice and reaction time data from the even numbered trials of the pooled data from the 39 subjects.

The maximum likelihood estimation procedure was implemented as follows. First, we set apart half of our data to estimate the model (the half was given by the even trials). Then

for each set of parameters and pair of liking ratings in the data we ran 1000 simulations of the model. In the simulations we randomly sampled fixation times from the empirical distribution conditional on the measure of choice difficulty given by  $r_{best} - r_{worst}$ . Given that first fixations were generally shorter than the rest, we sampled first fixation durations separately from the rest. We also used the empirical fact that subjects looked left first 74% of the time and that the first fixations were independent of value. Finally, the simulations assume instantaneous transitions between fixations while in the data there are often delays between fixations. To compensate, we calculated the total amount of “transition” time in each trial, randomly sampled from the empirical distribution of those “transition” times, and added them to the simulated reaction times. Note that the simulations used a time step size of 1 ms, which is significantly shorter than the average amount of time between measured fixations (20 ms).

Second, we computed the probability of each data point for each set of parameters as follows. The empirical spread of reaction times ranged from 525 ms to 25 s so in the fitting procedure we discarded any simulation trials below 500 ms or above 25 s. The rest of the reaction times were separated into bins from 500-6000 ms, each one spanning 100 ms, except for the first bin which went from 500-1000 ms and the final bin, which went from 7000-25000 ms. For each combination of liking ratings, we then split the data into the trials where Left was chosen and where Right was chosen, and then for each group, and counted the number of data trials in each reaction time bin, and similarly calculated the probability that a simulation trial occurs in each reaction time bin.

Third, we computed the set of parameters that maximized the log-likelihood of the data by taking the logarithms of each of these probabilities and summing them up. The resulting number is used to assess how well the model fit the data, with larger numbers (closer to zero) indicating better fits.

In the simulations, we vary  $\sigma$  as a function of the slope  $d$ , rather than absolutely. Therefore, we let  $\sigma = d * \mu$  and perform a grid search over values of  $d$ ,  $\mu$ , and  $\theta$ . Given the computational expense of this estimation procedure, the search for the maximum likelihood parameters was carried out in three steps. First we did a coarse grid search with  $d$  in  $\{0.0001, 0.00015, 0.0002, 0.00025\}$ ,  $\mu$  in  $\{80, 100, 120, 140\}$  and  $\theta$  in  $\{0.3, 0.5, 0.7, 0.9\}$ . Second, we used the results from the first search to define a finer search with  $d$  in  $\{0.000175, 0.0002, 0.000225\}$ ,  $\mu$  in  $\{90, 100, 110\}$  and  $\theta$  in  $\{0.2, 0.3, 0.4\}$ . The resulting log-likelihood value was -3704.

*Group likelihood ratio tests.* We tested whether  $\theta$  was significantly different from 1 and 0, by performing likelihood ratio tests. These tests use the results from the MLE described above, as well as those from another MLE model in which  $\theta$  was fixed to 0 or 1. This procedure was carried out exactly the same as before, using only the even numbered trials and starting with a coarse search with  $d$  in  $\{0.0001, 0.00015, 0.0002, 0.00025\}$  and  $\mu$  in  $\{80, 100, 120, 140\}$ , followed by a finer search with  $d$  in  $\{0.000175,$

$0.0002, 0.000225\}$  and  $\mu$  in  $\{90, 100, 110\}$ . The best-fitting set of parameters in both cases was  $d = 0.0002ms^{-1}$  and  $\mu = 100$ , with a log-likelihood value of  $-3708$  for  $\theta = 1$  and  $-3710$  for  $\theta = 0$ .

From these log-likelihood values we calculated the likelihood ratio statistic, which for the case of  $\theta = 1$  is given by:

$$LR = 2 * (x(\theta = 0.3) - x(\theta = 1))$$

Here,  $x$  is just the log-likelihood value for each set of parameters. This test statistic is distributed as  $\chi^2(1)$ .

*Group simulations.* We carried out 1000 simulations for every combination of values in the data set using the maximum likelihood parameter estimates and sampling fixation durations from the odd numbered trials.

*Individual model fitting.* An important concern with the group fits above is that they do not provide a good description of the underlying distribution of parameters in the subject population. We investigated this issue with two analyses.

In the first analysis (Fig. S11) we set  $d = 0.0002ms^{-1}$  and  $\mu = 100$  from the group level analysis and performed an MLE grid search over  $\theta$  in  $\{0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7,$



0.8, 0.9, 1} using all trials (even and odd), and 1000 simulations for every combination of values.

In the second analysis (Fig. S12) we calculated the average difference in left-choice probability between last-fixation-left trials and last-fixation-right trials, for each subject. This amounts to calculating the average difference between the two curves shown in Fig. 5A. Subjects with  $\theta=1$  should show no difference between these two curves, whereas subjects with  $\theta=0.3$  and  $\theta=0$  should show differences of 0.47 and 0.58, respectively (assuming  $d = 0.0002ms^{-1}$  and  $\mu = 100$ ).

*Goodness-of-fit calculations.* For Fig. 2B, 2C, 4C, 5C, and 5E we could not compute  $\chi^2$  goodness-of-fit statistics because the dependent variables are not binary.  $R^2$  statistics were also uninformative because of the high variability in average fixation duration from subject to subject. Therefore, we devised a different goodness-of-fit statistic that works as follows: (1) For each value of the independent variable we “correct” the dependent variable by subtracting the average simulated value from each subject’s average value. (2) We then run a weighted least-squares regression, regressing the “corrected” dependent variable on the independent variable. The weights in the WLS regression were equal to the inverse of the variance.

If the simulations fit the data well, then the “corrected” data should be a flat line at 0. On the other hand, if the simulation fits poorly, then the WLS coefficient should be non-zero.

So for goodness-of-fits, we report the p-values for the coefficients of these WLS regressions. If the p-values are less than 0.05 then we reject that the model fits the data.

*Fitting the fixation distributions.* In order to determine the best-fitting distributions for the first and middle fixation durations, we used a log-likelihood method to fit several different types of distributions to all the trials, as well as dividing trials by the absolute difference in the liking ratings. Table S1 summarizes the best-fitting parameters from log-normal distributions (which were consistently the best or near-best distribution) and the log-likelihoods for the different distributions. Figures S16-S22 show the log-normal fits to the data.

*Mixed effect regressions.* All mixed effect regressions mentioned in the chapter were run using R's *lmer* function, with random effects for subject-specific constants and slopes as well as fixed effects for the relevant independent variable. The three regression coefficients pertaining to Fig. 3 (and similarly S13-S15) all came from the same mixed effects regression, which included as independent variables the fixated item rating, the non-fixated item rating, and the absolute difference between the ratings.

*Alternative model simulations.* The three alternative models displayed in Fig. 6 and Fig. S23-25 were each simulated using 500 runs for every combination of values in the

dataset, and fixation durations and reaction times were sampled (where appropriate) from the odd numbered trials.

For the alternative model, the probability of the fixation ending at each time point was

$$p = \frac{0.002}{3 - k}$$

where  $k$  is the magnitude of the distance between the RDV and the choice barrier for the currently fixated item. These parameter values were chosen to roughly fit the psychometric patterns seen in Fig. 2.

For the second alternative model, reaction times were sampled from the empirical distribution conditional on the difference in liking ratings. The RDV was allowed to evolve for the entire reaction time. At the end of the trial, if the RDV was positive then the left item was chosen, if the RDV was negative then the right item was chosen.

For the third alternative model, when a fixation was made to the left item, the left choice barrier was lowered to 0.8. When a fixation was made to the right item, the right choice barrier was lowered to -0.8. These values were chosen to roughly fit the psychometric patterns seen in Fig. 2.

