Estimating Information Cost Functions in Models of Rational Inattention

Ambuj Dewan and Nathaniel Neligh*

October 2, 2019

Abstract

Models of costly information acquisition have grown in popularity in economics. However, little is known about what form information costs take in reality. We show that under mild assumptions on costs, including continuity and convexity, gross payoffs to decision makers are non-decreasing and continuous in potential rewards. We conduct experiments involving simple perceptual tasks with fine-grained variation in the level of potential rewards. Most subjects exhibit monotonicity in performance with respect to potential rewards, and evidence on continuity and convexity of costs is mixed. Moreover, subjects' behavior is consistent with a subset of cost functions commonly assumed in the literature.

*Dewan: Neuroscience Institute, NYU School of Medicine, New York University. Neligh: Economic Science Institute, Chapman University. We are grateful to Mark Dean and Navin Kartik for their guidance on this project. We thank Marina Agranov, David Rojo Arjona, Jushan Bai, Ethan Bromberg-Martin, Jisu Cao, Alessandra Casella, Yeon-Koo Che, Giorgio Coricelli, Chad Kendall, Judd Kessler, José Luis Montiel Olea, Pietro Ortoleva, Hashem Pesaran, Andrea Prat, Miikka Rokkanen, Christoph Rothe, Simon Smith, Michael Woodford, two anonymous referees, and seminar and colloquium participants at Columbia University and the University of Southern California for their helpful comments and suggestions. We would also like to thank the numerous colleagues who helped us test our experiments. We acknowledge the financial support of grants from the Columbia Experimental Laboratory for the Social Sciences and the Microeconomic Theory Colloquium at Columbia University. Portions of this research were completed while Dewan was a postdoctoral scholar research associate at USC Dornsife INET, University of Southern California. The authors affirm no financial interest in the outcomes of the experiments detailed herein. All data were collected with the approval of the Columbia University Institutional Review Board.

1 Introduction

It has been observed in many settings that people have a limited capacity for attention, and this has a strong impact on their decision-making. For example, Chetty et al. (2009) demonstrate that consumers underreact to non-salient sales taxes; De los Santos et al. (2012) show that people only visit a small number of websites before making online purchases; and Allcott and Taubinsky (2015) provide evidence that people do not fully account for energy efficiency when making purchasing decisions about light bulbs.[1] Several laboratory experiments have demonstrated evidence of limited attention, including Gabaix et al. (2006), Caplin and Martin (2017), and Dean and Neligh (2019).

An increasingly common explanation for this phenomenon is the theory of *rational inattention* (Sims, 2003, 2006; Caplin and Dean, 2015; Matějka and McKay, 2015). This theory posits that people rationally choose the information to which they attend, trading off the costs of paying more attention with the ensuing benefits of better decisions. This decision-making process occurs in two stages. In the first stage, the decision-maker chooses what information to acquire and pays costs accordingly. In the second stage, the decision-maker uses the information she acquired to make decisions. Some authors make minimal assumptions on first-stage costs and derive the resulting behavioral implications (e.g. Caplin and Dean, 2015; Chambers et al., 2019). Other authors assume a specific functional form for these costs, such as mutual information (Matějka and McKay, 2015; Steiner et al., 2017; Caplin et al., 2019), or channel capacity (Woodford, 2012a). However, little is known about what form these costs take in reality, and different assumptions on these costs can lead to starkly different predictions. For instance, the properties of the information cost function can determine the multiplicity of equilibria in a global game (Hellwig et al., 2012; Morris and Yang, 2019), or whether financial investments are diversified or concentrated (Van Nieuwerburgh and Veldkamp, 2010).

In this paper, we use a laboratory experiment to investigate these first-stage information costs,[2] a crucial input for a rational inattention framework. Subjects complete a series of tasks where they must identify the numerosity of an arrangement of randomly-placed dots, as in Saltzman and

[1] Other empirical studies that find evidence of limited attention include: Hossain and Morgan (2006), Pope (2009), Lacetera et al. (2012), and McDevitt (2014) (consumer choice); Bernard and Thomas (1989), Huberman (2001), DellaVigna and Pollet (2007), Malmendier and Shanthikumar (2007), Hirshleifer et al. (2009), and Ehrmann and Jansen (2017) (financial markets); Bartoš et al. (2016) (housing and labor); and Ho and Imai (2008) and Shue and Luttmer (2009) (voting behavior). For a survey that discusses additional field studies, see DellaVigna (2009).

[2] We also conducted an online experiment, the results of which we report in Supplementary Appendix S5.

Garner (1948) and Kaufman et al. (1949), but without time pressure. Our tasks are also similar to the ball-counting tasks of Caplin and Dean (2014) and Dean and Neligh (2019).

This series of tasks has fine-grained variation in the level of potential rewards for a correct answer. For each reward level, we observe both the correct answer and the subject's response. We interrogate these data in two ways: (1) testing various properties of cost functions; (2) determining which functional forms for information costs are most consistent with observed behavior.[3]

For the first set of analyses, the data allow us to recover several properties of each subject's underlying information cost function. The analyses proceeds hierarchically. For each subject, we are concerned with: (1) whether an information cost function exists, i.e. whether the subject's behavior is consistent with a rational inattention framework; (2) if it exists, whether it produces behavior that is responsive to incentives; and (3) if it induces responsiveness, whether it is "well-behaved," i.e. continuous and convex. In the paper, we derive conditions on subjects' data for testing each of these properties.

The reason we are interested in well-behavedness is because continuity and convexity are important characteristics of many cost functions and are often assumed in economic analysis. The convexity[4] of an information cost function can greatly affect model predictions. For example, Van Nieuwerburgh and Veldkamp (2010) study a portfolio choice problem in which investors choose which assets to learn about and how much to learn about each of them. Depending on the convexity of the investor's utility and cost functions, it can be optimal for the investor to learn about all available assets or to simply concentrate their attention on a single asset; utility and cost functions that imply concave objective functions result in generalized learning, whereas those that imply convex objective functions result in specialized learning. Convexity also has implications for comparative statics in a model of rational inattention. As we prove in this paper, continuity and convexity together imply that gross payoffs (excluding information costs) change continuously in incentives. This has crucial implications for economic analysis: an elasticity-based approach to welfare analysis, which is based around local properties of agents' behavior, can be deeply misleading if there are discontinuities in that behavior. As an illustration of the importance of these properties, we

[3]Using fine-grained variation in incentives provides several advantages for the second set of analyses. For details, please refer to Appendix subsection A4.2.

[4]Note that a finite, convex function is continuous on the interior of the space on which it is defined; in most cases of interest, continuity will be a necessary condition for convexity.

show in Supplementary Appendix S6 that the violation of them can have profound implications in a simple contracting model.

In our sample, over 85% of subjects exhibit behavior consistent with having an information cost function, but just over half exhibit responsiveness. We also find that roughly one-third of responsive subjects (those whose performance on the tasks improves with increasing potential rewards) have behavior that is consistent with well-behaved cost functions.

The second important set of analyses in our paper fits various classes of cost functions to our subjects' data and selects the best fit for each subject. From the accuracy of subjects' responses for each reward level, we infer how their performance in the experimental tasks changes with potential rewards; put differently, we estimate a *performance function* that traces out the relationship between the potential reward and the probability of success. From this relationship, we can recover estimates of the parameters of a subject's information cost function. Comparing the subjects' performance functions to those predicted by a range of information cost functions allows us to find the best fit for each individual.

In this paper, we provide a general result for recovering well-behaved, differentiable information cost functions from performance functions, and we derive functional forms for the performance functions associated with a range of information cost functions, some well-behaved and some not. Of particular interest to us are cost functions that have previously been used in the economic literature: mutual information (cf. Matějka and McKay, 2015), which is a scaling of the expected reduction in entropy from a decision-maker's prior beliefs to their posterior beliefs; Tsallis entropy costs (cf. Caplin et al., 2019), which generalize mutual information; fixed costs for information acquisition (e.g. Grossman and Stiglitz, 1980; Barlevy and Veronesi, 2000; Hellwig et al., 2012); and costs for increasing the precision of normally distributed signals (e.g. Verrecchia, 1982; Van Nieuwerburgh and Veldkamp, 2010). As we show in the paper: the first implies a logistic performance function; the second implies a sigmoid, inverse-S, or concave performance function; the third implies a binary performance function with two levels of performance; and the fourth implies a concave performance function. Of the set of models we estimate, we find that the data of the subjects who are responsive to incentives are best fit by one of these four models, with roughly two-thirds of subjects best fit by the first two models, a quarter of subjects best fit by the third model, and one-seventh of subjects best fit by the fourth model. Thus, while there is some heterogeneity in the population

with respect to which cost functions best reflect human behavior, the set of potential cost functions that we need to consider can reasonably be reduced to four of the cost functions commonly found in the literature.

The main advantage of using an experiment to characterize information costs is that it allows us to observe many decisions from the same individual, over a small time frame, in an environment where we can control the information available to subjects, thereby giving us a rich data set from which to recover the properties and parameters of each individual's cost functions. This is simply not possible with an administrative data set that contains a small number of decisions made by each individual. The experimental methodology we use is also highly adaptable and can accommodate a wide range of information acquisition tasks that may be of interests to researchers. Thus, our approach can be seen as a "testing bed" for theories of limited attention. Specifically using perceptual tasks with clear correct and incorrect answers, rather than choices between goods or gambles, allows us to estimate information costs separately from gross utility.

Furthermore, to our knowledge, our paper is the first to use an experiment with fine-grained variation in incentives to infer properties of information cost functions. This fine-grained variation is essential for estimating subjects' performance functions, which is crucial for our model-fitting exercise. Although several papers have examined competing hypotheses of dynamic evidence accumulation using perceptual data (e.g. Ratcliff and Smith, 2004; Smith and Krajbich, 2019), and some have used such data to fit a single model of static information acquisition (e.g. Shaw and Shaw, 1977; Pinkovskiy, 2009; Cheremukhin et al., 2015; Dean and Neligh, 2019), ours is the first to run a model selection exercise between a large number of types of cost functions in a static rational inattention framework.[5] Moreover, in contrast to previous experimental work, the tasks in our experiment involve more than two options, which allows us to differentiate between information cost functions that are not readily distinguished from each other under simple binary choice.

Our experiment also provides significant advantages relative to experiments that use choice over gambles to study information costs (e.g. Pinkovskiy, 2009; Cheremukhin et al., 2015). In

[5]Cheremukhin et al. (2015) assume a specific but flexible functional form for information costs and perform two types of model selection exercises. One is based on model fits and is used to select between expected utility and rank-dependent utility. The other is based on parameter estimates and is used to distinguish between additively separable Shannon entropy-based costs (Matějka and McKay, 2015) and a capacity constraint on mutual information (Sims, 2003), since the functional form they consider nests the former and approximately nests the latter. This analysis omits most of the information cost functions considered in our paper.

a perceptual task, the correct answer is objective and known to the experimenter, whereas in a choice between gambles, the "correct answer" is determined by the subject's preferences, which the experimenter does not directly observe and cannot easily be disentangled from the information costs entailed by the subject in reaching a decision. Moreover, because of our incentivization scheme, any subset of perceptual tasks in our experiment can be unambiguously ranked in terms of potential rewards, which cannot in general be done with choices over gambles. Therefore, using perceptual tasks allows us to not only more cleanly estimate information costs but also to generate objective measures of performance and see how they vary with potential rewards, which as we demonstrate in Section 2 of the paper, is crucial for testing whether subjects even are rationally inattentive (i.e. have information costs) in the first place.

The remainder of the paper proceeds as follows. Section 2 presents the theoretical framework that we use in this paper. Section 3 introduces various models of cost functions and applies them to the tasks of our experiment. Section 4 presents our experimental design. Section 5 presents and discusses basic experimental results and categorizes subjects according to the behaviors they exhibit. Section 6 fits various models of cost functions to the subjects' data and runs a model selection exercise to determine which is the best fit for each subject. Section 7 concludes. A more general version of this paper's theoretical framework, most proofs, experimental instructions, and robustness checks are included in appendices that can be found in the online supplement. Additional experimental results and an application to the delegation of investment are presented in unpublished supplementary appendices.[6]

2 Theoretical Framework

2.1 Uniform Guess Tasks

In this section, we present a simplified rational inattention framework that is adapted to the tasks we use in our experiment. A fuller treatment of the theory for a general discrete rational inattention framework can be found in Appendix A1.

Consider a task where there is some unknown true state of the world $\theta \in \Theta$ that a decision-maker (DM) has to identify, and learning about the true state is costly. There are n possible states,

[6]Supplementary appendices can be found at https://sites.google.com/view/ambuj-dewan/research.

each of which is *a priori* equally likely, i.e. $\Pr(\theta) = \frac{1}{n}\ \forall\, \theta \in \Theta$. In other words, the DM's *prior belief* on the state of the world is uniform. The DM receives a reward r for correctly identifying the state and no reward for incorrectly identifying the state. Therefore, the DM's goal is to maximize her probability of correctly identifying the state, which we call her *performance*, net of whatever costs she incurs in gathering information about the true state. We refer to tasks with this setup as *uniform guess tasks*.

The DM's performance is determined by her choice of *information structure*, which lists how likely guessing each state is, given the true state. Denote by $a \in \Theta$ the DM's guess of the state. Formally, an information structure is a collection of conditional probabilities $(q_{i,j}), i, j = 1, \ldots, n$, where $q_{i,j} = \Pr(a = \theta_i | \theta = \theta_j)$.

When the DM makes her guess, she has a belief about the likelihoods of each of the possible states, given by $\Pr(\theta = \theta_i | a = \theta_j)$. Applying Bayes' rule, it can be shown that this is equal to $\frac{q_{i,j}}{\sum_{k=1}^{n} q_{k,j}}$. We refer to this probability distribution of the state of the world conditional on the DM's guess as her *posterior belief*.

The DM's guess is correct when $a = \theta$ and is incorrect when $a \neq \theta$. Therefore, her performance is:

$$P = \frac{1}{n} \sum_{i=1}^{n} q_{i,i} \tag{1}$$

The DM's objective is to choose P maximize:

$$rP - C(P) \tag{2}$$

where $r > 0$ is the reward, $P \in [0, 1]$ is the DM's chosen performance, and the function $C(\cdot)$ is her associated cost. Denote by $P(r)$ the DM's choice of P for a given r, and call the resulting mapping from reward to performance the *performance function*.

An example of a uniform guess task is the type of task we implement in our experiment. In this type of task, which we refer to as the "dots" task, the DM is shown a screen with a random arrangement of dots. Her goal is to determine the number of dots on the screen, which is between 38 and 42, inclusive, with each possible number equally likely. She receives a reward r for correctly guessing the number of dots and no reward otherwise. In our example, information costs

could include the cost of effort exerted in counting dots, cognitive costs incurred in employing an estimation heuristic, or the opportunity cost of time spent trying to determine the number of dots.

2.2 Testing for Rational Inattention

In order to be able to say anything about the properties of the DM's information cost function, one must first determine whether such a function even exists. As Caplin and Dean (2015) demonstrate in their Theorem 1, observed behavior is consistent with a rational inattention framework with additively separable costs if and only if it satisfies their "no improving attention cycles" (NIAC) and "no improving action switches" (NIAS) conditions. Roughly speaking, NIAC ensures that attention is allocated efficiently, and NIAS ensures that guesses of the state are made optimally, given the information that the DM has obtained. We refer the reader to Caplin and Dean (2015) for formal definitions of these conditions, though the equivalent conditions we present in the propositions of this subsection will suffice for understanding the present paper.

In uniform guess tasks, the efficient allocation of attention can be thought of as paying more attention when it is more valuable to do so, i.e. when the rewards are higher. This is formalized in the following proposition.

Proposition 1. *The DM's behavior is consistent with NIAC iff $P(r)$ is non-decreasing in r.*

What NIAC rules out is negative responses to increased incentives, e.g. by being stressed out by higher stakes.

In a uniform guess task, making optimal guesses means that a correct guess is more likely than any individual incorrect guess. In other words, the DM cannot perform better by switching her guesses. This is formalized in the following proposition.

Proposition 2. *The DM's behavior is consistent with NIAS iff $\forall\, r$, $\forall\, x \in \Theta$, and $\forall\, y \in \Theta$,* $\Pr(\theta = x | a = x) \geq \Pr(\theta = y | a = x)$.

Put differently, NIAS is satisifed in uniform guess tasks if and only if the DM's empirical posterior beliefs are maximized at the guessed state. What this rules out is the systematic misuse of information, e.g. by mentally exchanging two states of the world.

Taking these results together, a DM completing a set of uniform guess tasks is rationally inattentive iff the conditions of Propositions 1 and 2 are satisfied.

2.3 Responsiveness

A set of behaviors that is trivially consistent with rational inattention is one where the DM selects the same performance level for each reward. This is consistent with frameworks such as signal detection theory, where the DM's information structure is exogenously given. More interesting are cases where the DM does modify her behavior in response to changes in the level of incentives.

Definition 1. A DM is *responsive (to incentives)* in a uniform guess task if for some $r_2 > r_1$, $P(r_2) > P(r_1)$.

Put differently, a DM is responsive to incentives if $P(r)$ exhibits an observable region of strict increase.

2.4 Continuity and Convexity

Continuity and convexity are assumptions made on costs in much of economic analysis. In a rational inattention framework, continuity of the cost function implies that gathering a small amount of additional information increases the total cost of information by only a small amount, and convexity implies that the marginal cost of information is increasing; the more information is acquired, the harder it is to acquire additional information. These properties have testable implications for the DM's behavior. Denote by $P^*(r)$ the DM's optimal choice of performance for each r.

Definition 2. $C(\cdot)$ is *well-behaved* if it is continuous and convex on $[0, 1]$, is strictly increasing and strictly convex on $\left(\frac{1}{n}, 1\right)$, and has a global minimum at $\frac{1}{n}$.

Proposition 3. *If $C(\cdot)$ is well-behaved, then $P^*(r)$ is continuous.*

Therefore, assuming the DM is utility-maximizing, one can reject the well-behavedness of $C(\cdot)$ if it is observed that her performance function $P(r)$ is discontinuous.