## References

1. Rangel, A., Camerer, C. & Montague, P.R. A framework for studying the neurobiology of value-based decision making. *Nature Reviews* **9**, 545-556 (2008).
2. Wallis, J.D. Orbitofrontal cortex and its contribution to decision-making. *Annual Review of Neuroscience* **30**, 31-56 (2007).
3. Padoa-Schioppa, C. & Assad, J.A. Neurons in the orbitofrontal cortex encode economic value. *Nature* **441**, 223-226 (2006).
4. Kahneman, D. & Tversky, A. Prospect theory: an analysis of decision under risk. *Econometrica* **4**, 263-291 (1979).
5. Mas-Colell, A., Whinston, M. & Green, J. *Microeconomic Theory* (Cambridge University Press, Cambridge, 1995).
6. Luce, R.D. *Response Times: Their Role in Inferring Elementary Mental Organization* (Oxford University Press, Oxford, 1986).
7. Stone, M. Models for choice-reaction time. *Psychometrika* **25**, 251-260 (1960).
8. Ratcliff, R. A theory of memory retrieval. *Psychological Review* **85**, 59-108 (1978).
9. Ratcliff, R., Cherian, A. & Segraves, M. A comparison of macaque behavior and superior colliculus neuronal activity to predictions from models of two-choice decisions. *Journal of Neurophysiology* **90**, 1392-1407 (2003).
10. Ratcliff, R. & Smith, P. A comparison of sequential sampling models for two-choice reaction time. *Psychological Review* **111**, 333-367 (2004).
11. Laming, D. A critical comparison of two random-walk models for choice reaction time. *Acta Psychologica* **43**, 431-453 (1979).
12. Link, S.W. *The Wave Theory of Difference and Similarity* (Lawrence Erlbaum, Hillsdale, NJ, 1992).
13. Usher, M. & McClelland, J. The time course of perceptual choice: the leaky, competing accumulator model. *Psychological Review* **108**, 550-592 (2001).
14. Smith, P. Psychophysically-principled models of visual simple reaction time. *Psychological Review* **102**, 567-593 (1995).
15. Smith, P. Stochastic dynamic models of response time and accuracy: a foundational primer. *Journal of Mathematical Psychology* **44**, 408-463 (2000).
16. Ditterich, J. Stochastic models of decisions about motion direction: behavior and physiology. *Neural Networks* **19**, 981-1012 (2006).
17. Bogacz, R. Optimal decision-making theories: linking neurobiology with behaviour. *Trends in Cognitive Sciences* **11**, 118-125 (2007).
18. Gold, J.I. & Shadlen, M.N. Neural computations that underlie decisions about sensory stimuli. *Trends in Cognitive Sciences* **5**, 10-16 (2001).
19. Gold, J.I. & Shadlen, M.N. Banburisms and the brain: decoding the relationship between sensory stimuli, decisions, and reward. *Neuron* **36**, 299-308 (2002).

20. Churchland, A.K., Kiani, R. & Shadlen, M.N. Decision-making with multiple alternatives. *Nature Neuroscience* **11**, 693-702 (2008).
21. Ditterich, J. Stochastic models of decisions about motion direction: behavior and physiology. *Neural Networks* **19**, 981-1012 (2006).
22. Ditterich, J., Mazurek, M.E. & Shadlen, M.N. Microstimulation of visual cortex affects the speed of perceptual decisions. *Nature Neuroscience* **6**, 891-898 (2003).
23. Gold, J.I. & Shadlen, M.N. Banburisms and the brain: decoding the relationship between sensory stimuli, decisions, and reward. *Neuron* **36**, 299-308 (2002).
24. Gold, J.I. & Shadlen, M.N. The neural basis of decision making. *Annual Review of Neuroscience* **30**, 535-574 (2007).
25. Hanks, T.D., Ditterich, J. & Shadlen, M.N. Microstimulation of macaque area LIP affects decision-making in a motion discrimination task. *Nature Neuroscience* **9**, 682-689 (2006).
26. Huk, A.C. & Shadlen, M.N. Neural activity in macaque parietal cortex reflects temporal integration of visual motion signals during perceptual decision making. *Journal of Neuroscience* **25**, 10420-10436 (2005).
27. Mazurek, M.E., Roitman, J.D., Ditterich, J. & Shadlen, M.N. A role for neural integrators in perceptual decision making. *Cerebral Cortex* **13**, 1257-1269 (2003).
28. Palmer, J., Huk, A.C. & Shadlen, M.N. The effect of stimulus strength on the speed and accuracy of a perceptual decision. *Journal of Vision* **5**, 376-404 (2005).
29. Roitman, J.D. & Shadlen, M.N. Response of neurons in the lateral intraparietal area during a combined visual discrimination reaction time task. *Journal of Neuroscience* **22**, 9475-9489 (2002).
30. Shadlen, M.N. & Newsome, W.T. Motion perception: seeing and deciding. *Proceedings of the National Academy of Sciences of the USA* **93**, 628-633 (1996).
31. Shadlen, M.N. & Newsome, W.T. Neural basis of a perceptual decision in the parietal cortex (area LIP) of the rhesus monkey. *Journal of Neurophysiology* **86**, 1916-1936 (2001).
32. Armel, K.C., Beaumel, A. & Rangel, A. Biasing simple choices by manipulating relative visual attention. *Judgment and Decision Making* **3**, 396-403 (2008).
33. Shimojo, S., Simion, C., Shimojo, E. & Sheier, C. Gaze bias both reflects and influences preference. *Nature Neuroscience* **6**, 1317 - 1322 (2003).
34. Bogacz, R. Optimal decision-making theories: linking neurobiology with behaviour. *Trends in Cognitive Science* **11**, 118-125 (2007).
35. Bogacz, R., Brown, E., Moehlis, J., Holmes, P. & Cohen, J.D. The physics of optimal decision making: a formal analysis of models of performance in two-alternative forced choice tasks. *Psychological Review* **113**, 700-765 (2006).
36. Reddi, B.A.J. & Carpenter, R.H.S. The influence of urgency on decision time. *Nature Neuroscience* **3**, 827-830 (2000).

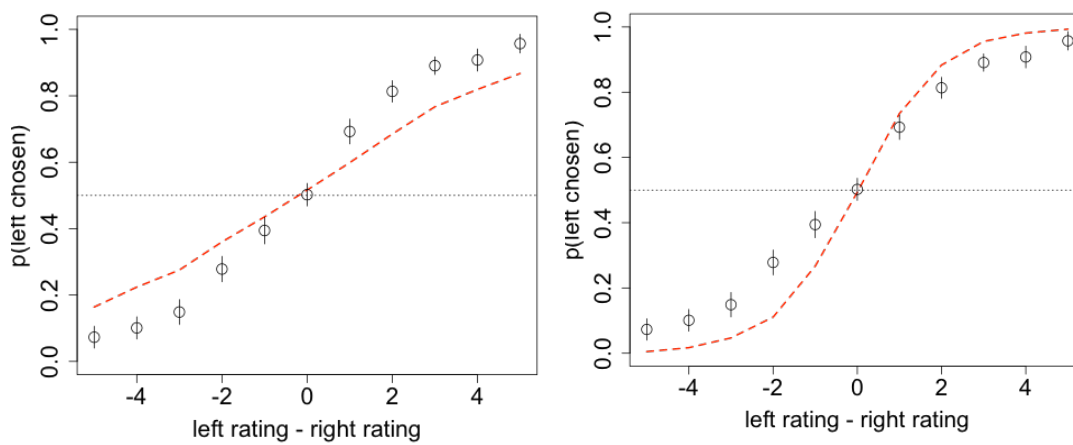
37. Roe, R.M., Busemeyer, J. & Townsend, J.T. Multialternative decision field theory: a dynamic connectionist model of decision making. *Psychological Review* **108**, 370-392 (2001).
38. Busemeyer, J. & Townsend, J.T. Decision field theory: a dynamic-cognitive approach to decision-making in an uncertain environment. *Psychological Review* **100**, 432-459 (1993).
39. Busemeyer, J. & Johnson, J.G. Computational models of decision making. in *Handbook of Judgment and Decision Making* (eds. D. Koehler & N. Narvey) 133-154 (Blackwell Publishing Co., New York, NY, 2004).
40. McClelland, J. & Rumelhart, D. An interactive activation model of context effects in letter perception: I. An account of basic findings. *Psychological Review* **88**, 375-407 (1981).
41. Plassmann, H., O'Doherty, J. & Rangel, A. Orbitofrontal cortex encodes willingness to pay in everyday economic transactions. *Journal of Neuroscience* **27**, 9984-9988 (2007).
42. Hare, T.A., O'Doherty, J., Camerer, C.F., Schultz, W. & Rangel, A. Dissociating the role of the orbitofrontal cortex and the striatum in the computation of goal values and prediction errors. *Journal of Neuroscience* **28**, 5623-5630 (2008).
43. Erk, S., Spitzer, M., Wunderlich, A., Galley, L. & Walter, H. Cultural objects modulate reward circuitry. *NeuroReport* **13**, 2499-2503 (2002).
44. Arana, F.S., *et al.* Dissociable contributions of the human amygdala and orbitofrontal cortex to incentive motivation and goal selection. *Journal of Neuroscience* **23**, 9632-9638 (2003).
45. Paulus, M.P. & Frank, L.R. Ventromedial prefrontal cortex activation is critical for preference judgments. *NeuroReport* **14**, 1311-1315 (2003).
46. Valentin, V.V., Dickinson, A. & O'Doherty, J. Determining the neural substrates of goal-directed learning in the human brain. *Journal of Neuroscience* **27**, 4019-4026 (2007).
47. Hare, T.A., Camerer, C.F. & Rangel, A. Self-control in decision-making involves modulation of the vmPFC valuation system. *Science* **324**, 646-648 (2009).
48. Itti, L. & Koch, C. Computational modelling of visual attention. *Nature Reviews* **2**, 194-203 (2001).
49. Spezio, M.L., Adolphs, R., Hurley, R.S.E. & Piven, J. Abnormal use of facial information in high-functioning autism *Journal of Autism and Developmental Disorders* **37**, 929-939 (2007).