3 Cost Functions

The space of admissible cost functions is vast. Indeed, any cost function $C : [0, 1] \longrightarrow \bar{\mathbb{R}}$ leads to behavior consistent with NIAS and NIAC. In this subsection, we introduce the classes of cost functions that are most relevant for our analysis and derive their behavioral implications. We

also outline how to recover these cost functions from data that fit the corresponding performance functions.

3.1 A General Recovery Result for Differentiable Cost Functions

If C is differentiable, then the DM's problem can be solved by appealing to the calculus. Not only does this allow one to solve for the DM's optimal performance function, but it also allows an analyst who does not observe C to recover it from observed behavior.

Proposition 4. *Suppose that C is well-behaved and differentiable. Then $P^*(r) = (C')^{-1}(r)$ for all r such that $C'\left(\frac{1}{n}\right) < r < \lim_{x\uparrow 1} C'(x)$. Moreover, $P^*(r) = 1$ for $r \geq \lim_{x\uparrow 1} C'(x)$.*

This follows from taking the first-order condition in the maximization of (2) and the fact that performance cannot exceed 1. Therefore, assuming the DM is utility-maximizing, her cost function can be recovered by inverting and integrating her observed performance function, provided that her performance is strictly increasing and continuous in incentives.

For example, suppose that that C is quadratic:

$$C(P) = \begin{cases} 0, & P \leq d \\ c(P-d)^2, & P > d \end{cases} \tag{3}$$

where $\frac{1}{n} \leq d < 1$. d represents the amount of information that is freely available to the DM, and c regulates the marginal cost of information. Applying Proposition 4, we have:

$$P^*(r) = \begin{cases} \frac{r}{2c} + d, & r \leq 2c(1-d) \\ 1, & r > 2c(1-d) \end{cases} \tag{4}$$

This performance function is affine (where performance would not exceed 1), and it is depicted in Figure 1 along with the corresponding cost function.. Note that (3) can be recovered from (4) by inverting and integrating the non-constant segment. This procedure is general, but it does not always yield a closed form for the recovered C. For instance, if $P^*(r)$ is a polynomial of degree 5 or higher (where it is non-constant), then a general algebraic closed form does not exist for C.

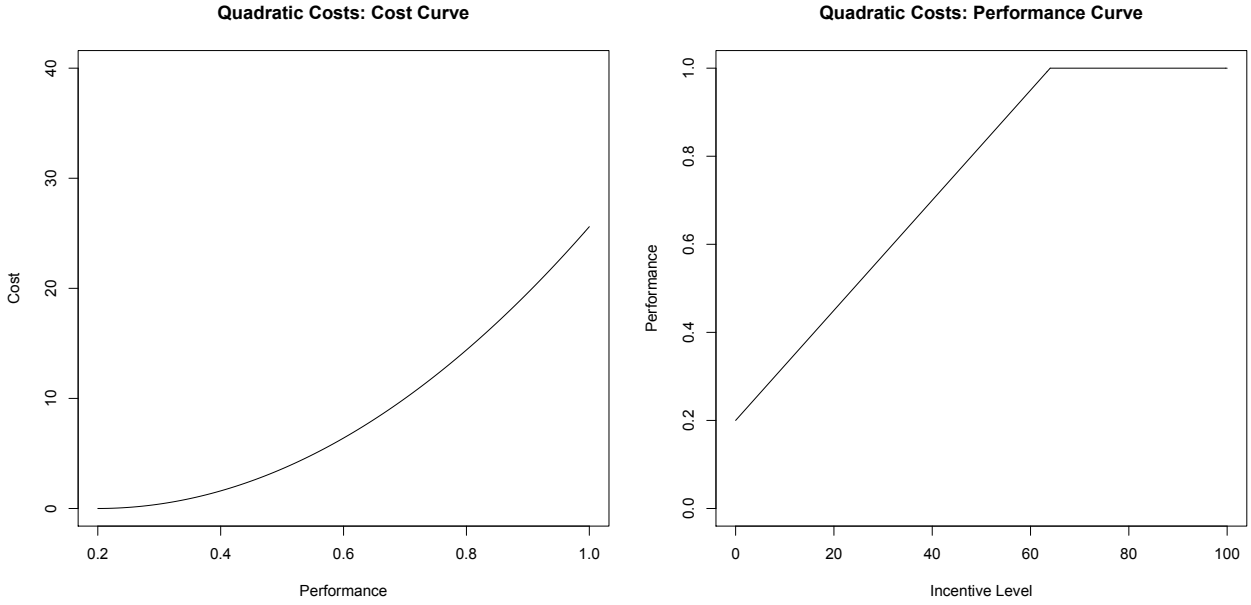


Figure 1: Quadratic costs. The left panel shows the cost function for $c = 40$ and $d = 0.2$, and the right panel shows the resulting performance curves.

3.2 Entropy-Based Cost Functions

One way of modeling the cost of information is to measure how much uncertainty or "randomness" the DM reduces when she acquires information. Learning is effortful, and greater reductions of uncertainty require greater effort. The reduction of uncertainty is usually measured as a difference between her *prior* uncertainty and her *posterior* uncertainty. Formally, let $H : \Delta^{n-1} \longrightarrow \mathbb{R}_{\geq 0}$ be concave. Denoting her belief by p and her information structure by q, this difference is:

$$H(p) - \mathbb{E}[H(p|q)] \tag{5}$$

In general rational inattention problems, this form of cost for the information structure q is called *posterior-separable* (Gentzkow and Kamenica, 2014; Caplin et al., 2019). In uniform guess tasks, we can write (5) as:

$$H\left(\frac{1}{n}, \ldots, \frac{1}{n}\right) - \frac{1}{n}\sum_{j=1}^{n}\left[\sum_{i=1}^{n} q_{i,j}\right] H\left(\frac{q_{1,j}}{\sum_{k=1}^{n} q_{k,j}}, \ldots, \frac{q_{n,j}}{\sum_{k=1}^{n} q_{k,j}}\right) \tag{6}$$

Of course, the choice of H is important; different H functions measure "randomness" in different

ways. One popular choice is *Shannon entropy* (Shannon, 1948). It is defined as follows:

$$H^S(p) := -\alpha \sum_{i=1}^{n} p_i \ln(p_i) \tag{7}$$

where α is a strictly positive constant. The posterior-separable cost function that uses Shannon entropy is known as *mutual information*, and it has been widely studied in the literature (Matějka and McKay, 2015; Caplin et al., 2019). In uniform guess tasks, by substituting (7) into (6), mutual information can be written as:

$$\alpha \left[\ln(n) + \frac{1}{n} \sum_{j=1}^{n} \sum_{i=1}^{n} q_{i,j} \ln \left(\frac{q_{i,j}}{\sum_{k=1}^{n} q_{k,j}} \right) \right] \tag{8}$$

There exist several generalizations of Shannon entropy. The one we study here is known as *Tsallis entropy* (Tsallis, 1988), and it is defined as follows:

$$H^T(p) := \frac{\alpha}{\sigma - 1} \left(1 - \sum_{i=1}^{n} p_i^{\sigma} \right) \tag{9}$$

for $\sigma \neq 1$, where α and σ are strictly positive constants. It can be shown that $H^T(p)$ converges pointwise to $H^S(p)$ as $\sigma \to 1$. Thus, Tsallis entropy generalizes Shannon entropy. In uniform guess tasks, by substituting (9) into (6), the posterior-separable cost function that uses it can be written as:

$$\frac{\alpha}{\sigma - 1} \left[1 - n^{1-\sigma} - \frac{1}{n} \sum_{j=1}^{n} \left(\sum_{i=1}^{n} q_{i,j} \right) \left(1 - \sum_{i=1}^{n} \left(\frac{q_{i,j}}{\sum_{k=1}^{n} q_{k,j}} \right)^{\sigma} \right) \right] \tag{10}$$

for $\sigma \neq 1$ and as (8) for $\sigma = 1$.

It can be shown that in the case of uniform guess tasks and Tsallis entropy costs, it is optimal for $q_{i,i}$ to be the same for all i and $q_{i,j}$ to be the same for all $i \neq j$.[7] Therefore, we can rewrite (8)

[7] For a formal statement and proof of this result, see Lemma A2 in Appendix Subsection A2.5. This implication does not in general hold empirically; for details, see Appendix S1.

and (10) in terms of performance as:

$$C(P) = \begin{cases} \frac{\alpha}{\sigma - 1}\left[P^{\sigma} + (n-1)^{1-\sigma}(1-P)^{\sigma} - n^{1-\sigma}\right], & \sigma \neq 1 \\ \alpha\left[\ln(n) + P\ln(P) + (1-P)\ln\left(\frac{1-P}{n-1}\right)\right], & \sigma = 1 \end{cases} \tag{11}$$

where $\sigma = 1$ is the special case of Shannon entropy costs. With this cost function, we can obtain the optimal performance function that solves (2).

Proposition 5. *When $n \geq 3$, the performance function associated with Tsallis costs can be characterized as follows:*

- *For $\sigma \neq 1$, the performance function is $P^*(r) = \min\{\tilde{P}, 1\}$, where $\tilde{P}$ is a non-negative solution to $r + \frac{\alpha\sigma}{\sigma - 1}\left[\left(\frac{1-\tilde{P}}{n-1}\right)^{\sigma-1} - \tilde{P}^{\sigma-1}\right] = 0$.*
- *For $\sigma = 1$, $P^*(r) = \frac{\exp\left(\frac{r}{\alpha}\right)}{n - 1 + \exp\left(\frac{r}{\alpha}\right)}$.*

Table 1: Properties of Tsallis performance functions for different values of σ

$\boldsymbol{\sigma}$	**Shape**	$\boldsymbol{= 1}$ **for** $\boldsymbol{r \geq \frac{\alpha\sigma}{\sigma - 1}}$ **?**
$(0, 1)$	Sigmoidal	No
1	Logistic	No
$(1, 2)$	Sigmoidal	Yes
2	Affine	Yes
$(2, 3)$	Inverse-S	Yes
3	Square root	Yes
> 3	Concave	Yes

The shapes of these cost functions and the corresponding performance functions are displayed in Figure 2, and properties of the performance functions are listed in Table 1. The σ parameter allows for much flexibility in the performance function, with sigmoidal curves for low σ (between 0 and 2) and concave curves for high σ (greater than 3). For sufficiently high σ (greater than 1), perfect performance is attained for a high enough reward, $\frac{\alpha\sigma}{\sigma - 1}$.

3.3 Normal Signals

Some authors, such as Verrecchia (1982) and Van Nieuwerburgh and Veldkamp (2010), have assumed that the DM receives normally distributed signals about the underlying state of the world,

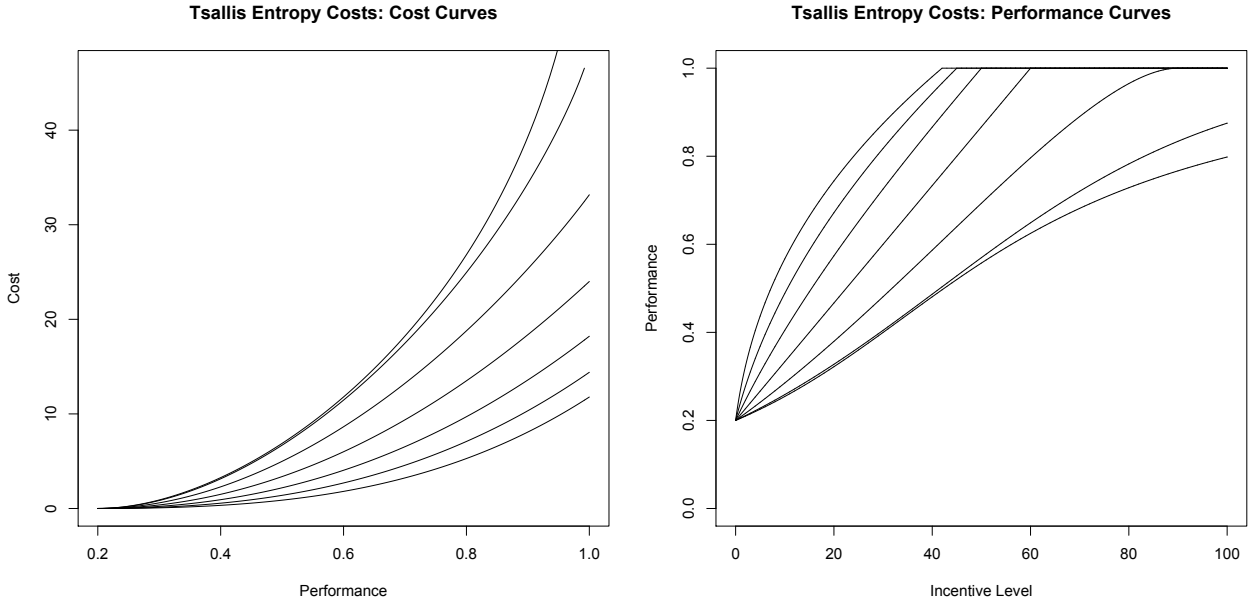


Figure 2: Tsallis entropy costs. The left panel shows the cost function for $\sigma \in \{0.5, 1, 1.5, 2, 2.5, 3, 3.5\}$, going clockwise, and the right panel shows the resulting performance curves for those values of σ, going counterclockwise. α is set at 30.

which she uses to update her prior beliefs, and she pays a higher cost for a more precise signal. In this subsection, we present such a setup.

Let $\Theta \subset \mathbb{R}$, so that we can order its elements from smallest to largest as $\theta_1 < \theta_2 < \ldots < \theta_n$, and suppose that the DM receives signals $\hat{m} \sim N(\theta, s^2)$ about the state of the world θ. The DM can choose the precision $\zeta^2 := s^{-2}$ of these signals, and she pays a cost $K(\zeta)$ accordingly, where K is increasing, convex, and differentiable.[8]

Suppose that the distance between consecutive states is constant so that $\exists\, \eta$ such that $\theta_i - \theta_{i-1} = 2\eta$ for $i \geq 2$. Then it can be shown that the DM's problem is:

$$\max_{\zeta \in [0,\infty)} \frac{r}{n} \left[2\Phi\left(\zeta\eta\right) + (n-2)\left(2\Phi\left(\zeta\eta\right) - 1\right)\right] - K(\zeta) \tag{12}$$

Each choice of ζ induces a performance $P = \breve{P}(\zeta) := \frac{1}{n}\left[2\Phi\left(\zeta\eta\right) + (n-2)\left(2\Phi\left(\zeta\eta\right) - 1\right)\right]$. This allows us to rewrite (12) in the form of (2) by rewriting the cost of information to be a function of P rather than ζ. Because it can be shown that $\breve{P}(\cdot)$ is one-to-one, this is accomplished by setting

[8]Note that K is defined as a function of the positive square root of the precision. However, for the sake of parsimony, we will refer to it as the "cost of precision."

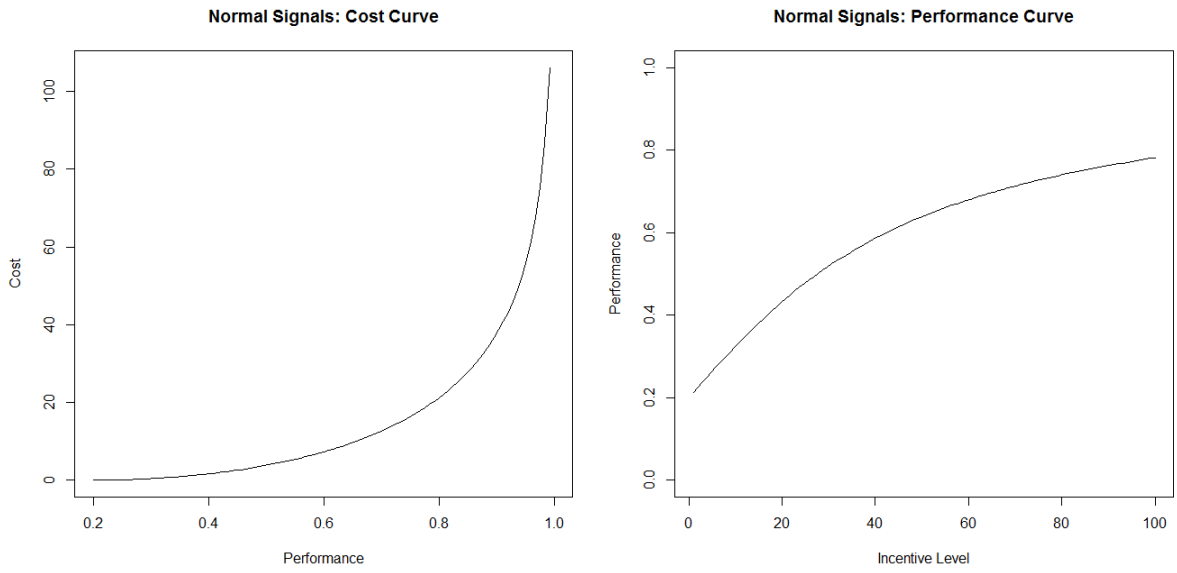


Figure 3: Normal signals with cost of precision given by $K(\zeta) = 4\zeta^2$. The left panel shows the cost function, and the right panel shows the resulting performance curve.

$C(P) = K(\breve{P}^{-1}(P))$. It can also be shown that the resulting $C(\cdot)$ is strictly convex. We then have the following proposition:

Proposition 6. *A DM with normal signals and cost of information $C(\cdot)$ such that $K(\cdot)$ is increasing, is convex, and has non-negative third derivative*[9] *has a strictly concave performance function.*

This type of performance function is depicted in the right-hand panel of Figure 3, for the case of linear K.

3.4 Fixed Costs

Another common model of information costs in the literature is "all-or-nothing" costs, where the DM begins with no information but can become completely informed about the state of the world if she pays a cost (e.g. Grossman and Stiglitz, 1980; Hellwig et al., 2012). Here, we generalize this form of costs by allowing for the DM to receive some information for free and pay a fixed cost to receive more information; we do not stipulate that she must become fully informed.[10]

[9] This assumption on the third derivative is a technical assumption. It holds if, for instance, K is linear in precision (i.e. quadratic in the square root of precision), as we assume later in the paper.