## APPENDIX

**Table S1.** Best-fitting parameters for various statistical distributions that were fit to the fixation data. The log-normal distributions consistently provided the best fit, as indicated by gray shading. Each row indicates the fit for either all the trials, or for only those trials with a particular absolute difference in liking ratings. In a couple cases the gamma distribution would not fit the data, resulting in a value of NaN.

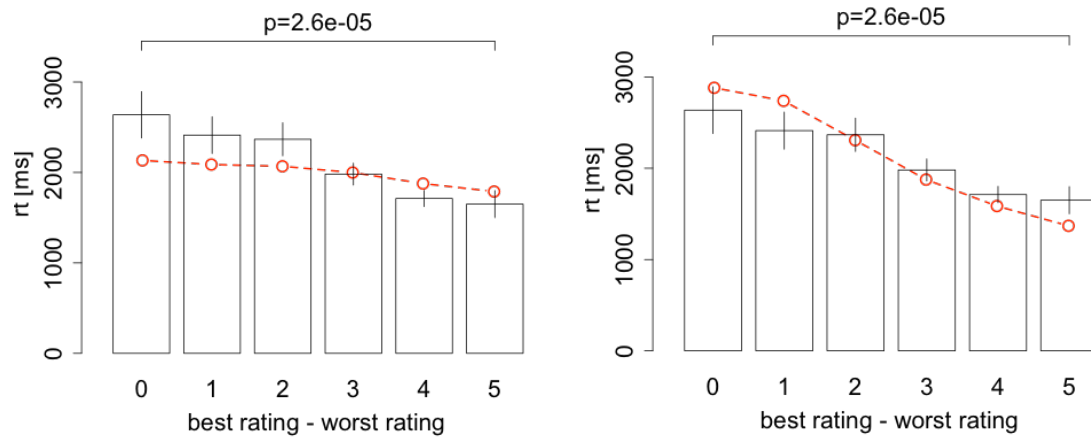
	log normal distribution			log likelihoods for other distributions			
	meanlog	sdlog	log likelihood	gamma	cauchy	negative binomial	normal
middle fixations							
all trials	6.39	0.63	-34802	-35295	-35109	-35008	-36849
diff = 0	6.46	0.65	-6139	NaN	-6160	-6180	-6559
diff = 1	6.45	0.66	-10093	-10203	-10243	-10149	-10624
diff = 2	6.39	0.64	-7988	NaN	-8056	-8050	-8507
diff = 3	6.33	0.58	-5126	-5158	-5163	-5136	-5340
diff = 4	6.28	0.54	-2736	-2742	-2762	-2735	-2822
diff = 5	6.28	0.53	-1866	-1887	-1898	-1878	-1952
first fixations							
all trials	5.83	0.62	-24740	-25021	-24884	-24755	-26223
diff = 0	5.83	0.58	-3460	-3541	-3448	-3479	-3727
diff = 1	5.83	0.66	-6170	-6231	-6205	-6155	-6546
diff = 2	5.87	0.59	-5429	-5541	-5504	-5471	-5814
diff = 3	5.84	0.56	-4188	-4205	-4239	-4187	-4356
diff = 4	5.8	0.64	-2762	-2759	-2767	-2748	-2863
diff = 5	5.8	0.68	-2064	-2096	-2074	-2068	-2207

**Figure S1:** Reproduction of Fig. 2A from the text but with the simulation data from  $\theta = 0$  (left) and  $\theta = 1$  (right). The  $\chi^2$  goodness-of-fit statistics for  $\theta = 0$  and  $\theta = 1$  were 39.6 ( $p < 10^{-5}$ ) and 192 ( $p < 10^{-16}$ ), respectively.

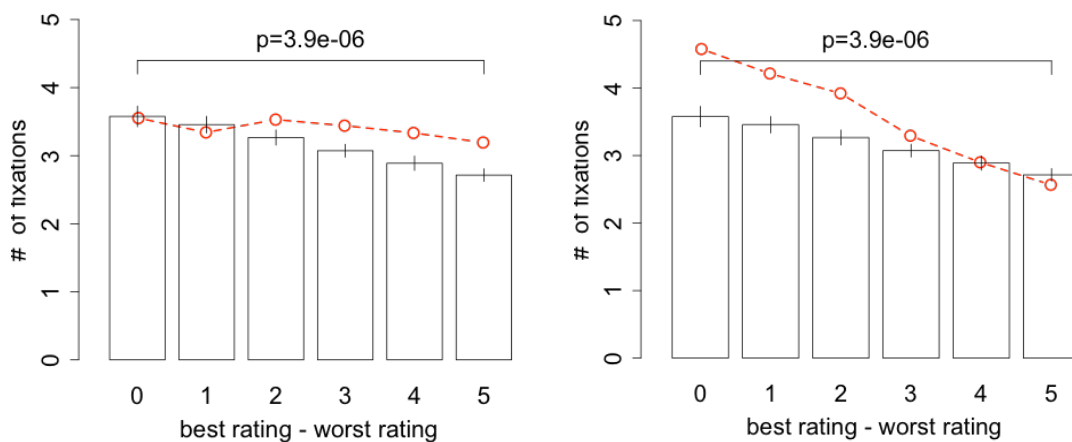




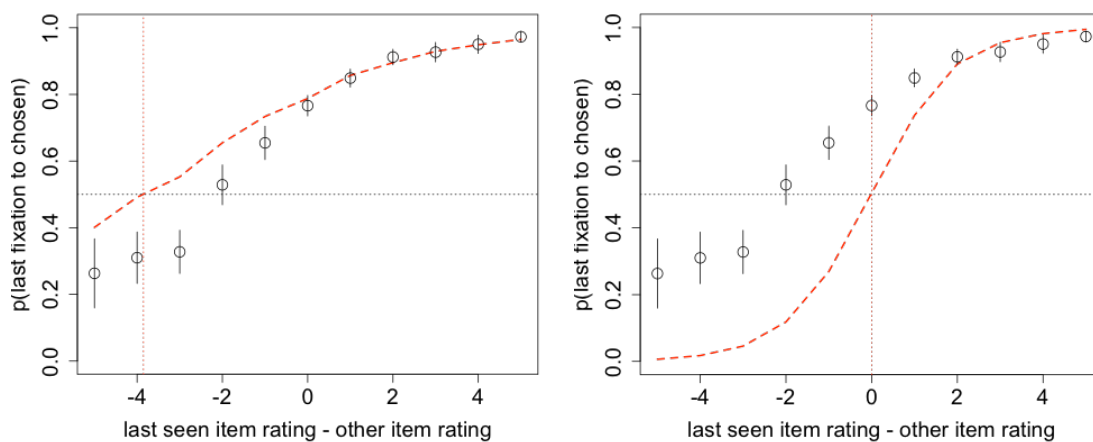
**Figure S2:** Reproduction of Fig. 2B from the text but with the simulation data from  $\theta = 0$  (left) and  $\theta = 1$  (right). The goodness-of-fit statistics for  $\theta = 0$  and  $\theta = 1$  were  $p = .01$  and  $p = 0.0007$ , respectively.



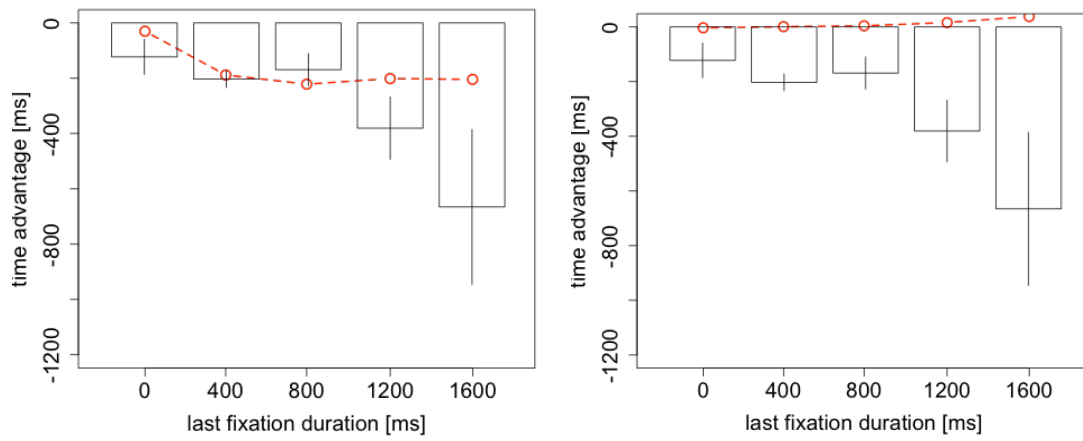
**Figure S3:** Reproduction of Fig. 2C from the text but with the simulation data from  $\theta = 0$  (left) and  $\theta = 1$  (right). The goodness-of-fit statistics for  $\theta = 0$  and  $\theta = 1$  were  $p = 10^{-5}$  and  $p = 10^{-15}$ , respectively.