[10] A similar modeling assumption is made by Admati and Pfleiderer (1988) in an asset market model, where a trader can choose either to remain uninformed about asset returns or to receive a noisy signal about returns at a fixed cost.

We can represent this situation as follows. Let there exist $\underline{q}, \bar{q}$ such that $\frac{1}{n} \leq \underline{q} < \bar{q} \leq 1$ and:

$$C(P) = \begin{cases} 0, & P \leq \underline{q} \\ \kappa, & P \in (\underline{q}, \bar{q}] \\ \infty, & P > \bar{q} \end{cases} \tag{13}$$

According to this cost function, the DM can receive information up to an accuracy of $\underline{q}$ for free, but she must pay a fixed cost κ to acquire information up to an accuracy of $\bar{q}$.

Cost functions with fixed costs such as these can be seen as representing dual-system cognitive processes (cf. Stanovich and West, 2000; Kahneman, 2003). In such processes, a small amount of information may be acquired at a very low cost, but there is a fixed cost to acquiring more information. This implies a discontinuity in the cost function between information structures with "low" informativeness and those with "high" informativeness.

In uniform guess tasks, the DM is willing to pay the cost κ of acquiring information only when the rewards are sufficiently high, i.e. when $r\bar{q} - \kappa \geq r\underline{q}$. This implies a binary performance function: for $r \leq \frac{\kappa}{\bar{q}-\underline{q}}$, the DM acquires no information and achieves $\underline{q}$, and for $r > \frac{\kappa}{\bar{q}-\underline{q}}$, the DM acquires enough information to achieve $\bar{q}$.

Cost functions of this subclass are easily recoverable from data by estimating the relationship depicted in the right panel of Figure 4 and finding the incentive level threshold at which the DM's performance level jumps.

3.5 Other Non-convexities

Other non-convex cost functions can also produce discontinuous performance functions. In fact, the cost function need not even be discontinuous for this occur. To illustrate this, consider a DM who has a cost function C that is concave in performance, as depicted in Figure 5.

Net payoffs are maximized when the positive distance between gross payoffs and costs is largest. For low reward levels (such as r_1), this happens at the no-information performance level, 0.2. For high reward levels (such as r_2), this happens at the full-information performance level, 1. In this manner, just as in the fixed-cost case, a binary performance function obtains, with the DM acquiring no information if the incentive is low and acquiring full information if the incentive level is high.

More complicated performance functions are also possible. Consider a richer representation

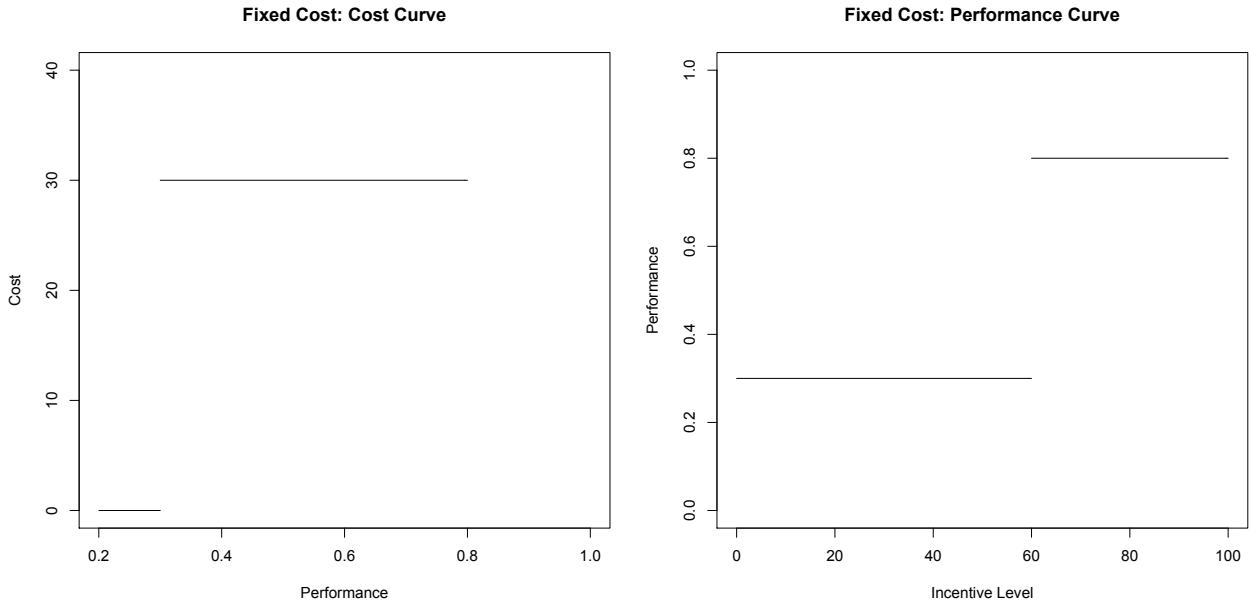


Figure 4: Fixed cost for information acquisition. The left panel shows the cost function, and the right panel shows the resulting performance curve. Parameters are $\kappa = 30$, $\underline{q} = 0.3$, and $\bar{q} = 0.8$.

of a dual-system cognitive process (cf. Kahneman, 2003) than was presented in the preceding subsection:

$$C(P) = \begin{cases} 0, & P \leq d_1 \\ c_1(P - d_1)^2, & d_1 < P \leq d_2 \\ c_1(d_2 - d_1)^2, & d_2 < P \leq d_3 \\ c_2(P - d_3)^2 + c_1(d_2 - d_1)^2, & P > d_3 \end{cases} \tag{14}$$

where $c_1 > 0$, $c_2 > 0$, and $\frac{1}{n} \leq d_1 < d_2 \leq d_3 < 1$. According to this cost function, information is available for free up to a performance level of d_1. In the interval $(d_1, d_2]$, performance can be adjusted up to a level of d_2. This represents the "automatic" system of the dual-system process, and can be thought of as a subconscious process to which the brain can variably allocate mental resources. Exerting effort beyond d_2 can be seen as actively thinking about the problem at hand, or engaging the "controlled" system of the dual-system process. Thinking allows a performance of at least d_3 to be achieved, with higher performance levels attainable with more effort. This "hybrid" cost function that concatenates two quadratic cost curves induces a discontinuous performance

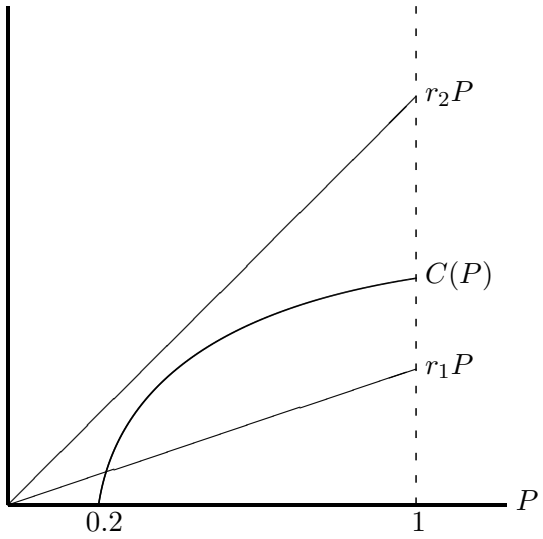


Figure 5: Concave costs

function.

Proposition 7. *Suppose that:*[11]

$$d_3 \in \left(\frac{c_1 d_1 + c_2 - \sqrt{(c_1 d_1 + c_2)^2 + (c_1 + c_2)(c_1(1 - 2d_1 - (d_2 - d_1)^2) - c_2)}}{c_1 + c_2}, \frac{c_1 d_1 + c_2 + \sqrt{(c_1 d_1 + c_2)^2 + (c_1 + c_2)(c_1(1 - 2d_1 - (d_2 - d_1)^2) - c_2)}}{c_1 + c_2} \right)$$

Then the cost function (14) *yields the following performance function:*

$$P^*(r) = \begin{cases} \frac{r}{2c_1} + d_1 & r < \delta \\ \min\left\{ \frac{r}{2c_2} + d_3, 1 \right\}, & r \geq \delta \end{cases} \tag{15}$$

where $\delta = \frac{c_1(d_2 - d_1)^2}{d_3 - d_1}$ *if* $c_1 = c_2$ *and* $\delta = \frac{2c_1 c_2}{c_1 - c_2} \left[\sqrt{(d_3 - d_1)^2 + \frac{(c_1 - c_2)(d_2 - d_1)^2}{c_2}} - (d_3 - d_1) \right]$ *if* $c_1 \neq c_2$.

This performance function consists of two affine segments separated by a jump discontinuity, and it flattens out once the upper bound of perfect performance is reached. It is depicted in Figure 6 along with its corresponding cost function.

[11] This condition on d_3 ensures that the performance function has two separate regions of strict increase rather than just one.

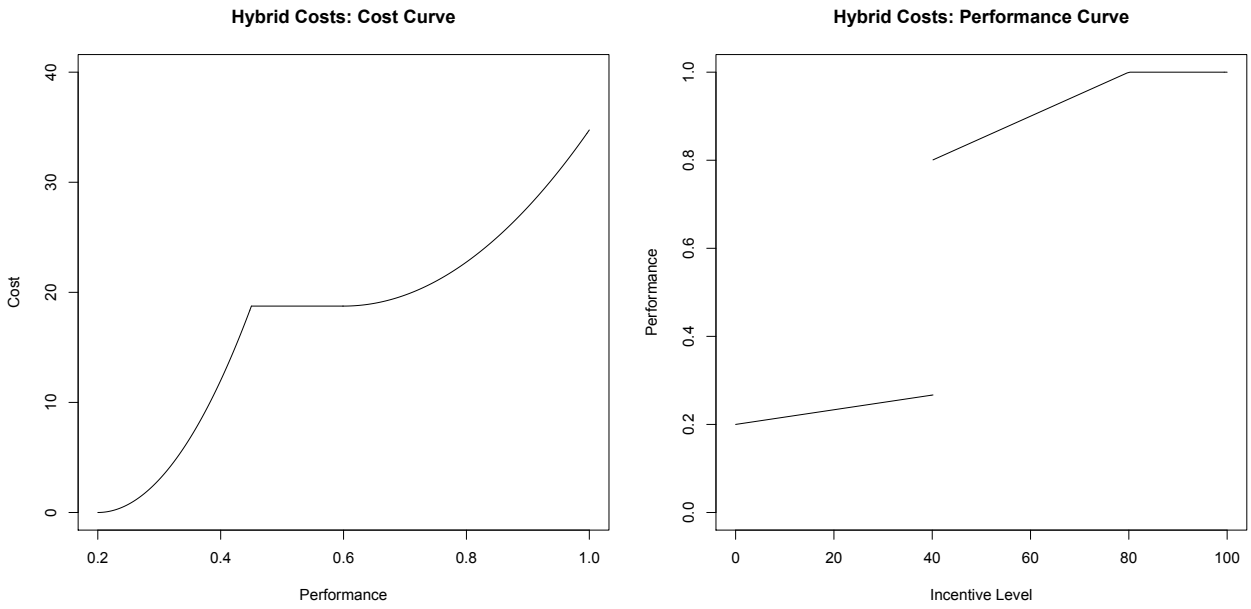


Figure 6: Hybrid costs. The left panel shows the cost function, and the right panel shows the resulting performance curve. Parameters are $c_1 = 300$, $c_2 = 100$, $d_1 = 0.2$, $d_2 = 0.45$, and $d_3 = 0.6$.

3.6 Summary

Table 2 summarizes the properties of the classes of cost functions discussed in this section and lists the corresponding performance functions. While each cost function generates a unique performance function, recovery of a cost function from a performance function is not in general unique, as illustrated in Subsection 3.5. However, a best-fitting performance function can be selected based on a DM's observed behavior, and this datum can be used to determine which of a set of plausible classes of cost functions could have generated the observed behavior. This is the exercise we perform in Section 6.

4 Experimental Design

4.1 Description

The experiment we implemented involved a series of perceptual tasks, each for a potential reward. In each of these tasks, subjects were shown a screen with a random arrangement of dots and were asked to determine the number of dots on the screen. The number of dots was between 38 and

Table 2: Properties of cost functions

Cost Function	**Continuous**	**Convex**	**Performance Function**
Differentiable and well-behaved	Yes	Yes	Inverse of derivative
Quadratic	*Yes*	*Yes*	*Affine*
Tsallis entropy	Yes	Yes	Sigmoid/inverse-S/concave (SIC)
Mutual information	*Yes*	*Yes*	*Logistic*
Normal signals	Yes	Yes	Concave
Dual-process	Can be	No	Discontinuous
Fixed costs	*No*	*No*	*Binary*
Hybrid	*Yes*	*No*	*Piecewise affine*

Note: Performance-function properties of normal-signal costs are for state spaces with equidistant spacing.

42, inclusive, and each number was equally likely.[12] Subjects were informed of these facts; there was no deception or withholding of information about the structure of the tasks. Subjects also completed a second set of tasks involving the identification of angles. We refer to the first type of task as the "dots" task and to the second as the "angle" task. For the sake of brevity, we relegate the description and results of the "angle" task to Supplementary Appendix S4.

Each task had a potential reward in an experimental currency called "points." At the start of each task, subjects were shown this reward, which we refer to as the *incentive level*, in large characters for three seconds (e.g. Figure 7), before it was replaced with the random dot arrangement (e.g. Figure 8). Displaying the incentive level before the dot arrangement ensured that subject looking at the screen would see the incentive level before being able to start the task. While the dot arrangement was on screen, the incentive level continued to be displayed to the right of the screen, ensuring that subjects would not have to memorize this number. Subjects then had as much time as they desired to determine the number of dots on the screen before proceeding to the next task. If they answered correctly, then they earned the potential reward; if not, then they earned no points for that task. Feedback was not given until the end of the experiment.

Subjects completed 200 tasks, each at an integer incentive level between 1 and 100, inclusive. They were randomly shown either all 100 "dots" tasks or all 100 "angle" tasks first. Blocks of tasks were balanced by incentive level to ensure roughly the same level of variation in incentives throughout the experiment. Subjects were first shown each of the 50 odd incentive levels between

[12]These numerosities were selected to be in line with previous experiments with similar tasks (e.g. Caplin and Dean, 2014).

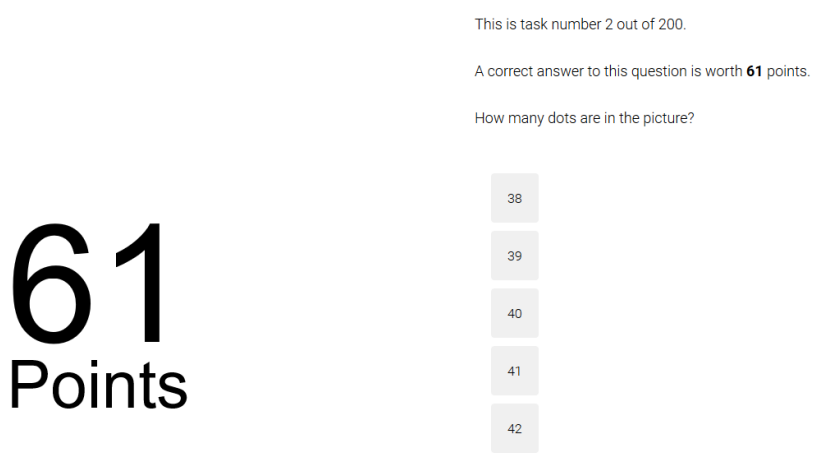


Figure 7: Incentive display for a task

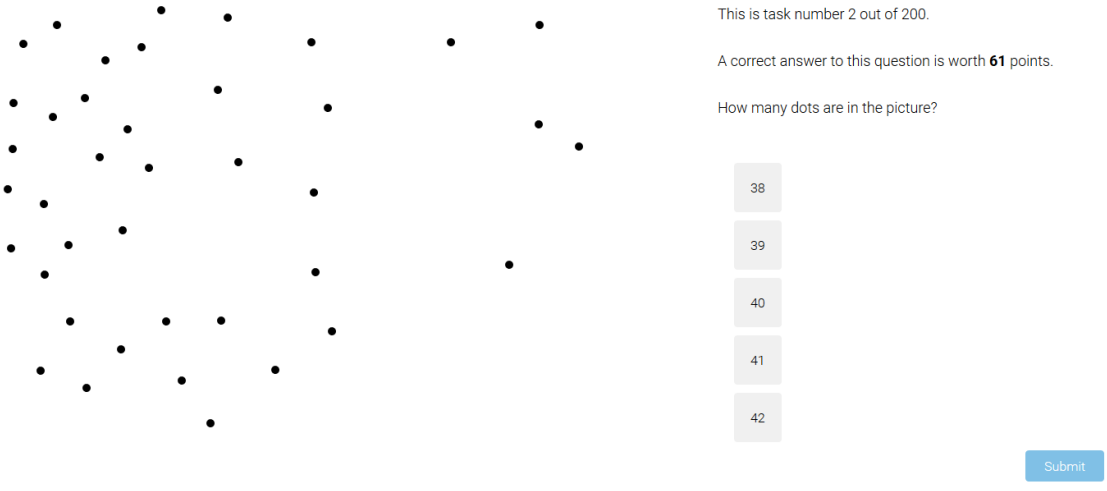


Figure 8: Arrangement of dots for a task

1 and 100 in a random order, and were then shown each of the 50 even incentive levels between 1 and 100 in a random order. This was repeated (in a different random order) for the next 100 tasks.[13]

Experimental earnings were determined as follows. One task from the first half the experiment and one task from the second half of the experiment were randomly selected for payment. The incentive level of each selected task determined the probability of winning one of two monetary prizes. For example, if the first selected task had an incentive level of 84 and was answered correctly, and the second selected task had an incentive level of 33 and was answered incorrectly, then this would give the subject an 84% probability of winning the first prize and a 0% probability of winning the second prize. Determining earnings in this manner ensured that expected earnings were linear in the incentive level, which obviated the need to elicit risk preferences.[14] In other words, this ensured that under the assumption of expected utility theory, the subjects' utilities (excluding information costs) were known to us (up to a multiplicative constant).[15] Thus, the estimated relationship between performance and incentive level for each subject could be considered a valid estimate of their performance function, without the need to apply any additional transformation.