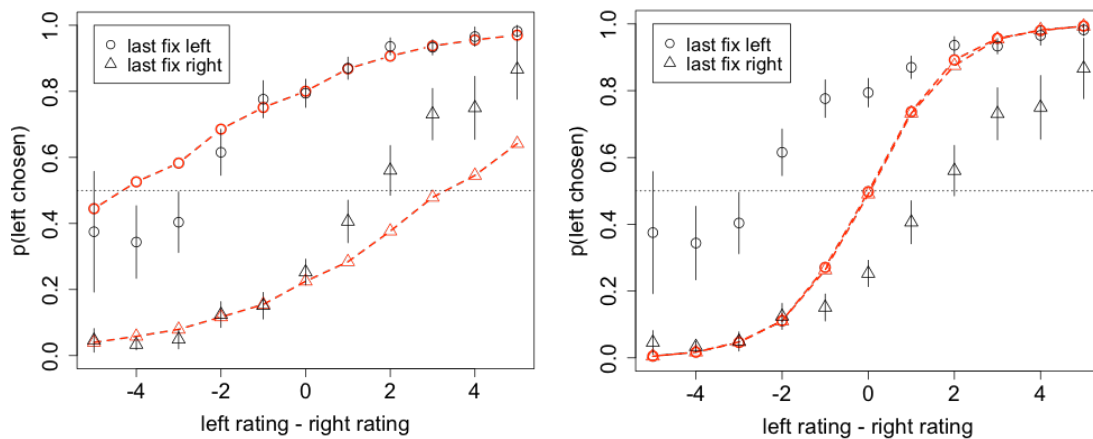
**Figure S4:** Reproduction of Fig. 4B from the text but with the simulation data from  $\theta = 0$  (left) and  $\theta = 1$  (right). The  $\chi^2$  goodness-of-fit statistics for  $\theta = 0$  and  $\theta = 1$  were 22.9 ( $p < .01$ ) and 622 ( $p < 10^{-16}$ ), respectively.



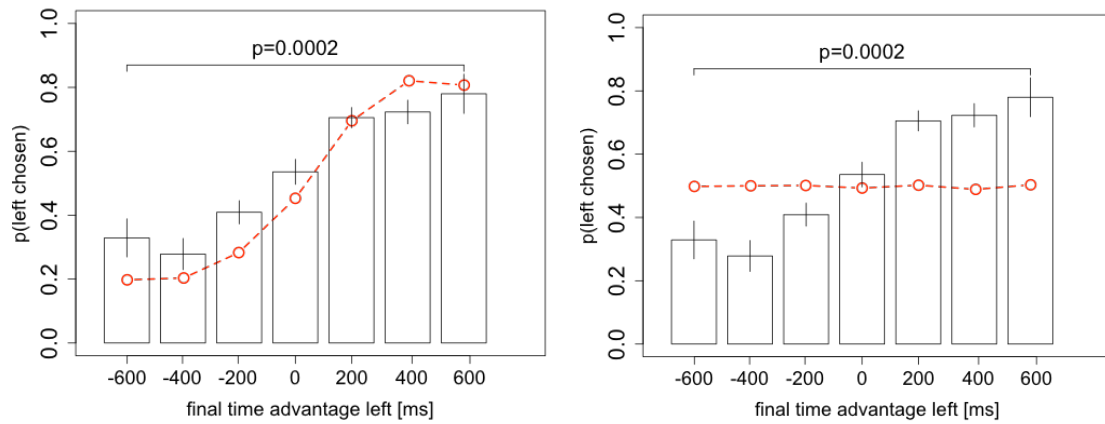
**Figure S5:** Reproduction of Fig. 4C from the text but with the simulation data from  $\theta = 0$  (left) and  $\theta = 1$  (right). The goodness-of-fit statistics for  $\theta = 0$  and  $\theta = 1$  were  $p = 0.96$  and  $p = 0.04$ , respectively.



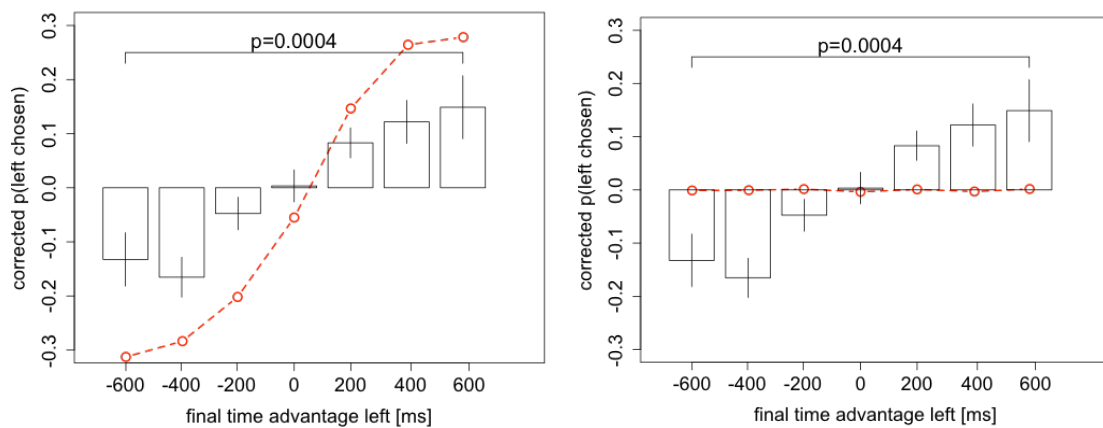
**Figure S6:** Reproduction of Fig. 5A from the text but with the simulation data from  $\theta = 0$  (left) and  $\theta = 1$  (right). The  $\chi^2$  goodness-of-fit statistics for  $\theta = 0$  and  $\theta = 1$  were 5.76 ( $p < .83$ ) and 495 ( $p < 10^{-16}$ ) respectively for last-fixation-left, and 13.6 ( $p < 0.19$ ) and 34.7 ( $p < 0.0001$ ), respectively for last-fixation-right.



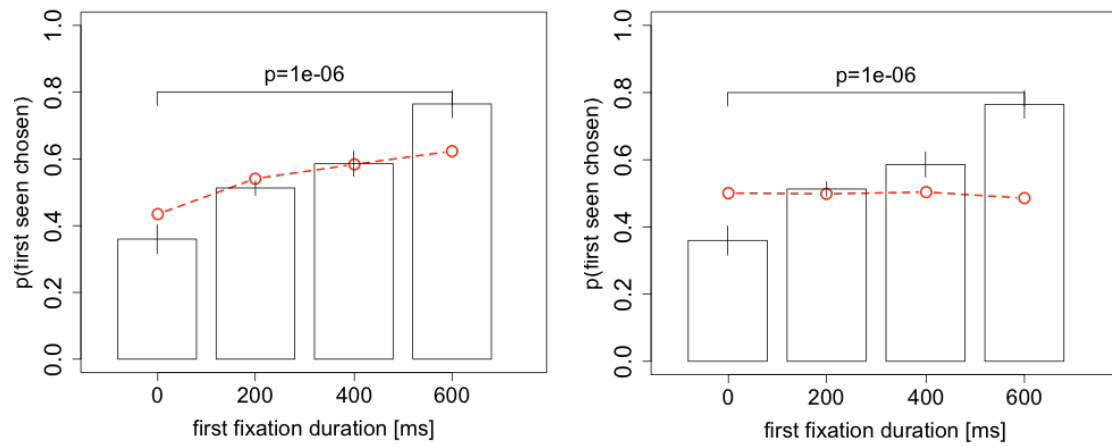
**Figure S7:** Reproduction of Fig. 5B from the text but with the simulation data from  $\theta = 0$  (left) and  $\theta = 1$  (right). The  $\chi^2$  goodness-of-fit statistics for  $\theta = 0$  and  $\theta = 1$  were 25.8 ( $p < 0.0002$ ) and 74 ( $p < 10^{-13}$ ), respectively.



**Figure S8:** Reproduction of Fig. 5C from the text but with the simulation data from  $\theta = 0$  (left) and  $\theta = 1$  (right). The goodness-of-fit statistics for  $\theta = 0$  and  $\theta = 1$  were  $p = 10^{-13}$  and  $p = 10^{-11}$ , respectively.

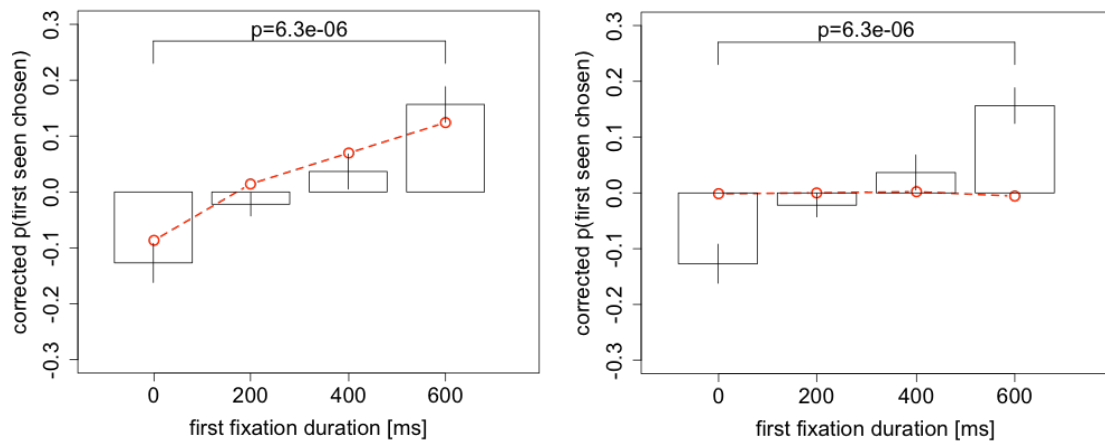


**Figure S9:** Reproduction of Fig. 5D from the text but with the simulation data from  $\theta = 0$  (left) and  $\theta = 1$  (right). The  $\chi^2$  goodness-of-fit statistics for  $\theta = 0$  and  $\theta = 1$  were 1.16 ( $p < 0.76$ ) and 16.5 ( $p < 0.0009$ ), respectively.

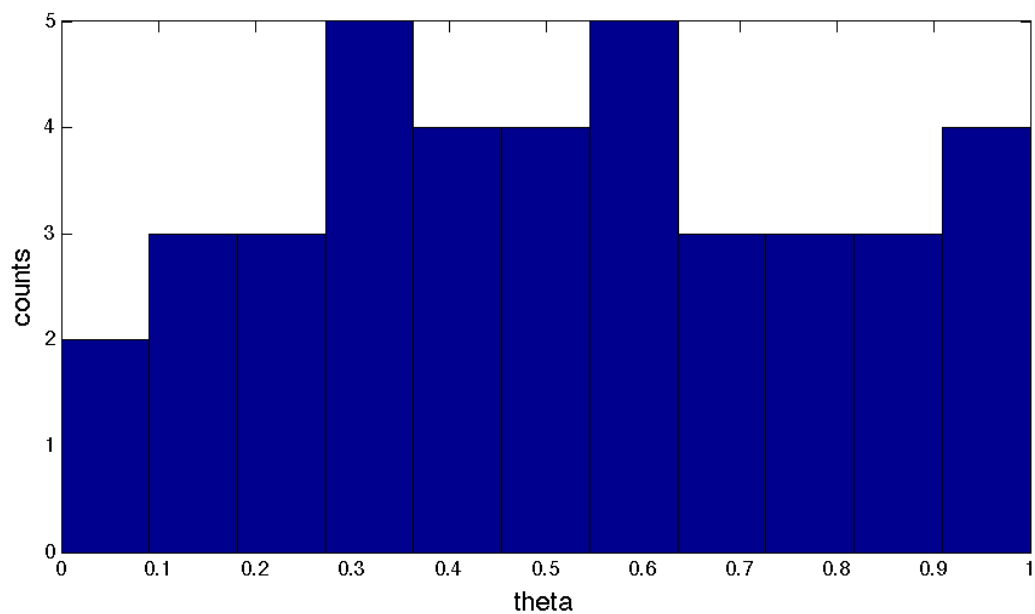




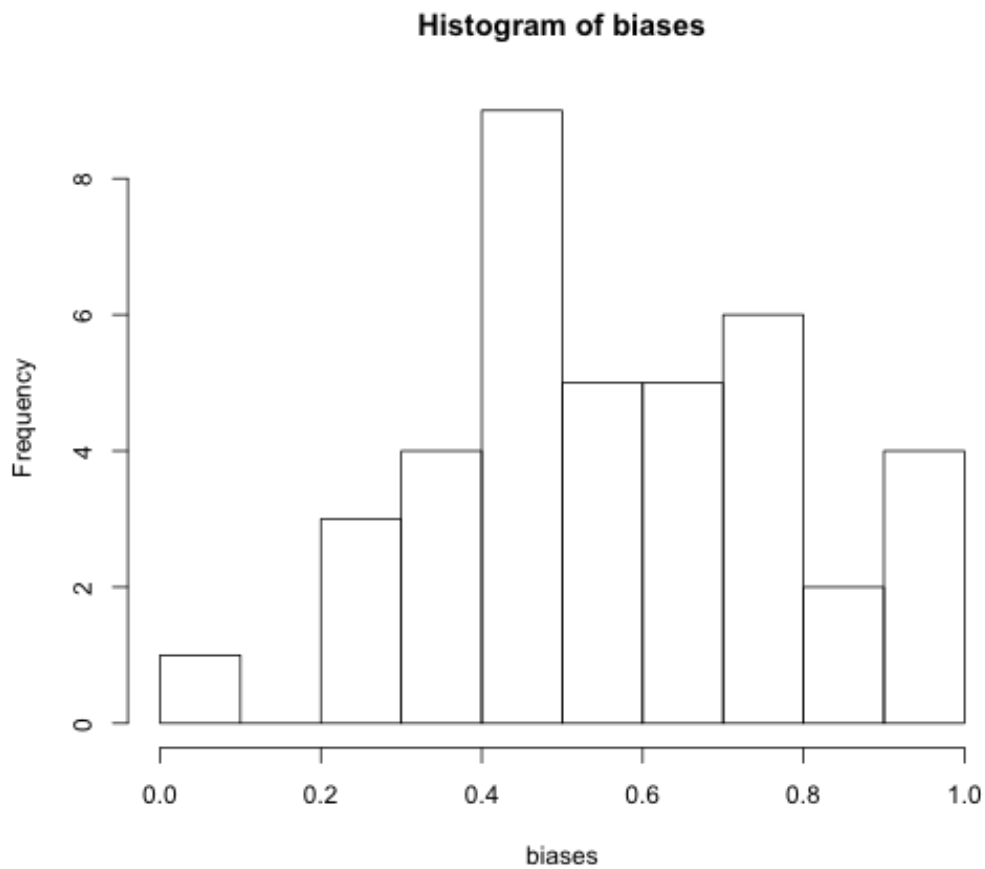
**Figure S10:** Reproduction of Fig. 5E from the text but with the simulation data from  $\theta = 0$  (left) and  $\theta = 1$  (right). The goodness-of-fit statistics for  $\theta = 0$  and  $\theta = 1$  were  $p = 0.1$  and  $p = 10^{-9}$ , respectively.



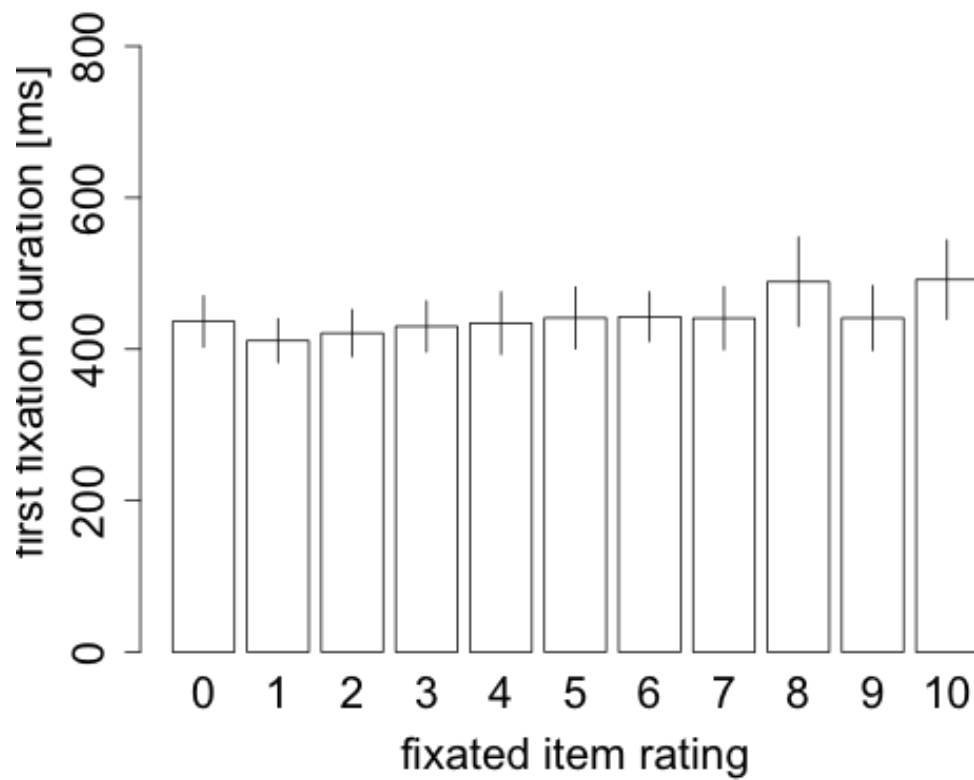
**Figure S11:** Histogram of the best-fitting  $\theta$  parameters based on a subject-by-subject MLE analysis where  $d$  and  $\sigma$  were fixed at their values (0.0002 and 0.02) from the group-level analysis, and we searched for  $\theta$  from 0 to 1, in increments of 0.1.