As mentioned above, subjects completed 200 tasks in total: 100 "dots" tasks and 100 "angle" tasks. They either completed all the "dots" tasks or all the "angle" tasks first, and this order was randomly determined.[16] For 41 subjects, the prizes were \$10 US, and for 40 subjects, the prizes were \$20 US. In addition, subjects were paid a \$10 participation fee.

All sessions were conducted at the Columbia Experimental Laboratory in the Social Sciences (CELSS) at Columbia University, using the Qualtrics platform. We ran 8 sessions with a total of 81 subjects, who were recruited via the Online Recruitment System for Economics Experiments (ORSEE) (Greiner, 2015).

[13] A detailed explanation of the advantages and disadvantages of this kind of fine-grained incentive structure can be found in Appendix subsection A4.2. To summarize: while fine-grained incentives present some drawbacks in terms of testing cost function properties, they provide significant benefits when it comes to recovering a subject's cost function from observed behavior.

[14] This binary lottery incentivization technique was pioneered by Roth and Malouf (1979).

[15] We relax the assumption of expected utility theory and allow for incentives to be probability-weighted in Appendix Subsection A4.3. Our qualitative results remain largely unchanged.

[16] In the online version of this experiment, subjects completed 200 "dots" tasks and no "angle" tasks. Results and further details can be found in Supplementary Appendix S5.

4.2 Discussion

Our experimental design has several beneficial features as compared to previous experimental work in limited attention. Firstly, the departure from experiments involving binary choice with two states of the world (e.g. Ratcliff and Smith, 2004; Cheremukhin et al., 2015; Dean and Neligh, 2019) allows for distinguishing between certain types of cost functions in a manner that would otherwise not be possible. For instance, if there were only two states (i.e. two different possible numbers of dots), then entropy-based cost functions would not yield performance curves with inflection points, and it would therefore be difficult to distinguish between Tsallis entropy costs with $\sigma \in (0, 2)$ and normal signal precision costs. Furthermore, having more than two states, some of which were closer to each other than others (e.g. 39 is closer to 38 than to 42) allows us to study perceptual distance, which we discuss further in Supplementary Appendix S1.

Secondly, using perceptual tasks instead of value-based choices, such as choices over gambles (e.g. Pinkovskiy, 2009; Cheremukhin et al., 2015), allows for cleaner identification of information costs. On any given trial, the true state of the world is known to the experimenters; we need not infer it from choice data. Thus, any choice on a subject's part can be cleanly classified as either a correct choice or a mistake, and this classification is then used to estimate information costs. This stands in contrast to experiments with choice over gambles, where it is necessary for the analyst to simultaneously use choice data to estimate utility parameters and use the estimated utility to construct the classification of correct responses and mistakes that is used to estimate information costs.

Moreover, choices over gambles present a couple of conceptual issues in a rational inattention framework. In such a framework, the amount of attention paid to a particular decision problem depends on the marginal benefit of selecting one option over another. However, in choices over gambles, these benefits are not known to the DM *ex ante*, and practically speaking, they cannot be calculated independently of knowing which gamble is preferable to another. By contrast, in perceptual tasks with known rewards, the marginal benefit of answering correctly relative to answering incorrectly is known *ex ante*, regardless of what the correct answer actually is. Furthermore, if utility over gambles (excluding additively separable information costs) deviates from expected utility theory — for instance, with probability weighting — then this can itself be seen as the result of a rationally inattentive process (Woodford, 2012b; Steiner and Stewart, 2016), theoretically but not

necessarily practically distinct from the additively separable information costs that cause choice mistakes, thereby calling into question the descriptive validity of the model.

Finally, using fine-grained variation in incentives allows us to see how subjects' behavior changes in response to small changes in potential rewards. To illustrate, suppose that a subject's utility of money were linear, and they were participating in the \$10 prize treatment. In that case, since an incentive level of 63 gives a 1% higher chance of receiving the prize for answering correctly than an incentive level of 62, the former incentive level is worth 10 cents more to the subject than the latter. Observing how a subject's behavior changes in response to these small differences in incentives allows us to reliably trace out their performance curve and classify that curve according to the information cost function that generated it.

5 Categorization Results

In this section, we present the first set of results of our laboratory experiments. We perform an individual-level analysis to classify subjects according to whether their behavior is consistent with rational inattention, responsiveness to incentives, and well-behavedness of their cost functions. Additional experimental results related to demographics and aggregate behavior are provided in Supplementary Appendix S2.

5.1 Choice Data

The data we are most interested in for each task t are the incentive level r_t, the true state of nature θ_t, and the subject's response a_t. For each task, define the subject's *correctness* as $y_t := \mathbb{1}_{\{\theta_t\}}(a_t)$. That is, y_t takes the value 1 if the subject correctly determined the state of the world in task t and 0 otherwise.

We are primarily interested in the relationship between correctness and incentive level. We can think of the pattern of successes and failures that we observe as being generated by some underlying data-generating process that for every possible reward level tells us the probability of answering correctly. We denote this probability by $P_t := \Pr(y_t = 1|r_t) = \Pr(a_t = \theta_t|r_t)$ for each task t; in other words, the underlying data-generating process is the performance function. Using the correctness data allows us to infer whether subjects have behavior consistent with having an

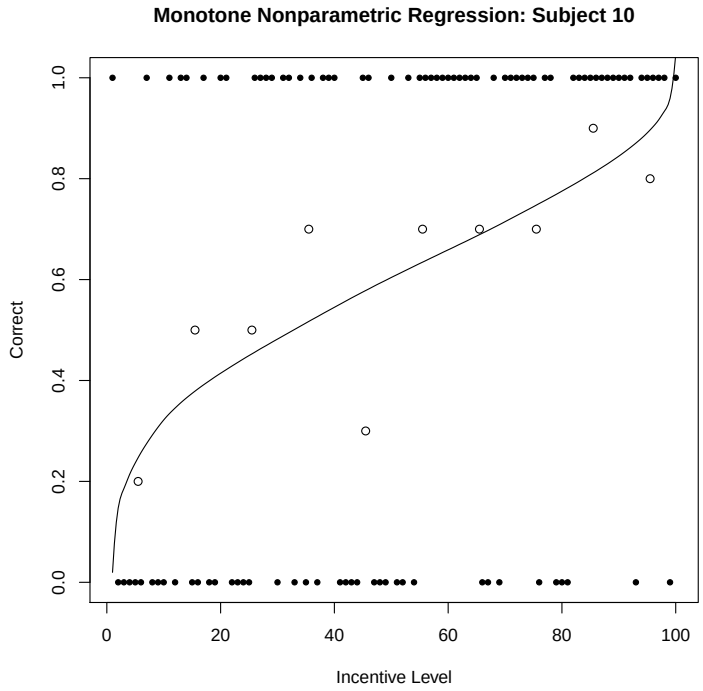


Figure 9: Isotone nonparametric regression of correctness on incentive level for Laboratory Subject 10 (Dette et al., 2006). Circles indicate average success rate within each bin of 10 incentives.

information cost function, and if so, what its properties are.

We are able to categorize subjects according to whether their behavior adheres to these properties. First, we classify them by whether or not they are rationally inattentive. Then, we classify rationally inattentive subjects by whether or not they are responsive to incentives. This subset of subjects is the subset of greatest interest to us; these are the subjects for whom we can estimate performance functions and back out corresponding information cost functions. Finally, we classify responsive subjects according to whether or not their behavior is consistent with well-behaved cost functions. This categorization scheme is illustrated in Figure 10.

5.2 Rational Inattentiveness

We now proceed with the individual-level categorization exercise.

Before testing the properties of the subjects' cost functions, it is necessary to determine whether there exists a cost function that rationalizes their data in the first place. To that end, we test the necessary and sufficient "no improving attention cycles" and "no improving action switches"

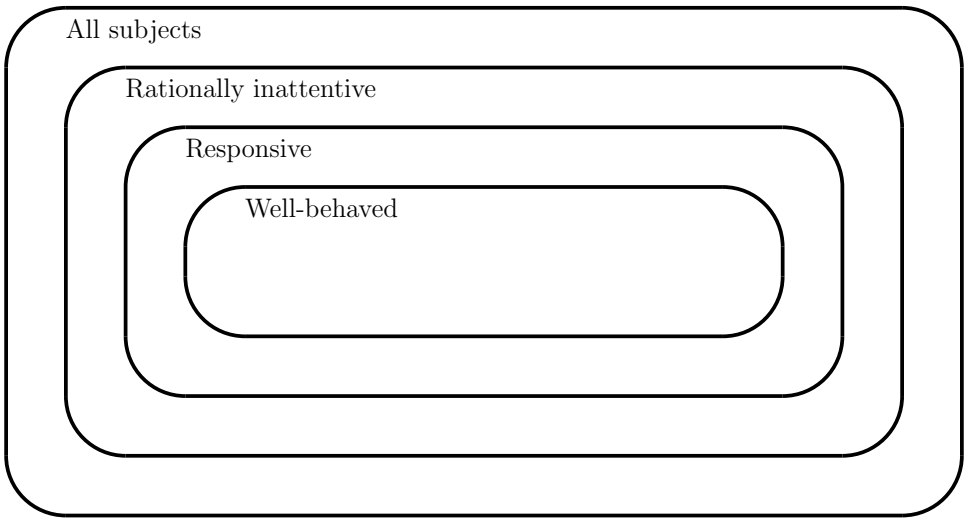


Figure 10: Categorization of subjects

conditions by testing the equivalent conditions established in Subsection 2.2.

5.2.1 No Improving Attention Cycles

As demonstrated in Proposition 1, a subject satisfies NIAC in our experiment if and only if their probability of correctly guessing the state is non-decreasing in the reward. This implies that rationally inattentive subjects have non-decreasing performance functions.

At this point, a clarification is in order. As we showed in Proposition 1, NIAC holds in a set of uniform guess tasks iff for any pair of incentive levels (r_1, r_2) with $r_1 > r_2$, we have that $P^*(r_1) \geq P^*(r_2)$. Observationally, this means that the subject had more correct answers under incentive level r_1 than incentive level r_2. However, in our experiment each subject is given each incentive level only once. Therefore, the empirically-observed probabilities of answering each decision problem correctly are either 0 or 1. If were to apply the NIAC condition directly to our data, this would mean that the only subjects whose behavior is consistent with NIAC would be those who always answer incorrectly up to some incentive threshold after which they always answer correctly. Given the stochasticity of choice under limited attention, this scenario is implausible.

Therefore, rather than strictly interpreting our data as stochastic choice data and making direct pairwise comparisons of decision problems to test NIAC, we adopt an estimation-based approach. We flexibly estimate the performance function given correctness data and see if this estimate is significantly different from a non-decreasing function, in which case we reject NIAC. In theory,

unless there is some reward threshold below which the subject is never correct and above which the subject is always correct, the fit of a monotone performance function can be improved by adding peaks and troughs. The question, then, that we wish to pose is not whether a non-monotone or decreasing function can fit the data, but whether we can reject the hypothesis that a non-decreasing function explains the data.

To test for weak positive monotonicity, we employ a method developed by Doveh et al. (2002) and compare the estimation of an unrestricted cubic polynomial regression of correctness on incentive level for each subject to one with a positive derivative restriction.[17] The null hypothesis for this test is that the response function is monotonic. At the 5% level, we fail to reject positive monotonicity for 77 out of 81 lab subjects (95.1%).[18] Examples of polynomial regressions and correctness data for two subjects, one who rejects NIAC and one who fails to reject NIAC, are depicted in Figure 11.

5.2.2 No Improving Action Switches

To test for the second necessary and sufficient condition for rational inattentiveness, NIAS, we cannot simply simply examine the estimated performance function; following Proposition 2, we must look at the posterior probabilities of each state given each response. We employ a bootstrap procedure. For each subject and action, we calculate the empirically observed distribution of true states, i.e. we calculate $\Pr(\theta = y|a = x)$ for each x and y.[19] We then simulate 499 bootstrap samples for each distribution. If the most common true state is the one corresponding to the action in at least 5% of samples for each action for a given subject, then that subject fails to reject NIAS. In other words, we check that $\Pr(\theta = x|a = x)$ is maximized at $\theta = x$, for each $a = x$, in at least

[17] Several other methods in the statistical and econometric literatures have been devised to test for the monotonicity of regression, including but not limited to Bowman et al. (1998), Ghosal et al. (2000), Hall and Heckman (2000), Birke and Dette (2007), and Chetverikov (2019), most of which are nonparametric. We use Doveh et al.'s (2002) parametric test because it is less prone to rejecting monotonicity when there are outliers, e.g. a lone failure in a region of success, or vice versa.

[18] The optimization in the computation of the restricted regression for lab subject 35 failed to converge, and so we did not perform the test for them. For that subject, a one-tailed t-test of the coefficient on incentive level in a linear regression of correctness on incentive level failed to reject the null of the coefficient being non-negative at the 5% level, and so we classify them as having a non-decreasing performance function.

[19] It should be noted that strictly speaking, the NIAS condition applies separately to each decision problem that the DM faces. Since each subject faces each incentive level only once, they actually face 100 different decision problems. For that reason, we test a slightly weaker condition: whether an individual exhibits overall systematic misuse of information. Overall systematic misuse of information implies systematic misuse of information in at least one decision problem.

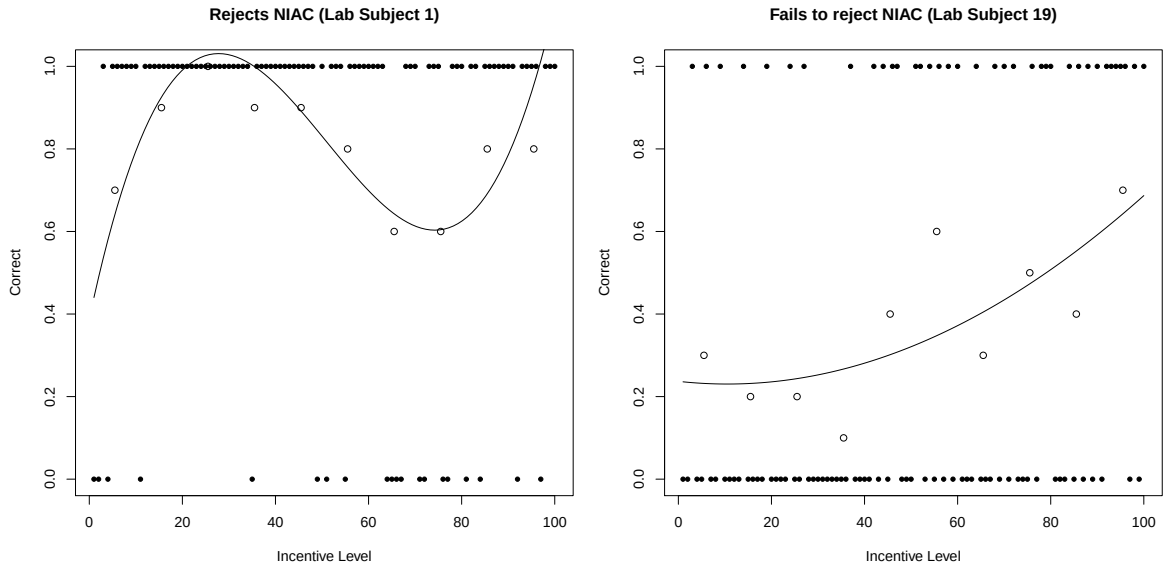


Figure 11: Unrestricted cubic polynomial regression of correctness on incentive level for Subjects 1 and 19. The former rejects NIAC (and therefore rejects rational inattentiveness), and the latter fails to reject NIAC. Circles indicate average success rate within each bin of 10 incentives.

5% of samples. Overall, we find that 74 out of 81 (91.3%) laboratory subjects fail to reject NIAS.

Overall, 70 out of 81 (86.4%) laboratory subjects fail to reject both NIAC and NIAS. We refer to these subjects as "rationally inattentive," or simply "rational," subjects.

5.3 Responsiveness

Of the subjects who fail to reject rational inattentiveness, some of them may have flat response functions, i.e. while they could be rationally inattentive, they do not actually respond to incentives (within the range of incentives presented to them).

To determine which subjects are responsive to incentives, for each subject who failed to reject rational inattentiveness, we run a linear weighted least squares regression of correctness on incentive level and run a one-sided t-test of the coefficient on incentive level with the null of non-positivity, i.e. non-responsiveness to incentives. However, this is insufficient to capture all responsive subjects; a subject may be responsive only within a small range of incentives. To address this issue, for each subject, we repeat this procedure on incentive levels 1 through 50 and on incentive levels 51 through 100.[20] If a subject has a significantly positive coefficient on incentive level in any of these three

[20]Further sample splitting leads to the spurious detection of responsiveness; it leads to some subjects with >95%

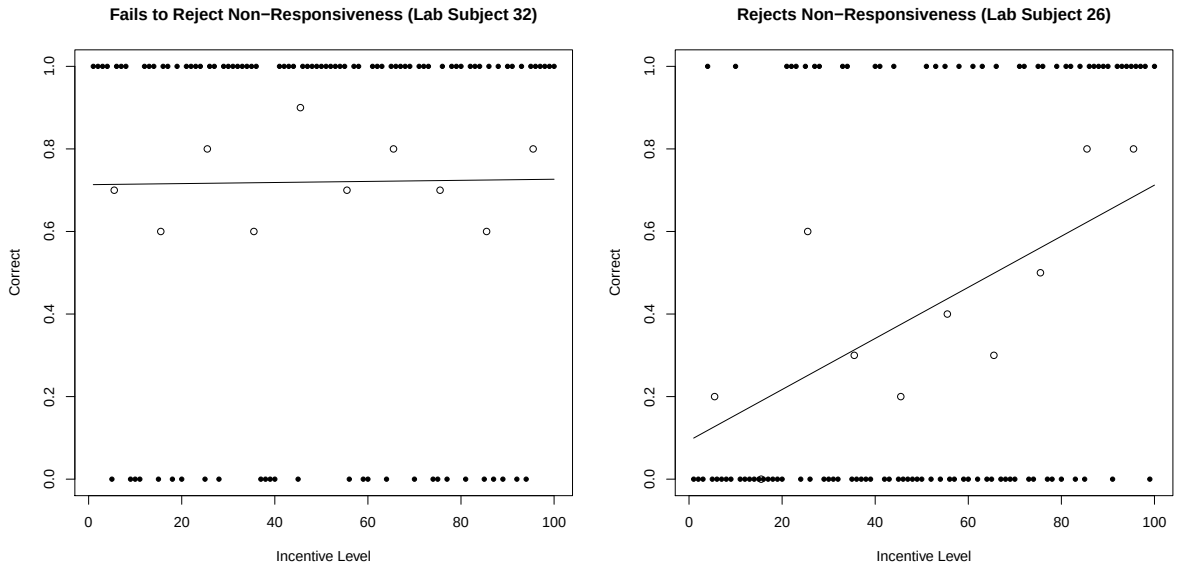


Figure 12: Linear regressions of correctness on incentive level for two subjects. The left panel shows an unresponsive subject, and the right panel shows a responsive one. Circles indicate average success rate within each bin of 10 incentives.

regressions, then we classify them as responsive.[21]

At the 5% significance level, 42 out of 70 lab subjects (60.0%) who fail to reject rational inattentiveness are responsive to incentives. Examples of full-sample linear regressions and correctness data for two subjects, one who fails to reject non-responsiveness and one who rejects non-responsiveness, are depicted in Figure 12.