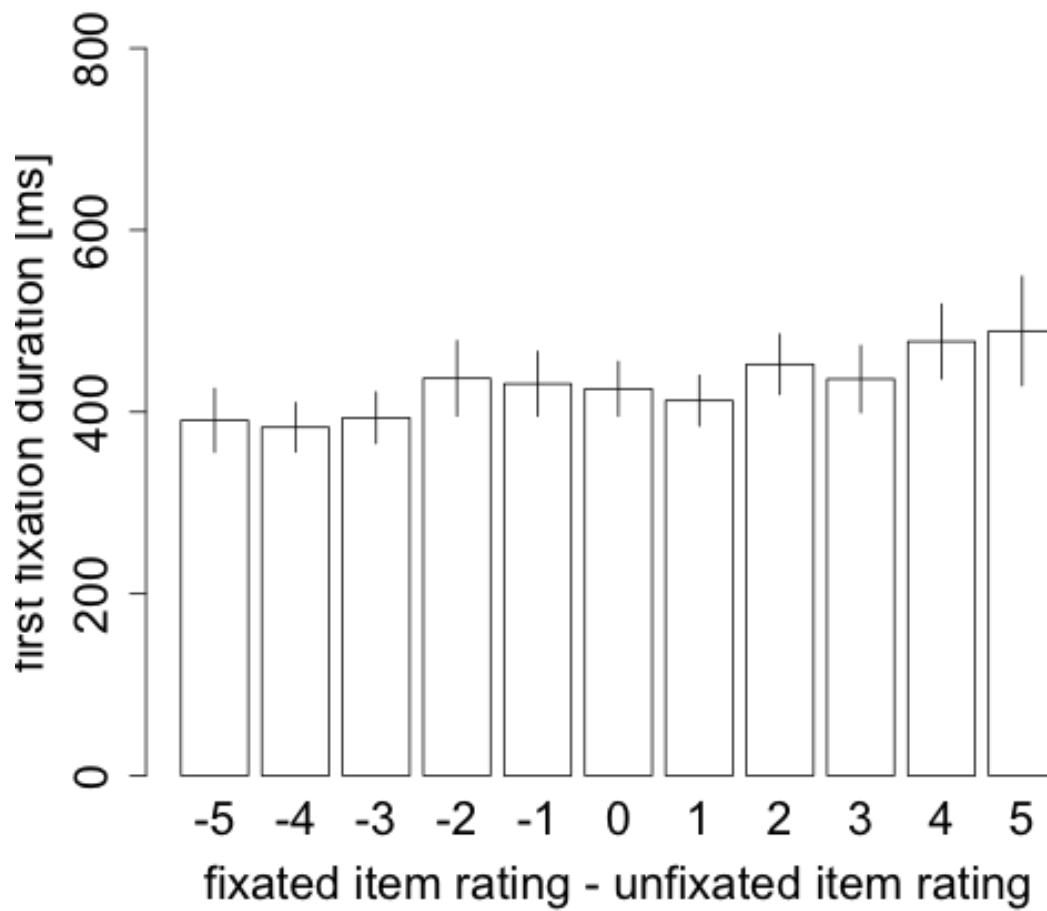
**Figure S12:** Histogram of the left-choice bias between last-fixation-left trials and last-fixation-right trials, subject by subject. This bias measure takes the average difference between the two curves in Fig. 5A. With  $d$  and  $\sigma$  fixed at their values (0.0002 and 0.02) from the group-level analysis, a subject with  $\theta = 1$  would show a bias of 0 (first bin with one subject), a subject with  $\theta = 0.3$  would show a bias of 0.47 (fifth bin with nine subjects), and a subject with  $\theta = 0$  would show a bias of 0.58 (sixth bin with five subjects).



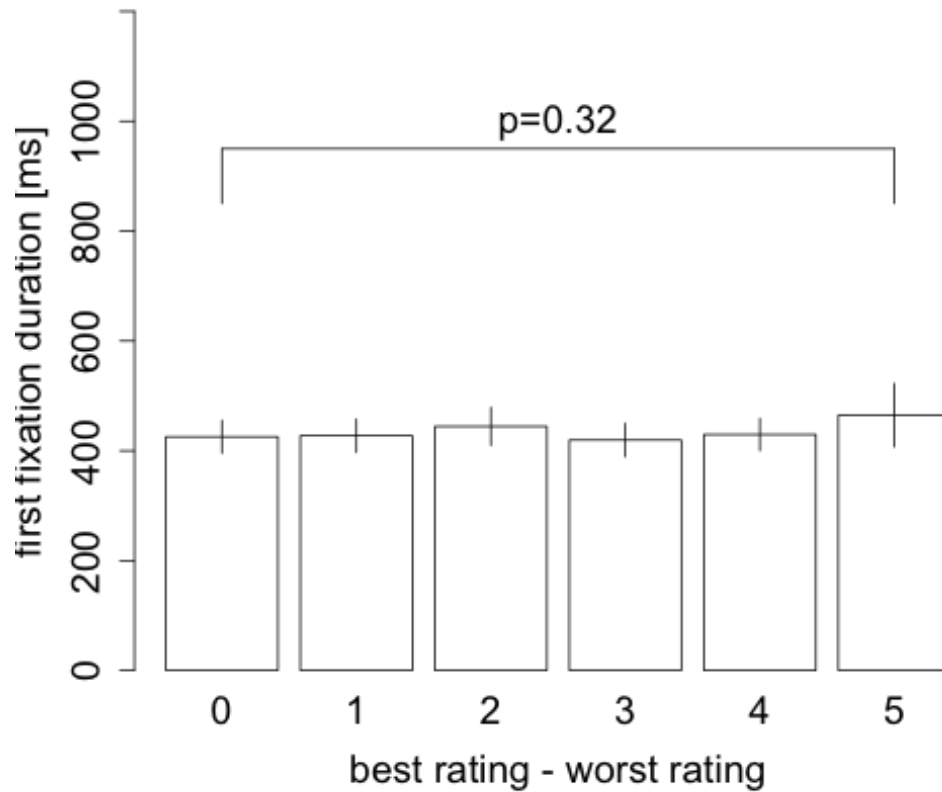
**Figure S13:** First fixation duration as a function of the fixated item's liking rating. A mixed effects regression for fixation duration on liking rating yielded a coefficient of 9.4 ms/rating ( $p < 0.004$ ).



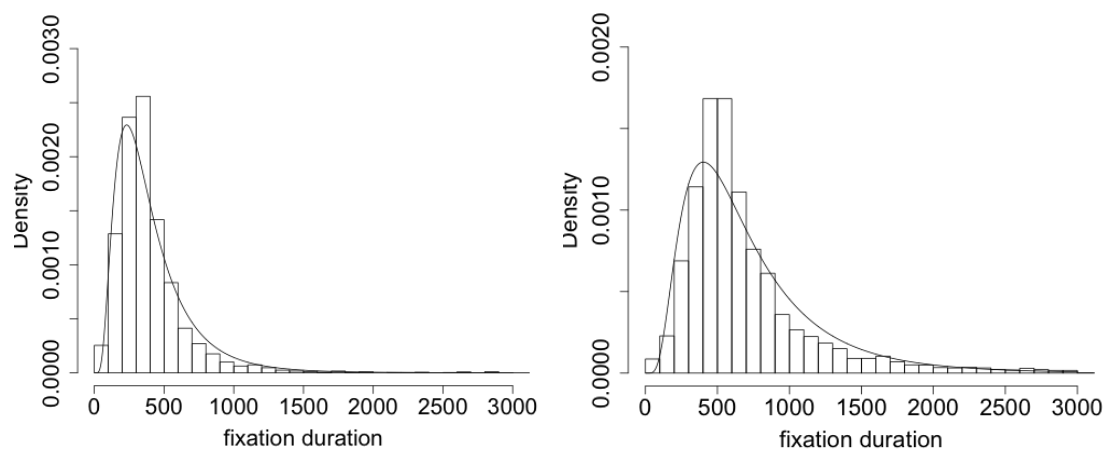
**Figure S14.** First fixation duration as a function of the difference in liking ratings between the fixated item and the unfixated item. A mixed effects regression for fixation duration on the difference in liking ratings yielded a coefficient of 4.3 ms/rating ( $p < 0.1$ ) suggesting no significant effect of relative value on first fixation duration.