5.4 Well-Behavedness

If the assumptions of Proposition 3 are satisfied, i.e. the cost function is well-behaved, then the performance function should be continuous in r. Therefore, observing a discontinuity in the performance function indicates a violation of convexity.

Strictly speaking, one cannot definitively observe a discontinuity without an infinite data set; a continuous function with a sufficiently steep slope at points of potential discontinuity can always be used to fit finite data. Therefore, for each subject, the question we wish to answer is whether it

success being classified as responsive.

[21] We must consider the full-sample regressions in tandem with the split-sample regressions. If we considered only the split-sample regressions, then we would classify subjects who have binary-response performance functions with thresholds around 50 as non-responsive.

is more plausible that a discontinuous performance function or a continuous performance function generated their correctness data. This implies a statistical test where the null hypothesis is that the performance function belongs to some class of discontinuous functions, and the alternative is that the performance function belongs to some class of continuous functions.

We test for the presence of a discontinuity by applying a likelihood ratio test. We estimate a step function of the form:[22]

$$P_t = \beta_0 + \beta_1 \mathbb{1}_{[\delta,\infty)}(r_t) \tag{16}$$

where β_0, β_1, and δ are the parameters to be estimated and compare its likelihood to an estimation of the following logistic relationship:

$$P_t = \frac{\beta_1}{1 + \exp(-\lambda(r_t - \delta))} + \beta_0 \tag{17}$$

In (16), δ is the location of the discontinuity, whereas in (17), it is a location parameter that determines the midpoint of the curve's upward sloping portion. It can be shown that (16) is the pointwise limit of (17) as λ goes to infinity. Therefore, (16) can be seen as the restricted null model, and a likelihood ratio test comparing these models is effectively a test of the null hypothesis that $\lambda = \infty$, i.e. it is a test against the null hypothesis that there is a jump discontinuity. Since we are performing this test only on responsive subjects, our estimates of β_1 for each subject should be positive, and therefore this procedure should not detect spurious downward jump discontinuities for those subjects.[23]

Using this test, at the 5% level we cannot reject that 29 out of 42 responsive lab subjects (69.0%) have discontinuities in their response functions.[24]

[22] We use a variant of the procedure of Bai and Perron (1998) for this estimation, *imposing* a discontinuity and determining its location instead of using their algorithm to determine whether such a discontinuity is present.

[23] Several procedures for detecting discontinuities have been proposed in the econometric literature. See, for example, Andrews (1993), Andrews and Ploberger (1994), Bai and Perron (1998), and Porter and Yu (2015). All of these procedures are designed to detect both positive and negative jump discontinuities, and so they are vulnerable to the detection of spurious negative jumps in our setting. A clarification is in order here. Bai and Perron (1998) propose both an estimation procedure and a testing procedure for models with structural breaks with unknown discontinuity points. We use their estimation procedure to estimate (16), but we do not use their testing procedure.

[24] As a robustness check, we also ran this test at the 10% level to gain additional statistical power. In this case, we cannot reject that 27 out of 42 responsive lab subjects (64.2%) are not well-behaved.

Table 3: Categorization of subjects

Category	**Of All Subjects**	**Of R.I. Subjects**	**Of Resp. Subjects**
All subjects	81 (100%)	—	—
R.I. subjects	70 (86.4%)	70 (100%)	—
Resp. subjects	///	42 (60.0%)	42 (100%)
W.B. subjects	///	///	13 (31.0%)

Note: "R.I." = rationally inattentive; "Resp." = responsive; "W.B." = well-behaved, i.e. subjects whose behavior is consistent with continuous, convex cost functions. — denotes that the column category is a subset of the row category, and /// denotes that the row category is defined only on a subset of the column category.

Table 4: Performance functions estimated and their corresponding cost functions

	Cost Function	**Performance Function**	**Ref.**
1	Very high or infinite marginal costs	Constant	N/A
2	Simple dual-process or concave	Binary	3.4/3.5
3	Hybrid dual-process	Affine with break	3.5
4	Quadratic in performance	Affine (without break)	3.1/3.2
5	Convex on the order of $P^{\frac{3}{2}}$	2nd degree polynomial	3.1
6	Convex on the order of $P^{\frac{4}{3}}$	3rd degree polynomial	3.1
7	Shannon mutual information	Logistic	3.2
8	Posterior-separable with Tsallis entropy	Sigmoid/inverse-S/concave (SIC)	3.2
9	Normal signals with linear cost of precision	Concave	3.3

Notes: The "Ref." column indicates in which subsections of Section 3 the relevant theoretical treatment can be found. Model 1 is included as a robustness check.

5.5 Summary of Categorization

Table 3 summarizes the results of preceding subsections. Each cell indicates the number and percentage of row category subjects in the column category. It should be noted that the vast majority (86.4%) of subjects are rationally inattentive, and moreover, most rationally inattentive subjects are responsive (60.0%).

6 Model Selection

In this section, for each responsive subject we fit several possible parametric functional forms for performance functions, each of which can be generated by some cost function. These models

are listed in Table 4.[25] Estimating equations, estimation methods, and mappings from estimated parameters to cost function parameters are detailed in Table 5. In contrast to existing experimental papers that estimate information costs on laboratory data (e.g. Pinkovskiy, 2009; Cheremukhin et al., 2015; Dean and Neligh, 2019), we estimate not just models that nest Shannon mutual information, but also models that do not, such as fixed costs for information (Model 2) and normal signal costs (Model 9).

Since the models are non-nested and are estimated using different methods, we cannot use a traditional auxiliary regression method for model selection. To determine which model is the best fit for each responsive subject, we estimate each model for each such subject and then compare their Akaike Information Criteria (AIC) (Akaike, 1974), selecting the model that yields the lowest AIC. The results of this selection are given in Table 6.[26]

All responsive subjects are best fit by binary (fixed costs), logistic (mutual information), SIC (Tsallis entropy costs) or concave (normal signals with linear precision cost) performance. The first implies some sort of non-convexity or discontinuity in the cost function, whereas the latter three are consistent with convex cost functions. Figures 13, 14, 15, and 16 show what these performance functions look like for four subjects, each best fit by a different model.

Table 7 shows the average estimated AIC and rank of each model in the selection exercise.[27] Models 2 (binary), 7 (logistic), and 8 (SIC) have the lowest ranks on average. Flexible polynomial fits do quite poorly; the average rank of a cubic performance function (Model 6) is higher than that of the constant performance model (Model 1).

[25] The reason that we do not consider the channel capacity cost function is because since the prior distribution in our task is uniform, channel capacity would be consistent with the same behavior as mutual information (cf. Section 1.2.3 of Woodford, 2012a)

[26] As a robustness check, we also perform the analysis with the small sample-corrected AIC (AICc), where AICc $= AIC + \frac{2k(k+1)}{T-k-1}$, T is the number of tasks, and k is the number of parameters in the model (Technically, the small sample correction should depend on the underlying model, but this particular correction formula is said to be appropriate for a wide variety of settings. For more information, refer to Subsection 7.4.1 of Burnham and Anderson, 2002). Our qualitative findings are completely unaffected. In particular, none of the subjects have a different best-fitting model under the AICc than under the AIC.

[27] A lower rank implies a lower AIC and therefore a better fit.

Table 5: Estimating equations for performance functions

	Perf. F'n	**Estimating Equation**	**Method**	**Estimated Cost Function**
1	Constant	$P_t = \beta_0$	OLS	N/A
2	Binary	$P_t = \beta_0 + \beta_1 \mathbb{1}_{[\delta,\infty)}(r_t)$	BP98	$\hat{C}(P) = \begin{cases} 0, & P \leq \hat{\beta}_0 \\ \hat{\delta}(\hat{\beta}_1 - \hat{\beta}_0), & P \in (\hat{\beta}_0, \hat{\beta}_0 + \hat{\beta}_1] \\ \infty, & P > \hat{\beta}_1 \end{cases}$
3	Affine w/ break	$P_t = \beta_0 + \beta_1 \mathbb{1}_{[\delta,\infty)}(r_t) + \beta_2 r_t + \beta_3 r_t \cdot \mathbb{1}_{[\delta,\infty)}(r_t)$	BP98	$\hat{C}(P) = \begin{cases} 0, & P \leq \hat{\beta}_0 \\ \frac{1}{2\hat{\beta}_1}(P - \hat{\beta}_0)^2, & \hat{\beta}_0 < P \leq \hat{d}_2 \\ \frac{1}{2\hat{\beta}_1}(\hat{d}_2 - \hat{\beta}_0)^2, & \hat{d}_2 < P \leq \hat{\beta}_0 + \hat{\beta}_2 \\ \frac{1}{2(\hat{\beta}_1 + \hat{\beta}_3)}(P - (\hat{\beta}_0 + \hat{\beta}_2))^2 + \frac{1}{2\hat{\beta}_1}(\hat{d}_2 - \hat{\beta}_0)^2, & P > \hat{\beta}_0 + \hat{\beta}_2 \end{cases}$ where $\hat{d}_2 = \hat{\beta}_0 + \sqrt{\hat{\beta}_1 \hat{\delta}\left(\hat{\beta}_3 \hat{\delta} + 2\hat{\beta}_2\right)}$
4	Affine	$P_t = \beta_0 + \beta_1 r_t$	WLS	$\hat{C}(P) = \frac{1}{2\hat{\beta}_1}(P^2 - 0.04) - \frac{\hat{\beta}_0}{\hat{\beta}_1}(P - 0.2)$
5	2nd deg. poly.	$P_t = \beta_0 + \beta_1 r_t + \beta_2 r_t^2$	WLS	$\hat{C}(P) = \frac{\hat{\beta}_1}{2\hat{\beta}_2}(0.2 - P) \pm \frac{1}{24\hat{\beta}_2}\left(\left(4\hat{\beta}_2(P - \hat{\beta}_0) + \hat{\beta}_1^2\right)^{\frac{3}{2}} - \left(4\hat{\beta}_2(0.2 - \hat{\beta}_0) + \hat{\beta}_1^2\right)^{\frac{3}{2}}\right)$ where the $+$ is taken if $\hat{\beta}_2 > 0$ and the $-$ is taken if $\hat{\beta}_2 < 0$
6	3rd deg. poly.	$P_t = \beta_0 + \beta_1 r_t + \beta_2 r_t^2 + \beta_3 r_t^3$	WLS	Omitted for legibility
7	Logistic	$P_t = \frac{1}{4\exp(-\frac{r_t}{\alpha})+1}$	MLE	$\hat{C}(P) = \hat{\alpha}\left[\ln(n) + P\ln(P) + (1-P)\ln\left(\frac{1-P}{n-1}\right)\right]$
8	SIC	$r_t + \frac{\alpha\sigma}{\sigma-1}\left[\left(\frac{1-P_t}{4}\right)^{\sigma-1} - P_t^{\sigma-1}\right] = 0$	MLE	$\hat{C}(P) = \begin{cases} \frac{\hat{\alpha}}{\hat{\sigma}-1}\left[P^{\hat{\sigma}} + (n-1)^{1-\hat{\sigma}}(1-P)^{\hat{\sigma}} - n^{1-\hat{\sigma}}\right], & \hat{\sigma} \neq 1 \\ \hat{\alpha}\left[\ln(n) + P\ln(P) + (1-P)\ln\left(\frac{1-P}{n-1}\right)\right], & \hat{\sigma} = 1 \end{cases}$
9	Concave	$P_t = \frac{8}{5}\Phi\left(\frac{\zeta^*(r_t)}{2}\right) - \frac{3}{5}$, where $\alpha\zeta^* = \frac{2}{5} r_t \phi\left(\frac{\zeta^*}{2}\right)$	MLE	$\hat{C}(P) = \hat{\alpha}\breve{P}^{-1}(P)$, where $\breve{P}(\zeta) := \frac{1}{5}\left[2\Phi\left(\frac{1}{2}\zeta\right) + 3\left(2\Phi\left(\frac{1}{2}\zeta\right) - 1\right)\right]$

Notes on estimation: "BP98" = Bai and Perron (1998). Model 2 is estimated using a variant of their algorithm where a structural break is imposed rather than detected. Notes on nesting of models: Model 2 nests Model 1. Model 3 nests Models 2 and 4. Model 6 nests Model 5, which nests Model 4, which nests Model 1. Model 8 nests Models 4 and 7.

Table 6: Model Selection for Responsive Subjects

Model	Binary (2)	Logistic (7)	SIC (8)	Concave (9)
Number of Subjects	10 (23.8%)	19 (45.2%)	7 (16.7%)	6 (14.3%)

Table 7: Average AIC and Rank for Estimated Models

	Model	AIC	Rank
1	Constant	131.165	7.476
2	Binary	114.060	2.881
3	Affine w/ break	119.982	4.429
4	Affine	117.832	5.833
5	Quadratic	131.123	6.595
6	Cubic	132.929	7.643
7	Logistic	116.008	2.714
8	SIC	113.672	2.310
9	Concave	121.967	5.119

Note that the average AIC and rank for Model 7, the logistic performance function, is only slightly lower than that of Model 2, the binary performance function, despite the fact that substantially more subjects are best fit by Model 7 than by Model 2. This may be due to the binary model being a decent fit when the best-fitting model is logistic, but the logistic model being a poor fit when the best-fitting model is binary. When the logistic model is the best fit, the average rank of the binary model is 3.368; however, when the binary model is the best fit, the average rank of the logistic model is 4.600. Note also that the average rank of Model 9 (normal signals with linear precision cost) is fairly high at 5.119. This indicates that when Model 9 is not the best fit for a subject, it is a poor fit.

7 Conclusion

This paper has provided a schema for testing properties of and estimating information cost functions in a rational inattention framework. To the extent that the presence or absence of characteristics such as continuity and convexity can have an impact on people's decisions, it is worth knowing whether their cost functions satisfy such conditions. Decision-makers' cost functions are not directly observable, so instead we must infer their characteristics from observed behavior. We conducted a set of experiments that allowed us to implement tests of the properties of interest and perform a

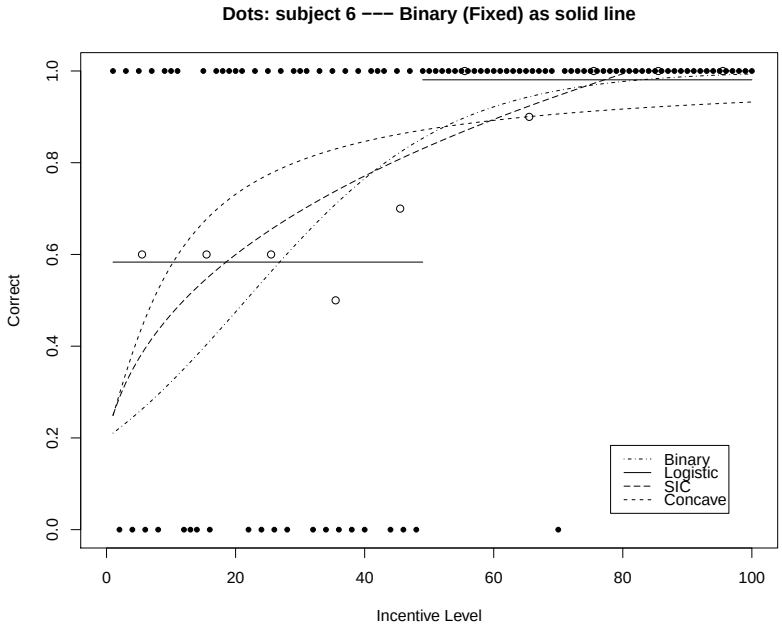


Figure 13: Fits of binary, logistic, SIC, and concave performance for Subject 6, with the best-fitting binary (fixed cost) model as a solid line. Solid dots indicate correctness for each incentive. Circles indicate average success rate within each bin of ten incentives.

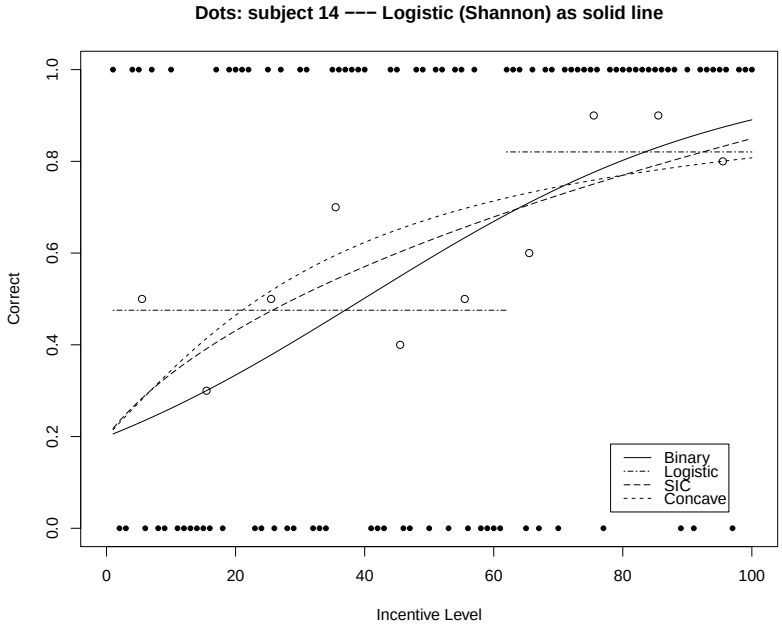


Figure 14: Fits of binary, logistic, SIC, and concave performance for Subject 14, with the best-fitting logistic (Shannon) model as a solid line. Solid dots indicate correctness for each incentive. Circles indicate average success rate within each bin of ten incentives.

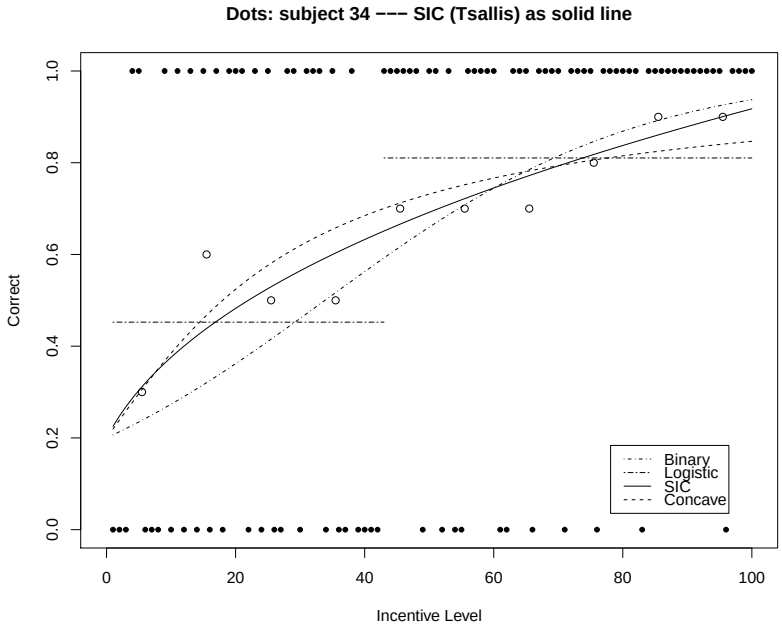


Figure 15: Fits of binary, logistic, SIC, and concave performance for Subject 34, with the best-fitting SIC (Tsallis) model as a solid line. Solid dots indicate correctness for each incentive. Circles indicate average success rate within each bin of ten incentives.

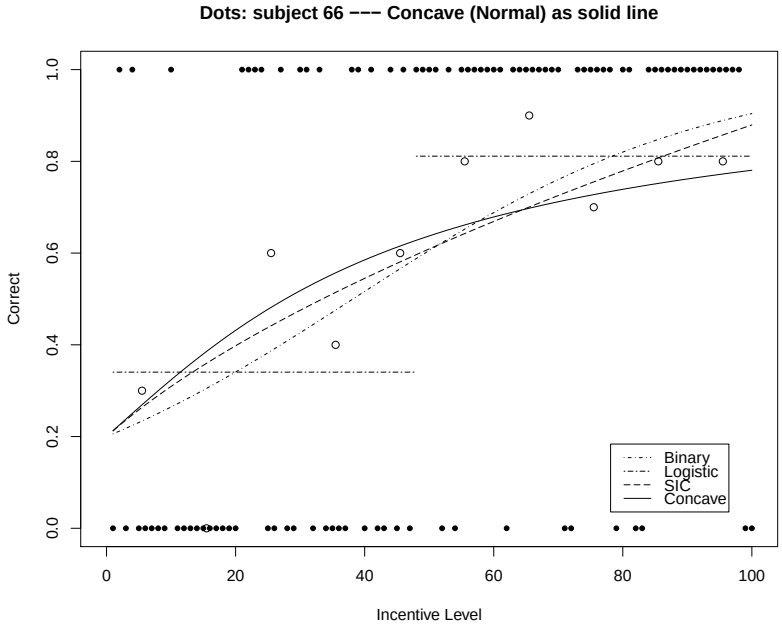


Figure 16: Fits of binary, logistic, SIC, and concave performance for Subject 66, with the best-fitting concave (normal) model as a solid line. Solid dots indicate correctness for each incentive. Circles indicate average success rate within each bin of ten incentives.

model selection exercise.

These experiments reveal substantial heterogeneity in behavior. Most subjects are rationally inattentive, but only about half are actually responsive to incentives. Many subjects have behavior that is consistent with continuous, convex cost functions, but a substantial fraction do not. Moreover, there is considerable heterogeneity in how subjects adjust their attention in response to incentives, though this heterogeneity is limited to four classes of cost functions: fixed costs, mutual information, Tsallis entropy costs (which nest mutual information), and normal signals are the only best-fitting cost functions for responsive subjects in terms of performance.

The fact that there is a significant presence of both binary performance and continuous performance functions in the population has important implications for economic modeling. In Supplementary Appendix S6, we present an application of rational inattention to a principal-agent framework of investment delegation and show that the principal's optimal payment schedule crucially depends on the shape of the agent's information cost function, and moreover, equilibrium robustness in this model relies on continuity; if an agent's information cost function is discontinuous, infinitesimal deviations from the optimal contract can lead to large welfare losses for the principal. Our experimental results also indicate that if a modeler wishes to use a single cost function for all agents for the sake of simplicity, then Tsallis costs, which have the lowest average rank and AIC, may be a good compromise, due to their flexibility.

Three possible avenues for future experimental research present themselves. The first is to obtain more detailed data on what subjects are actually paying attention to. Eyetracking has already been used in several economics experiments (e.g. Wang et al., 2010; Krajbich et al., 2010; Arieli et al., 2011) to track subjects' gaze, which allows researchers to find out what visual information the subjects are acquiring. Tracking subjects' mouse movements in computer-based tasks (e.g. Gabaix et al., 2006; Reeck et al., 2017) is another potential approach, since those movements indicate to which areas of their computer monitors they are paying attention. The second is to use choice data in tandem with reaction time data to fit models of dynamic information acquisition (e.g. Ratcliff and Smith, 2004; Clithero, 2018; Webb, 2018). This would also allow researchers to determine to what extent subjects trade off speed and accuracy in their decision-making. The third is to exploit the structure of the state space and data on subjects' mistakes to study how subjects perceive distance and dissimilarity (e.g. Natenzon, 2019; Pomatto et al., 2019).

References

Anat R. Admati and Paul Pfleiderer. A theory of intraday patterns: Volume and price variability. *The Review of Financial Studies*, 1(1):3–40, 1988.

Hirotugu Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, 1974.

Hunt Allcott and Dmitry Taubinsky. Evaluating behaviorally motivated policy: Experimental evidence from the lightbulb market. *American Economic Review*, 105(8):2501–38, 2015.

Donald W. K. Andrews. Tests for parameter instability and structural change with unknown change point. *Econometrica*, 61(4):821–856, 1993.

Donald W. K. Andrews and Werner Ploberger. Optimal tests when a nuisance parameter is present only under the alternative. *Econometrica*, 62(6):1383–1414, 1994.

Amos Arieli, Yaniv Ben-Ami, and Ariel Rubinstein. Tracking decision makers under uncertainty. *American Economic Journal: Microeconomics*, 3(4):68–76, 2011.

Jushan Bai and Pierre Perron. Estimating and testing linear models with multiple structural changes. *Econometrica*, 66(1):47–78, 1998.

Gadi Barlevy and Pietro Veronesi. Information acquisition in financial markets. *Review of Economic Studies*, 67(1):79–90, 2000.

Vojtěch Bartoš, Michal Bauer, Julie Chytilová, and Filip Matějka. Attention discrimination: Theory and field experiments with monitoring information acquisition. *American Economic Review*, 106(6):1437–1475, 2016.

Victor L. Bernard and Jacob K. Thomas. Post-earnings-announcement drift: Delayed price response or risk premium? *Journal of Accounting Research*, 27:1–36, 1989.

Melanie Birke and Holger Dette. Testing strict monotonicity in nonparametric regression. *Mathematical Methods of Statistics*, 16(2):110–123, 2007.

A.W. Bowman, M.C. Jones, and I. Gijbels. Testing monotonicity of regression. *Journal of Computational and Graphical Statistics*, 7(4):489–500, 1998.

Kenneth P. Burnham and David R. Anderson. *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. Springer, 2002.

Andrew Caplin and Mark Dean. Revealed preference, rational inattention, and costly information acquisition. *Working paper*, 2014.

Andrew Caplin and Mark Dean. Revealed preference, rational inattention, and costly information acquisition. *American Economic Review*, 105(7):2183–2203, 2015.

Andrew Caplin and Daniel Martin. Defaults and attention: The drop out effect. *Revue économique*, 68(5):747–755, 2017.

Andrew Caplin, Mark Dean, and John Leahy. Rationally inattentive behavior: Characterizing and generalizing Shannon entropy. *Working paper*, 2019.

Christopher P. Chambers, Ce Liu, and John Rehbeck. Costly information acquisition. *Working paper*, 2019.

Anton Cheremukhin, Anna Popova, and Antonella Tutino. A theory of discrete choice with information costs. *Journal of Economic Behavior & Organization*, 113:34–50, 2015.

Raj Chetty, Adam Looney, and Kory Kroft. Salience and taxation: Theory and evidence. *American Economic Review*, 99(4):1145–1177, 2009.

Denis Chetverikov. Testing regression monotonicity in econometric models. *Econometric Theory*, 35(4):729–776, 2019.

John A. Clithero. Improving out-of-sample predictions using response times and a model of the decision process. *Journal of Economic Behavior & Organization*, 148:344–375, 2018.

Babur De los Santos, Ali Hortaçsu, and Matthijs R. Wildenbeest. Testing models of consumer search using data on web browsing and purchasing behavior. *American Economic Review*, 102 (6):2955–2980, 2012.

Mark Dean and Nathaniel Neligh. Experimental tests of rational inattention. *Working Paper*, 2019.

Stefano DellaVigna. Psychology and economics: Evidence from the field. *Journal of Economic Literature*, 47(2):315–372, 2009.

Stefano DellaVigna and Joshua M. Pollet. Demographics and industry returns. *American Economic Review*, 97(5):1667–1702, 2007.

Holger Dette, Natalie Neumeyer, and Kay F. Pilz. A simple nonparametric estimator of a strictly monotone regression function. *Bernoulli*, 12(3):469–490, 2006.

E. Doveh, A. Shapiro, and P.D. Feigin. Testing of monotonicity in parametric regression models. *Journal of Statistical Planning and Inference*, 107(1–2):289–306, 2002.

Michael Ehrmann and David-Jan Jansen. The pitch rather than the pit: Investor inattention, trading activity, and FIFA World Cup matches. *Journal of Money, Credit and Banking*, 49(4): 807–821, 2017.

Xavier Gabaix, David Laibson, Guillermo Moloche, and Stephen Weinberg. Costly information acquisition: Experimental analysis of a boundedly rational model. *American Economic Review*, 96(4):1043–1068, 2006.

Matthew Gentzkow and Emir Kamenica. Costly persuasion. *American Economic Review*, 104(5): 457–462, 2014.

Subhashis Ghosal, Arusharka Sen, and Aad W. van der Vaart. Testing monotonicity of regression. *Annals of Statistics*, 28(4):1054–1082, 2000.

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

Sanford J. Grossman and Joseph E. Stiglitz. On the impossibility of informationally efficient markets. *American Economic Review*, 70(3):393–408, 1980.

Peter Hall and Nancy E. Heckman. Testing for monotonicity of a regression mean by calibrating for linear functions. *Annals of Statistics*, 28(1):20–39, 2000.

Christian Hellwig, Sebastian Kohls, and Laura Veldkamp. Information choice technologies. *American Economic Review*, 102(3):35–40, 2012.

David Hirshleifer, Sonya Seongyeon Lim, and Siew Hong Teoh. Driven to distraction: Extraneous events and underreaction to earnings news. *Journal of Finance*, 64(5):2289–2325, 2009.

Daniel Ho and Kosuke Imai. Estimating causal effects of ballot order from a randomized natural experiment: The California alphabet lottery, 1978–2002. *Public Opinion Quarterly*, 72(2):216–240, 2008.

Tanjim Hossain and John Morgan. ...Plus shipping and handling: Revenue (non) equivalence in field experiments on eBay. *The B.E. Journal of Economic Analysis & Policy*, 6(2):1–30, 2006.

Gur Huberman. Familiarity breeds investment. *The Review of Financial Studies*, 14(3):659–680, 2001.

Daniel Kahneman. Maps of bounded rationality: Psychology for behavioral economics. *American Economic Review*, 93(5):1449–1475, December 2003.

E.L. Kaufman, M.W. Lord, T.W. Reese, and J. Volkmann. The discrimination of visual number. *American Journal of Psychology*, 62(4):498–525, 1949.

Ian Krajbich, Carrie Armel, and Antonio Rangel. Visual fixations and the computation and comparison of value in simple choice. *Nature Neuroscience*, 13(10):1292–1298, 2010.

Nicola Lacetera, Devin G. Pope, and Justin R. Sydnor. Heuristic thinking and limited attention in the car market. *American Economic Review*, 102(5):2206–2236, 2012.

Ulrike Malmendier and Devin Shanthikumar. Are small investors naive about incentives? *Journal of Financial Economics*, 85(2):457–489, 2007.

Filip Matějka and Alisdair McKay. Rational inattention to discrete choices: A new foundation for the multinomial logit model. *American Economic Review*, 105(1):272–298, 2015.

Ryan C. McDevitt. "A" business by any other name: Firm name choice as a signal of firm quality. *Journal of Political Economy*, 122(4):909–944, 2014.

Stephen Morris and Ming Yang. Coordination and continuous stochastic choice. *Working paper*, 2019.

Paulo Natenzon. Random choice and learning. *Journal of Political Economy*, 127(1):419–457, 2019.

Maxim L. Pinkovskiy. Rational inattention and choice under risk: Explaining violations of expected utility through a Shannon entropy formulation of the costs of rationality. *Atlantic Economic Journal*, 37(1):99–112, 2009.

Luciano Pomatto, Philipp Strack, and Omer Tamuz. The cost of information. *Working paper*, 2019.

Devin G. Pope. Reacting to rankings: Evidence from "America's Best Hospitals". *Journal of Health Economics*, 28(6):1154–1165, 2009.

Jack Porter and Ping Yu. Regression discontinuity designs with unknown discontinuity points: Testing and estimation. *Journal of Econometrics*, 189(1):132–147, 2015.

Roger Ratcliff and Philip L. Smith. A comparison of sequential sampling models for two-choice reaction time. *Psychological Review*, 111(2):333–367, 2004.

Crystal Reeck, Daniel Wall, and Eric J. Johnson. Search predicts and changes patience in intertemporal choice. *Proceedings of the National Academy of Sciences*, 114(45):11890–11895, 2017.

Alvin E. Roth and Michael W. Malouf. Game-theoretic models and the role of information in bargaining. *Psychological Review*, 86(6):574–594, 1979.

I.J. Saltzman and W.R. Garner. Reaction time as a measure of span of attention. *Journal of Psychology*, 25(2):227–241, 1948.

C.E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27(3): 379–423, 1948.

Marilyn L. Shaw and Peter Shaw. Optimal allocation of cognitive resources to spatial locations. *Journal of Experimental Psychology: Human Perception and Performance*, 3(2):201–211, 1977.

Kelly Shue and Erzo Luttmer. Who misvotes? The effect of differential cognition costs on election outcomes. *American Economic Journal: Economic Policy*, 1(1):229–257, 2009.

Christopher A. Sims. Implications of rational inattention. *Journal of Monetary Economics*, 50(3): 665–690, 2003.

Christopher A. Sims. Rational inattention: Beyond the linear-quadratic case. *American Economic Review*, 96(2):158–163, 2006.

Stephanie M. Smith and Ian Krajbich. Gaze amplifies value in decision making. *Psychological Science*, 30(1):116–128, 2019.

Keith E. Stanovich and Richard F. West. Individual differences in reasoning: Implications for the rationality debate? *Behavioral and Brain Sciences*, 23(5):645—665, 2000.

Jakub Steiner and Colin Stewart. Perceiving prospects properly. *American Economic Review*, 106 (7):1601–1631, 2016.

Jakub Steiner, Colin Stewart, and Filip Matějka. Rational inattention dynamics: Inertia and delay in decision-making. *Econometrica*, 85(2):521–553, 2017.

Constantino Tsallis. Possible generalization of Boltzmann-Gibbs statistics. *Journal of Statistical Physics*, 52(1-2):479–487, 1988.

Stijn Van Nieuwerburgh and Laura Veldkamp. Information acquisition and under-diversification. *Review of Economic Studies*, 77(2):779–805, 2010.

Robert E. Verrecchia. Information acquisition in a noisy rational expectations economy. *Econometrica*, pages 1415–1430, 1982.

Joseph Tao-yi Wang, Michael Spezio, and Colin F. Camerer. Pinocchio's pupil: Using eyetracking and pupil dilation to understand truth telling and deception in sender-receiver games. *American Economic Review*, 100(3):984–1007, 2010.

Ryan Webb. The (neural) dynamics of stochastic choice. *Management Science*, 65(1):230–255, 2018.

Michael Woodford. Inattentive valuation and reference-dependent choice. *Working paper*, 2012a.

Michael Woodford. Prospect theory as efficient perceptual distortion. *American Economic Review*, 102(3):41–46, 2012b.