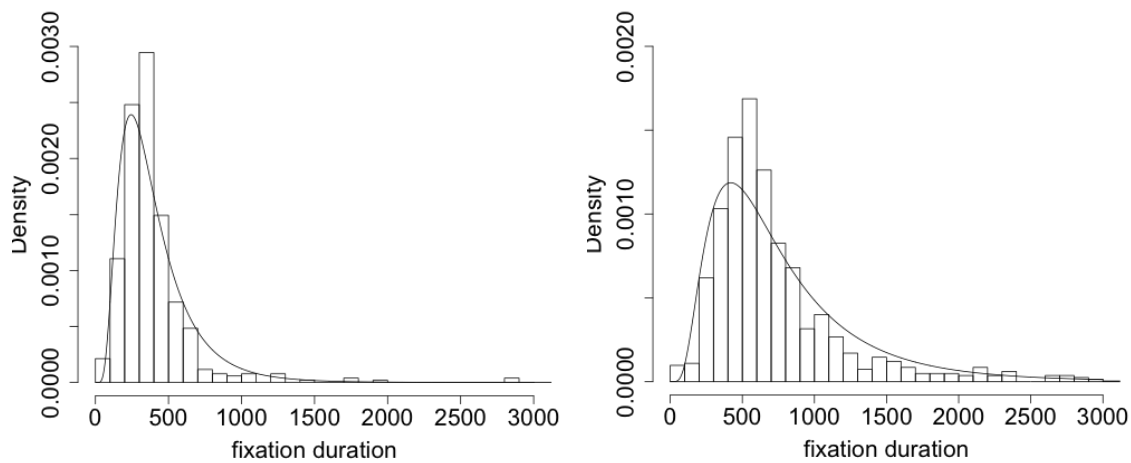
**Figure S15.** First fixation duration as a function of the difference in liking ratings of the best and worst rated items. A mixed-effects regression for fixation duration on the absolute difference in liking ratings yielded a coefficient of  $-0.11$  ms/rating ( $p < 0.98$ ) indicating no significant effect of absolute rating difference on first fixation duration.



**Figure S16:** Fixation duration histograms across all trials with best-fitting log-normal distributions superimposed (solid line) for first fixations (left) and middle fixations (right).

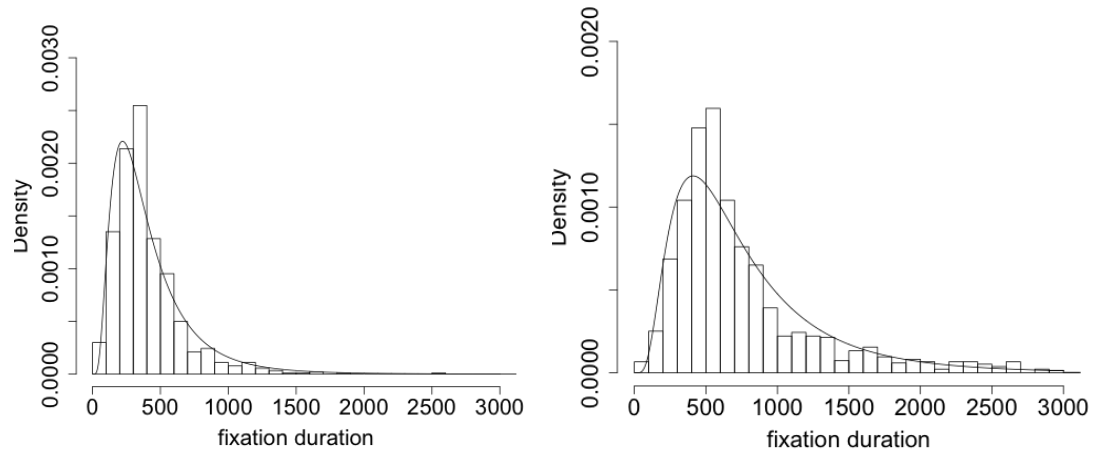


**Figure S17:** Fixation duration histograms for trials with an absolute difference in liking ratings of 0, with best-fitting log-normal distributions superimposed (solid line) for first fixations (left) and middle fixations (right).

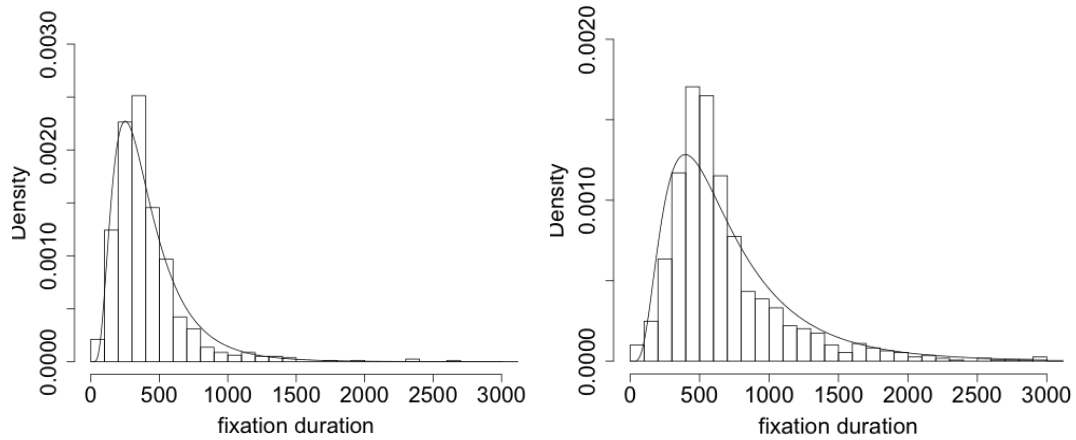




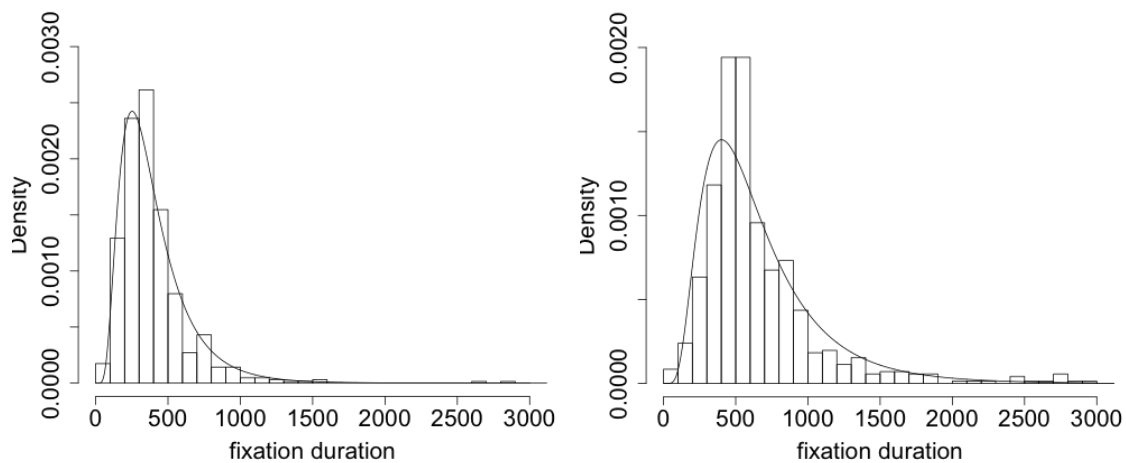
**Figure S18:** Fixation duration histograms for trials with an absolute difference in liking ratings of 1, with best-fitting log-normal distributions superimposed (solid line) for first fixations (left) and middle fixations (right).



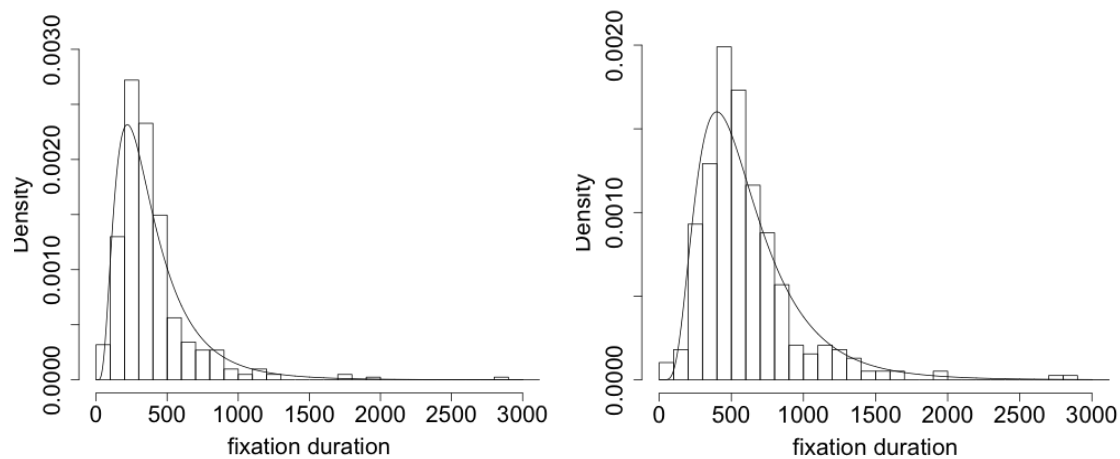
**Figure S19:** Fixation duration histograms for trials with an absolute difference in liking ratings of 2, with best-fitting log-normal distributions superimposed (solid line) for first fixations (left) and middle fixations (right).