Appendices for "Estimating Information Cost Functions in Models of Rational Inattention"

For Online Publication

Ambuj Dewan and Nathaniel Neligh

October 1, 2019

A1 A General Discrete Rational Attention Framework

In Section 2 of the paper, we presented a rational inattention framework that applied to a class of decision problems that we call uniform guess tasks. In this appendix, we present of a general version of that framework. It applies to decision problems with finite state spaces and action spaces.

A1.1 The Framework

In this framework, there is an unknown state of the world about which an decision-maker (DM) can choose to acquire information. This information affects her beliefs about the state of the world. After obtaining this information, she makes a decision that maximizes her payoff given her beliefs.

We model information as a collection of probabilistic mappings from states of the world to a set of subjective signals. We define an *information structure* to be a set of conditional distributions of signals given states. Given a prior belief, observing a signal generates a corresponding posterior belief over states, and given this posterior belief, the DM maximizes her payoff by selecting an optimal action. Each information structure has a cost associated with it.

We remain agnostic about what the exact source of information costs is. Information costs could represent cognitive or physical effort exerted in learning about the true state, as well as the opportunity cost of time spent doing so.

This framework has several beneficial features. Firstly, it has the same behavioral implications as the model of posterior-based information costs of Caplin and Dean (2015) (henceforth CD15),

which means we can apply their necessary and sufficient conditions for models of rational inattention to the problems we study. This assertion requires careful proof, as well as an explanation of exactly where our model departs from CD15's. For the sake of readability, we refer the interested reader to Appendix Subsection A1.6 for details. Secondly, as in some of the previous literature on stochastic choice (e.g. McGuire, 1972; Leshno and Spector, 1992), it expresses information structures as stochastic matrices, which as demonstrated later in the appendix, will permit us to easily compare information structures and to define a simple geometric notion of convexity of information costs; this is not possible with the posterior-based approach of CD15. Thirdly, it allows information costs to depend not just on the beliefs engendered by the information structure, but on *how* those beliefs are engendered. Though this does not rationalize any kind of behavior that is not rationalized under CD15, it allows for the costs leading to certain behaviors to be described more intuitively, especially those that encode the perception of distance, as we explain in Appendix Subsection A1.6 and Supplementary Appendix S1.[A1]

Let $\Theta = \{\theta_i\}_{i=1}^{|\Theta|}$ be a finite state space, let $M = \{m_i\}_{i=1}^{|M|}$ be a finite signal space,[A2] and let $A = \{a_i\}_{i=1}^{|A|}$ be a finite action space, with $|M| \geq |A|$ so that there are at least as many signals as there are actions. Let $\pi = (\pi_i)_{i=1}^{n} \in \Delta(\Theta)$, where $n := |\Theta|$, be the DM's prior over Θ. Each action-state pair (a, θ) has an associated utility $u(a, \theta)$. The DM maximizes:

$$\mathbb{E}_{\gamma_Q^\pi}\left[\mathbb{E}_{\langle \pi|m\rangle}\left[u(a,\theta)\right]\right] - C(\pi, Q) \tag{A1}$$

where Q is an information structure (a collection of conditional signal probabilities, given states), $\gamma_Q^\pi \in \Delta(\Delta(\Theta))$ is the distribution of posterior beliefs it induces given the prior π, $\langle \pi|m\rangle$ is the posterior belief associated with signal m, and C is a cost function that depends on both the prior and the information structure.

As explained above, the DM's problem has two stages. First, she selects an information structure

[A1] See Pomatto et al. (2019) for a further discussion of this point.

[A2] Given that the state space is finite, the finiteness of the signal space is not a substantive restriction. In fact, if we assume that more informative information structures are costlier (our Assumption E, presented later in the paper), it can be shown that given a finite state space, a DM never need use more than a finite number of signals. This follows from Proposition 4 of Kamenica and Gentzkow (2010). They study a game where the information structure and the action are chosen by different players, but if we assume those players' preferences are perfectly aligned, then ignoring information costs, our framework maps onto theirs. By their Proposition 4, if a DM employs an information structure with an infinite number of signals, then ignoring information costs, she could have done at least as well with an information structure with a finite number of signals. Moreover, since the former information structure is more informative than the latter, it is costlier. Therefore, the DM will choose to use a finite number of signals.

Q. She then observes a signal m according to that information structure, which gives her a posterior belief $\langle \pi | m \rangle$, derived by Bayes' rule. Second, given this posterior belief, she chooses an action a to maximize her expected payoff.

We can express this problem more formally using matrix notation. Let $\Pi = \text{diag}(\pi)$.[A3] Let $\mathscr{U} \subseteq M_{|A|\times|\Theta|}(\mathbb{R})$, and let $U \in \mathscr{U}$ be a matrix with entries $u_{i,j} := u(a_i, \theta_j)$, i.e. the utility of taking action i in state j. We refer to $\mathscr{U}$ as the *set of decision problems* and to U as a *payoff matrix*.

Let $\mathscr{Q}$ be the space of right-stochastic matrices of dimension $|\Theta| \times |M|$, and let $\mathscr{D}$ be the space of right-stochastic matrices of dimension $|M| \times |A|$.[A4] $C : \Delta(\Theta) \times \mathscr{Q} \longrightarrow \bar{\mathbb{R}}$ gives the cost[A5] of selecting an information structure from $\mathscr{Q}$, given a prior in $\Delta(\Theta)$.[A6]

The decision-maker's problem, then, is (cf. Leshno and Spector, 1992):[A7]

$$\max_{Q\in\mathscr{Q},D\in\mathscr{D}} \text{tr}(QDU\Pi) - C(\pi, Q) \tag{A2}$$

where the entries of Q are $q_{i,j} = \Pr(m_j|\theta_i)$, i.e. the probably of signal m_j in state θ_i, and the entries of D are $d_{i,j} = \Pr(a_j|m_i)$, i.e. the probability of selecting action a_j given signal m_i. The i-th row of Q represents the conditional distribution of signals given state θ_i, and so Q can be seen as a collection of signal distributions given states. We refer to D as the *decision matrix*.

We refer to the maximand in (A2) as the *net payoff* and its first component as the *ex-ante gross payoff*. Specific realizations of this payoff are called the *ex-post gross payoff*. Where it will not cause confusion, we will drop the "ex-ante" and "ex-post."

This setup allows us to index decision problems of the form of (A2) by (π, U). In this paper, we will hold π fixed, and thus we will simply index decision problems by U where it will cause no confusion. For a given finite sequence of decision problems $\{U_i\}$ drawn from $\mathscr{U}$ and a given true state of the world θ_i, we can observe the action a_i chosen by the DM. Using the data set

[A3] $\text{diag}(x)$ is the square matrix that has the entries of x in order on its diagonal and zeroes elsewhere.

[A4] Some authors require that a stochastic matrix be square. We allow for a stochastic matrix to have different numbers of rows and columns, provided that all its entries are non-negative and each of its rows sums to 1.

[A5] $\bar{\mathbb{R}} := \mathbb{R} \cup \{-\infty, \infty\}$ is the set of extended reals. If for some $\tilde{\pi}$ and $\tilde{Q}$, $C(\tilde{\pi}, \tilde{Q}) = \infty$, then the cost of the information structure $\tilde{Q}$ given $\tilde{\pi}$ is infinite, and the DM will never select it, provided there is at least one information structure available at a finite cost. This idea is formalized in Subsection A1.4.

[A6] In principle, though the cost-function approach implies flexibility in the selection of information structures, it can accommodate restrictions on the space of available information structures as well. For example, if a modeler wishes to impose an exogenous process of information acquisition, then he may set the cost of a corresponding information structure to be finite and the cost of all other information structures to be positive infinity.

[A7] $\text{tr}(X)$ denotes the trace of X, the sum of its diagonal entries.

$\{(U_i, \theta_i, a_i)\}$ will allow us to infer the properties of $C(\cdot, \cdot)$. Following CD15, we refer to a data set of this type as *state-dependent stochastic choice data*.

In this setup, for each $i \in \{1, \ldots, n\}$ and $j \in \{1, \ldots, |A|\}$, π_i and $u_{j,i}$, are exogenous parameters. For each $i \in \{1, \ldots, n\}, j \in \{1, \ldots, |A|\}$, and $k \in \{1, \ldots, |M|\}$, $q_{i,k}$ and $d_{k,j}$ are chosen by subjects. Though one cannot observe $q_{i,k}$ and $d_{k,j}$ separately, one can estimate the products $q_{i,k}d_{k,j}$; if a DM solves the same decision problem repeatedly, one can observe how often each action is chosen in each state.[A8]

A1.2 Testing for Rational Inattention

As CD15 demonstrate, observed behavior is consistent with their model if and only if it satisfies their "no improving attention cycles" (NIAC) and "no improving action switches" (NIAS) conditions.[A9] Their NIAC condition ensures that improvements to gross payoffs cannot be made by reallocating attention cyclically across decision problems, and their NIAS condition ensures that the DM's actions are optimal given the beliefs induced by her chosen information structure. Because our model is behaviorally equivalent to theirs, NIAC and NIAS are necessary and sufficient conditions for stochastic choice data to satisfy our model. Put differently, the DM fails to fulfill either of those two conditions if and only if there does not exist a cost function that rationalizes her stochastic choice data.

In our notation, the NIAC condition can be expressed as follows. Assume a fixed prior π, and let $U_0, U_1, \ldots, U_{J-1}$ be any subset of two or more of the payoff matrices faced by the DM. Let $Q_0, Q_1, \ldots, Q_{J-1}$ and $D_0, D_1, \ldots, D_{J-1}$ be the corresponding information structures and decision matrices selected by the DM, and let D_i^j be a decision matrix that maximizes the gross payoff given payoff matrix U_i and information structure Q_j. Then the NIAC condition states:

$$\sum_{j=0}^{J-1} \text{tr}\left(Q_j D_j U_j \Pi\right) \geq \sum_{j=0}^{J-1} \text{tr}\left(Q_{(j+1) \bmod J} D_j^{(j+1) \bmod J} U_j \Pi\right) \tag{A3}$$

The NIAS condition can be expressed as follows. Assume a fixed prior π. Then for any payoff matrix U, let Q^* be the information structure and D^* be the decision matrix chosen by the DM.

[A8] See Section II.A of CD15 for details.

[A9] The NIAS condition is due to Caplin and Martin (2015). It is the key condition that characterizes their Bayesian expected utility model.

Then the NIAS condition states that for any $k \in \{1, \ldots, |A|\}$ such that the k-th column of D^* (denoted by $d^*_{\bullet,k}$) has at least one nonzero entry and any $l \in \{1, \ldots, |A|\}$:

$$u_{k,\bullet} \Pi Q^* d^*_{\bullet,k} \geq u_{l,\bullet} \Pi Q^* d^*_{\bullet,k} \tag{A4}$$

where $u_{k,\bullet}$ and $u_{l,\bullet}$ are the k-th and l-th rows of U, respectively.

Proposition A1. *NIAC and NIAS are necessary and sufficient conditions for stochastic choice data to satisfy* (A2).

Proof. This follows directly from Proposition A5 in Appendix Subsection A1.6 and Theorem 1 of CD15. □

A1.3 Responsiveness

A set of behaviors that is trivially consistent with rational inattention is one where the DM's behavior is consistent with their posterior beliefs not changing across decision problems; regardless of the decision problem, she chooses the same information structure. This is consistent with models such as signal detection theory, where the DM's information structure is exogenously given. In particular, it does not become more informative if the DM's gross reward from choosing an optimal action increases. In those cases, the DM simply does not respond to changes in the level of incentives across decision problems. More interesting are cases where the DM does modify her behavior in response to changes in the level of incentives.

Definition 1. Suppose that a DM is given a set of decision problems $\mathfrak{U} := \{U_1, U_2, \ldots U_J\}$. Further suppose that $\exists\, U, \tilde{U} \in \mathfrak{U}$ satisfying the following: for each $i \in \{1, \ldots n\}$, let $\tau_i \in \underset{j \in \{1,\ldots,|A|\}}{\text{argmax}}\, u_{i,j}$; $\forall\, i \in \{1, \ldots n\}$, $\tilde{u}_{i,j} \geq u_{i,j}$ if $j = \tau_i$ and $\tilde{u}_{i,j} \leq u_{i,j}$ if $j \neq \tau_i$, with at least one strict inequality. Then we say the DM is *responsive (to incentives)*, or *exhibits responsiveness*, if her behavior is such that $\Pr\left(a \in \underset{z \in A}{\text{argmax}} \in \tilde{u}(z, \theta)\right) > \Pr\left(a \in \underset{z \in A}{\text{argmax}} \in u(z, \theta)\right)$.

Put differently, a DM is responsive to incentives if for some pair of decision problems, her probability of taking a (gross) payoff-maximizing action increases when the utility associated with payoff-maximizing actions increases and the utility associated with non-payoff-maximizing actions decreases.

Responsiveness is a fairly intuitive condition for human behavior to fulfill. Roughly speaking, it says that people perform better (by choosing the best option more often) when the stakes are higher.

A1.4 Continuity and Convexity

In this subsection, we establish sufficient conditions for a continuous relationship between gross payoffs and incentives in general rational inattention problems. Roughly speaking, continuity and convexity of the information cost function imply gross payoffs that are continuous in incentives.

Assumption A. ***Finiteness.*** *$\exists\, \tilde{\mathscr{Q}} \subseteq \mathscr{Q}$ closed, convex, and non-empty such that $\forall\, \pi \in \Delta(\Theta)$, $C(\pi, Q)$ is finite for $Q \in \tilde{\mathscr{Q}}$ and positive infinity otherwise.*

This assumption helps ensure that the DM's decision problem has a solution. We call signal structures in $\tilde{\mathscr{Q}}$ *admissible*.

Assumption B. ***Straightforwardness.*** *$\exists\, \varrho : \{1, \dots |M|\} \longrightarrow \{0, \dots |M|\}$ such that:*

- *$|\varrho(\{1, \dots |M|\})| \leq |A| + 1$, and $\forall\, i \in \{1, \dots, |M|\}, |\varrho^{-1}(i)| \leq 1$.*
- *Given $U \in \mathscr{U}$ and $Q \in \tilde{\mathscr{Q}}$, $\exists\, Q' \in \tilde{\mathscr{Q}}$ such that:*
  - *$\forall\, j \in \{1, \dots, |M|\}$, if $\varrho(j) \neq 0$, then $q'_{\bullet,j} = \sum_{\ell \in V_j^{\pi,U}(Q)} q_{\bullet,\ell}$, where $V_j^{\pi,U}(Q)$ is defined as a non-empty subset of $\{k \mid \varrho(j) \in \operatorname{argmax}_i z_{i,k}\}$, $Z = (z_{i,j}) := U\Pi Q$, and $q_{\bullet,j}$ and $q'_{\bullet,j}$ denote the j-th columns of Q and Q', respectively.*
  - *$\forall\, j \in \{1, \dots, |M|\}$, if $\varrho(j) = 0$, then $q'_{\bullet,j} = \vec{0}$.*
  - *$\forall\, h, j \in \{1, \dots, |M|\}$, if $h \neq j$, then $V_h^{\pi,U}(Q) \cap V_j^{\pi,U}(Q) = \emptyset$.*

Assumption B is a seemingly technical assumption that nonetheless has a simple interpretation:[A10] for any admissible signal structure Q, there is an equivalent admissible signal structure Q', in terms of the distribution of actions it induces the DM to take, where each action is induced by at most one signal. ϱ can be seen as a "standard" prescription given by the signal space: if a DM receives signal m_j according to Q', then she should take action $a_{\varrho(j)}$. Following Kamenica and

[A10] Note that B is an assumption on $\tilde{\mathscr{Q}}$, and $\tilde{\mathscr{Q}}$ can be described as the set where C is finite. Therefore, B is actually an assumption on the cost function C.

Gentzkow (2011), we call such Q' *straightforward*.[A11] A particularly salient example for ϱ in the case that $M = A$ is the identity mapping, in which case under a straightforward signal structure, the signal the DM receives is literally the action she should take.

Assumptions A and B need not be onerous. In particular, they are trivially satisfied whenever $\tilde{\mathscr{Q}} = \mathscr{Q}$.

Assumption C. ***Continuity.*** *$C(\pi, Q)$ is continuous in its second argument on $\tilde{\mathscr{Q}}$.*

Continuity is a typical assumption in much of economic analysis. In this case, it implies that gathering a small amount of additional information increases the total cost of information by only a small amount. This may seem like an innocuous assumption, but it precludes some plausible cost functions, such as those with fixed costs for information acquisition, as will be seen in Section 3.

Assumption D. ***Almost strict convexity.*** *$\forall\, \pi \in \Delta(\Theta), \forall\, \lambda \in (0,1), \forall\, Q_1, Q_2 \in \tilde{\mathscr{Q}}$, $C(\pi, \lambda Q_1 + (1-\lambda)Q_2) \leq \lambda C(\pi, Q_1) + (1-\lambda)C(\pi, Q_2)$, where the inequality is strict except possibly if Q_1 and Q_2 induce the same distribution of posteriors.*

This notion of convexity can be contrasted with CD15's. CD15 define a notion of convexity over the space of distributions of posteriors called "mixture feasibility";[A12] however, it has no empirical content, because distributions of posteriors and mixtures thereof are observationally equivalent given choice data. In our framework, cost functions are defined over signal structures instead of the distributions of posteriors they induce. Since the space of stochastic matrices can be identified with a subset of Euclidean space, Assumption D gives us an easily interpretable "geometric" notion of convexity. Moreover, taken in tandem with the other assumptions, Assumption D actually has empirical content, as shown in Proposition A2, which generalizes Proposition 3 in the paper.

In order to ensure that continuity and almost strict convexity imply continuous gross payoffs, we require one additional condition.

Assumption E. ***Monotonicity of information.*** *Let R be a right-stochastic matrix of dimension $|M| \times |M|$, which we refer to as a* garbling matrix. *Then for any $\pi \in \Delta(\Theta)$ and $Q \in \tilde{\mathscr{Q}}$, $C(\pi, Q) \geq C(\pi, QR)$, provided $QR \in \tilde{\mathscr{Q}}$.*

[A11] Straightforwardness is also related to the concept of the "revealed information structure" in CD15.