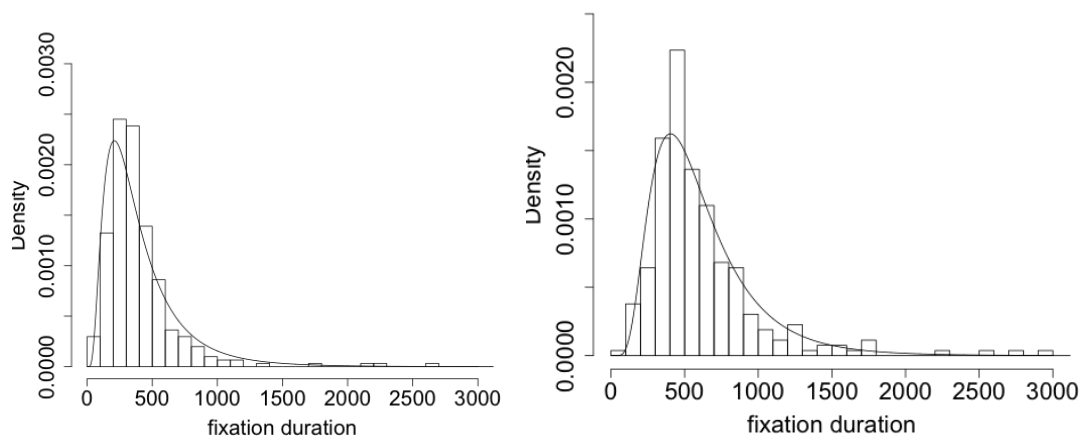
**Figure S20:** Fixation duration histograms for trials with an absolute difference in liking ratings of 3, with best-fitting log-normal distributions superimposed (solid line) for first fixations (left) and middle fixations (right).



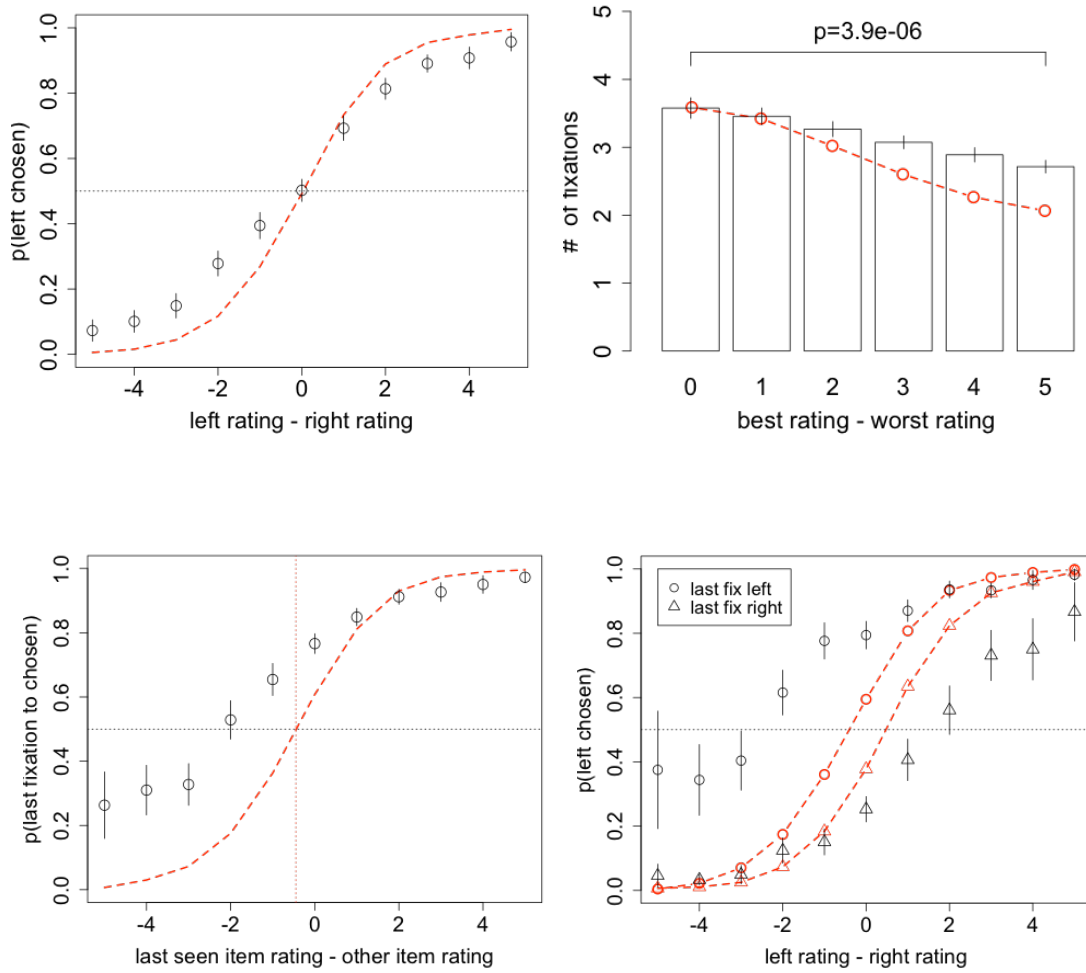
**Figure S21:** Fixation duration histograms for trials with an absolute difference in liking ratings of 4, with best-fitting log-normal distributions superimposed (solid line) for first fixations (left) and middle fixations (right).



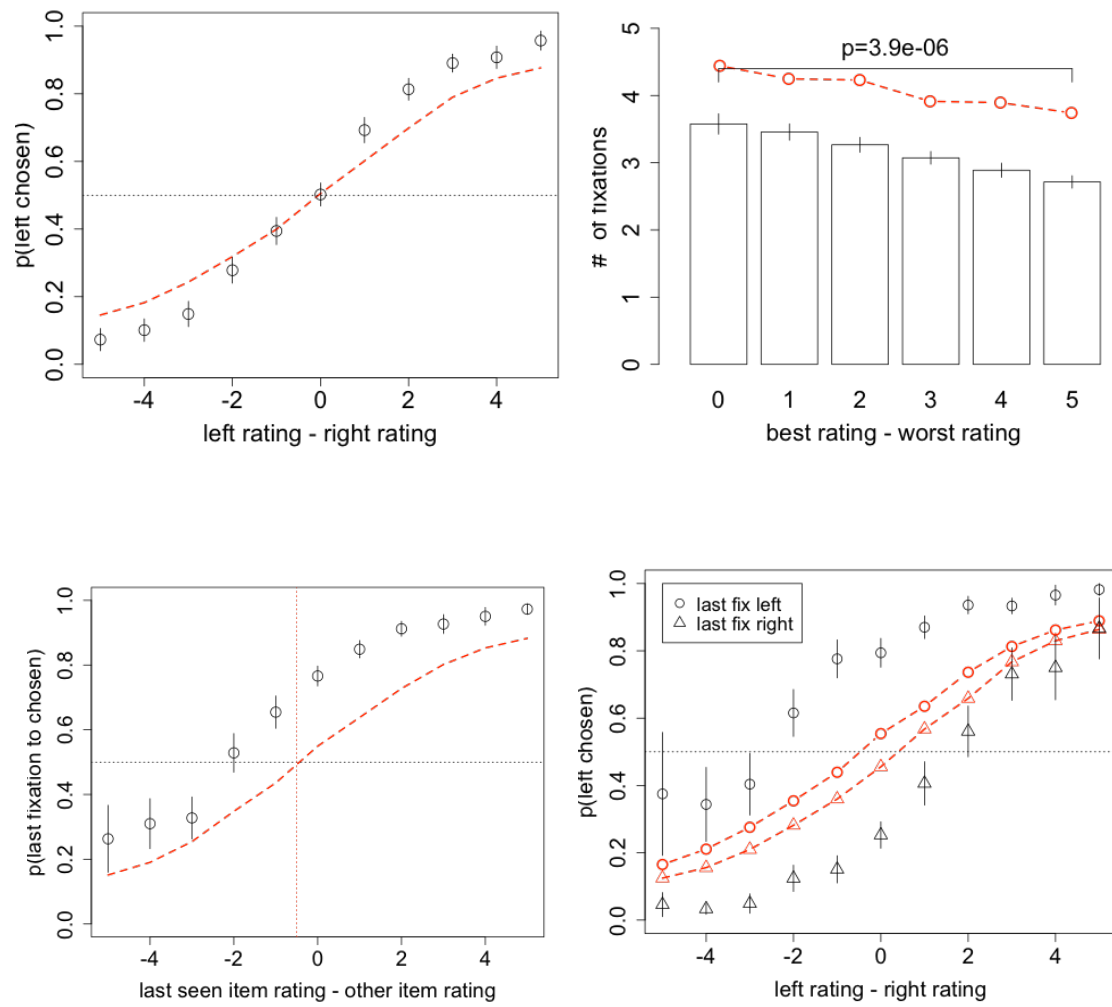
**Figure S22:** Fixation duration histograms for trials with an absolute difference in liking ratings of 5, with best-fitting log-normal distributions superimposed (solid line) for first fixations (left) and middle fixations (right).



**Figure S23:** Replication of Fig. 2A, 2C, 4B, and 5A with the alternative model 1.



**Figure S24:** Replication of Fig. 2A, 2C, 4B, and 5A with the alternative model 2.



**Figure S25:** Replication of Fig. 2A, 2C, 4B, and 5A with the alternative model 3.