[A12] Assumption D involves mixtures of conditional signal probabilities, which could yield posteriors not generated by either information structure in the mixture, whereas mixture feasibility involves mixtures of distributions of posteriors whose support is the union of the supports of the distributions in the mixture.

If Q is an information structure and R is a garbling matrix, then Q can be thought of as containing all the information contained in QR, i.e. QR simply adds noise to Q. In this case, we shall say that Q *Blackwell-dominates* QR. As shown by Blackwell (1953), Q yields a (weakly) higher gross payoff than QR for any decision problem, given an optimal selection of decision matrices. Therefore, Assumption E implies that if one information structure is more informative than another, then it is also costlier.[A13]

Assumption E is equivalent to Condition K1 of CD15, and they show that it is not testable; any stochastic choice data set that is consistent with some cost function C is also consistent with some cost function $\tilde{C}$ that satisfies Assumption E. Therefore, requiring it does not eliminate any additional sets of stochastic choice data from being consistent with a model of rational inattention.

We have now established a set of sufficient conditions that ensure that the DM's ex-ante gross payoff is continuous in incentives.

Proposition A2. *Suppose that π is fixed and C satisfies Assumptions A, B, C, D, and E. Then the ex-ante gross payoff is continuous in U.*

Proof. First we show that we may assume a fixed decision matrix D' such that:

- If $i \in \varrho^{-1}(\{1, \ldots, |M|\})$, then $d'_{i,\varrho(i)} = 1$.
- If $i \in \varrho^{-1}(0)$, then $d'_{i,1} = 1$.

By Assumption A, $Q \in \mathscr{Q} \setminus \tilde{\mathscr{Q}}$ will never be optimal and can be ignored. Given $Q \in \tilde{\mathscr{Q}}$ and $U \in \mathscr{U}$, Assumption B tells us that there is a $Q' \in \tilde{\mathscr{Q}}$ constructed by summing columns of Q such that if $a_{\varrho(j)}$ is an action optimally induced by the signals in $\left\{m_k \middle| k \in V_j^{\pi,U}(Q)\right\}$ under Q, then $a_{\varrho(j)}$ is also an action optimally induced by the signal m_j under Q'. Therefore, in constructing D', it is optimal for the DM to set $d'_{i,\varrho(i)} = 1$ if $i \in \varrho^{-1}(\{1, \ldots, |M|\})$.

By Assumption B, if $i \in \varrho^{-1}(0)$, then $q_{\bullet,i} = \vec{0}$, and the signal m_i is never used by the DM. Therefore, any assumption can be made about the action taken under this zero-probability event, and we may set $d'_{i,1} = 1$.

Thus, D' constructed in this manner is in $\operatorname{argmax}_D \operatorname{tr}(Q'DU\Pi)$. Moreover, by construction, for any $\theta \in \Theta$, $\Pr^{\pi}_{Q}(m_j|\theta) = \Pr^{\pi}_{Q'}\left(V_j^{\pi,U}(Q)\middle|\theta\right)$, so $\operatorname{tr}(Q'D'U\Pi) = \max_D \operatorname{tr}(QDU\Pi)$.

[A13] This assumption does not provide a complete order on information costs, since it is possible that two experiments are not ranked in the Blackwell sense. In other words, if Q_1 and Q_2 are information structures of the same dimension, there does not necessarily exist R of appropriate dimension such that $Q_1R = Q_2$ or $Q_2R = Q_1$.

Now we must show that we can restrict our focus to Q' constructed in this manner. Note that by construction, $Q' = QP$, where P is square and has entries $p_{i,j}$ such that $p_{i,j} = 1$ if $i \in V_j^{\pi,U}(Q)$ and $p_{i,j} = 0$ otherwise.[A14] Therefore, it is right-stochastic, and by Assumption E, $C(\pi, Q') \leq C(\pi, Q)$. Therefore, we need only consider information structures such that the optimal decision matrix is fixed as D'.

We may now simply consider the problem:

$$\max_{Q \in \tilde{\mathscr{Q}}} \operatorname{tr}(QD'U\Pi) - C(\pi, Q) \tag{A5}$$

Denote the maximand in (A5) by $F(Q)$. As a consequence of Assumption C, $F(Q)$ is continuous in Q and U. By Assumption A and the Heine-Borel theorem, $\tilde{\mathscr{Q}}$ is compact. Therefore, by the maximum theorem, the optimal choice of information structure for each payoff matrix, $Q^*(U)$, is upper hemicontinuous in U.

Since the first term of $F(Q)$ is linear and the second is almost strictly convex by Assumption D, it inherits its convexity properties from the second term. In other words, $F(Q)$ is almost strictly concave, with almost strict concavity defined analogously to almost strict convexity. For each U, either $Q^*(U)$ is unique or it is multivalued. Suppose it is multivalued, and $Q_1^*, Q_2^* \in Q^*(U)$. Then $F(Q_1^*) = F(Q_2^*)$. If Q_1^* and Q_2^* induce different distributions of posteriors, then $\forall\, \lambda \in (0,1), F(\lambda Q_1^* + (1-\lambda)Q_2^*) > \lambda F(Q_1^*) + (1-\lambda)F(Q_2^*) = F(Q_1^*)$, contradicting the optimality of Q_1^* and Q_2^*.

Now suppose that Q_1^* and Q_2^* induce the same distribution of posteriors and therefore induce the same gross payoffs. Then either $\nexists\, \lambda \in (0,1)$ such that $F(\lambda Q_1^* + (1-\lambda)Q_2^*) = \lambda F(Q_1^*) + (1-\lambda)F(Q_2^*)$, in which case the argument of the preceding paragraph applies, or else there does exist such λ, in which case $Q^*(U) \ni Q_\lambda := \lambda Q_1^* + (1-\lambda)Q_2^*$ as well. Then, by the linearity of the trace function and the fact that Q_1^* and Q_2^* induce the same distribution of posteriors, $\operatorname{tr}(Q_\lambda D'U\Pi) = \operatorname{tr}(Q_1^* D'U\Pi) = \operatorname{tr}(Q_2^* D'U\Pi)$.

This implies that $\operatorname{tr}(Q^*(U)D'U\Pi)$ is single-valued, and since it is the composition of a continuous function (which can be viewed as an upper hemicontinuous correspondence) with an upper hemicontinuous correspondence, it is itself upper hemicontinuous (cf. Theorem 14.1.5 of Sydsæter

[A14] P can be seen as the matrix that takes column i of Q to column j of Q' if $p_{i,j} = 1$.

et al., 2008). Together, its upper hemicontinuity and single-valuedness imply that it is a continuous function of U, thereby completing the proof. □

To summarize the proof, straightforwardness and monotonicity of information ensure that the decision matrix chosen by the DM can be fixed, which in turn ensures the convexity of the problem. While almost strict convexity does not ensure a unique solution to the problem, it does ensure that the optimal ex-ante gross payoff is single-valued, which together with the continuity of the cost function implies the result.[A15]

The assumptions necessary for Proposition A2 are satisfied by many different cost functions. For example, it is easily shown that cost functions that can be expressed as a sum of strictly convex functions of the entries of a stochastic matrix are almost strictly convex.

A1.5 Uniform Guess Tasks

The general framework presented above can be applied to the uniform guess tasks introduced in Section 2. In these decision problems, since the DM is trying to determine the true state, we can set $A = \Theta$. Moreover, $U = rI_n$ for some $r > 0$. Therefore, the DM's ex-ante gross payoff in this task can be written as $r\text{tr}(QD\Pi)$. Since the prior is uniform, $P := \text{tr}(QD\Pi) = \frac{1}{n}\text{tr}(QD)$ is the ex-ante probability of correctly guessing the state, or their performance, and the ex-ante gross payoff can be written as $\frac{r}{n}\text{tr}(QD)$.

In Section 2, we defined costs as depending on the performance P rather than the entire information structure. We can equivalently define costs on the entire information structure as follows. Suppose that $|\Theta| = |M|$. Then Q is square, and we can impose that $C(\pi, Q) = \check{C}\left(\sum_{i=1}^{n} \pi_i q_{i,i}\right)$ if Q is such that the diagonal entries of ΠQ are maximal in their columns,[A16] and the off-diagonal entries of Q are fixed fractions of the "remaining" probability in each row.[A17] The DM's optimal

[A15] At this point, a clarification is in order. Proposition A2 is a statement about what the properties of an information cost function imply about behavior. To obtain a statement about what behavior implies about the properties of cost functions, we invoke the contrapositive: if gross payoffs are discontinuous in incentives, then this behavior cannot be rationalized by an information cost function that satisfies Assumptions C, D, and E simultaneously. However, as we explained earlier in this subsection, Assumption E is not testable. Therefore, given stochastic choice data, we can assume the cost function that rationalizes it satisfies Assumption E, and so if we observe that ex-ante gross payoffs are discontinuous in incentives, then this implies that the DM's cost function either is discontinuous or fails almost strict convexity.

[A16] It can be shown that this implies $\sum_{i=1}^{n} \pi_i q_{i,i}$ is at least $\frac{1}{n}$.

[A17] In an experimental setting these fixed fractions could be estimated from a decision-maker's distribution of sub-optimal choices, i.e. mistakes.

D is then simply the identity matrix, and the argument of $\check{C}(\cdot)$ is equal to their performance P, and the DM's maximand can be expressed as $rP - \check{C}(P)$.

Stated more formally, fix Ω, an $n \times n$ stochastic matrix with zeroes on its diagonal and entries $\omega_{i,j}$. Then, set $\tilde{\mathscr{Q}} = \{Q \,|\, j \in \text{argmax}_i \pi_i q_{i,j} \ \forall\, j \ \text{ and } \ q_{i,j} = \omega_{i,j}(1 - q_{i,i}) \text{ for } i \neq j\}$, with $C(\pi, Q) = \check{C}\left(\sum_{i=1}^{n} \pi_i q_{i,i}\right)$ on $\tilde{\mathscr{Q}}$ and positive infinity otherwise. It can can shown that if $\check{C}$ is well-behaved, then the performance function is continuous, thereby proving Proposition 3. We provide the relevant proof in Appendix Subsection A2.3.

It should also be noted that while Tsallis entropy costs (including mutual information) can be written as depending directly on performance in uniform guess tasks, we can equivalently work with the more general definition given in Subsection 3.2. We illustrate this in the proof of Proposition 5 given in Appendix Subsection A2.5.

A1.6 Posterior-Equivalent Information Structures

This appendix subsection formalizes the relationship between information structures defined as conditional likelihoods on a signal space and information structures defined as distributions of posterior beliefs, which allows us to clarify the relationship between the present framework and that of CD15.

The framework outlined in the preceding subsections defines costs jointly on the DM's prior belief and information structures as conditional distributions of signals, given states. Defining information structures in this manner is the approach taken by McGuire (1972) and Leshno and Spector (1992), among others. From the ex-ante perspective (i.e. before signals are realized), each information structure corresponds to a distribution of posterior beliefs; each potential signal has a posterior belief associated with it, and the likelihood of each of these posterior beliefs is the likelihood of receiving the signal associated with it. If π is a prior belief on Θ and Q is an information structure that generates signals in M, then the distribution of posteriors γ_Q^π is defined by:

$$\Pr_Q^\pi(x) = \sum_{j \in \left\{\ell \,\middle|\, \exists\, \alpha \in \mathbb{R}_{>0} \text{ s.t. } \pi \circ q_{\bullet,\ell} = \alpha x\right\}} \sum_{i=1}^{|\Theta|} \pi_i q_{i,j} \tag{A6}$$

where $\Pr_Q^\pi(\cdot)$ denotes the probability of its argument, given prior π and information structure Q,

x is an element of $\Delta(\Theta)$, $\circ$ denotes the Hadamard (component-wise) matrix product, and empty sums are taken to be zero. Several authors, including Kamenica and Gentzkow (2011) and CD15, choose to work directly with these distributions of posteriors. In the case of CD15, information costs are defined on these distributions. From the perspective of pure Bayesian expected utility maximization (cf. Caplin and Martin, 2015), the two approaches are clearly equivalent in terms of the behaviors they imply.

From the perspective of rational inattention, however, when information structures have costs associated with them, this equivalence is less readily established. The cost of a particular distribution of posteriors may not just depend on the distribution itself, but also on *how* it was generated. Consider the following examples. Let $\Theta = \{X, Y\}$ and $M = \{x, y, z\}$, both indexed in those orders. Let $\pi = (0.5, 0.5)$, and let $Q_1 = \begin{pmatrix} 0.8 & 0.2 & 0 \\ 0.2 & 0.8 & 0 \end{pmatrix}$, $Q_2 = \begin{pmatrix} 0.2 & 0.8 & 0 \\ 0.8 & 0.2 & 0 \end{pmatrix}$, and $Q_3 = \begin{pmatrix} 0.8 & 0.1 & 0.1 \\ 0.2 & 0.4 & 0.4 \end{pmatrix}$. It is easily verified that each of these information structures generates the same distribution of posteriors. Under Q_1, x was most likely generated by X, and y was most likely generated by Y. Thus, the signals can be seen as a "natural" interpretation of the states. By contrast, under Q_2, x was most likely generated by Y, and y was most likely generated by X. This interpretation is "unnatural" and consequently may be more mentally costly for a DM to process. Now, consider Q_3, where again x was most likely generated by X, and y was most likely generated by Y, but there is also a third signal z generated with positive probability that is most likely to have been generated by Y. Though y and z both correspond to the same posterior belief, having to keep track of three signals may be more mentally taxing than keeping track of two, and so Q_3 might be costlier than Q_1. These examples serve to illustrate that assigning signals to states is not merely a matter of indexing when considering information costs; costs may depend on the interpretability of and meaning implied by an information structure.

This of course raises the question of whether our model implies the potential to accommodate behaviors that would not be feasible under the posterior-based approach to rational inattention. That is the problem to which we turn our attention in this appendix. Before proceeding, we require some preliminaries.

A1.6.1 Preliminaries

Let Q be a stochastic matrix. Denote its entries by $q_{i,j}$, its rows by $q_{i,\bullet}$, and its columns by $q_{\bullet,j}$. We define three operations on Q:

1. Q' is obtained from Q by *swapping* if $q'_{\bullet,j} = q_{\bullet,k}$, $q'_{\bullet,k} = q_{\bullet,j}$, and all other columns are the same.

2. Q' is obtained from Q by *summing* if for some j, k such that $q_{\bullet,j} = \alpha q_{\bullet,k}$ for some $\alpha > 0$, $q'_{\bullet,j} = q_{\bullet,j} + q_{\bullet,k}$, $q'_{\bullet,k}$ is a column of zeroes, and all other columns are the same.

3. Q' is obtained from Q by *splitting* if $\exists\, k$ and $\lambda \in (0,1)$ such that $q_{\bullet,k}$ is a column of zeroes, $q'_{\bullet,j} = \lambda q_{\bullet,j}$, $q'_{\bullet,k} = (1-\lambda) q_{\bullet,j}$, and all other columns are the same.

Note that each of these operations is reversible as one of the other operations. Swapping columns can be reversed by simply swapping the columns again. Summing columns can be reversed by splitting the summed column into the summands. Splitting columns can be reversed by summing the split columns.

Finally, let $\Diamond$ be a binary relation on the space of $|\Theta| \times |M|$ stochastic matrices, defined by $Q \Diamond R$ iff given some $\pi \in \text{int}(\Delta(\Theta))$ (i.e. π has full support on Θ),[A18] Q and R induce the same distribution of posteriors, i.e. $\gamma_Q^\pi = \gamma_R^\pi$. We will say Q and R are *posterior-equivalent* if $Q \Diamond R$.

A1.6.2 Posterior Equivalence and the Algebra of Stochastic Matrices

The posterior equivalence relation defined in the previous subsection is independent of the prior; if two information structures are posterior-equivalent for some prior with full support, then the posterior equivalence condition holds for all priors with full support. In other words, as the following proposition shows, two information structures could be said to be posterior-equivalent if they induce the same distribution of posteriors for *any* prior with full support on the state space.

Proposition A3. *If $Q \Diamond R$, then $\gamma_Q^\pi = \gamma_R^\pi \; \forall\, \pi \in \text{int}(\Delta(\Theta))$.*

[A18] We require that π have full support, because the probability distribution of signals conditional on a zero-probability state is irrelevant for determining the distribution of posteriors and can therefore be chosen arbitrarily.

Proof. (A6) can be rewritten by rearranging the order of summation:

$$\Pr_Q^\pi(x) = \sum_{i=1}^{|\Theta|} \pi_i \left(\sum_{j \in \left\{ \ell \middle| \exists\, \alpha \in \mathbb{R}_{>0} \text{ s.t. } \pi \circ q_{\bullet,\ell} = \alpha x \right\}} q_{i,j} \right) \tag{A7}$$

Now suppose $\bar{\pi}$ is a prior that generates posterior equivalence between Q and R, i.e. $\gamma_Q^{\bar{\pi}} = \gamma_R^{\bar{\pi}}$. Then, for a given posterior $\bar{x}$, we can write:

$$\sum_{i=1}^{|\Theta|} \bar{\pi}_i \left(\sum_{j \in \left\{ \ell \middle| \exists\, \alpha \in \mathbb{R}_{>0} \text{ s.t. } \pi \circ q_{\bullet,\ell} = \alpha \bar{x} \right\}} q_{i,j} \right) = \sum_{i=1}^{|\Theta|} \bar{\pi}_i \left(\sum_{j \in \left\{ \ell \middle| \exists\, \alpha \in \mathbb{R}_{>0} \text{ s.t. } \pi \circ q_{\bullet,\ell} = \alpha \bar{x} \right\}} r_{i,j} \right) \tag{A8}$$

Since the summands in the inner sums are entries from columns of Q and R whose Hadamard products with π are multiples of $\bar{x}$, the respective columns must be multiples of each other. To see this, suppose that q is one such column of Q and r is one such column of R. Then we can write $\pi \circ q = \alpha \bar{x}$ and $\pi \circ r = \beta \bar{x}$ for some $\alpha, \beta > 0$. Equivalently, we can write $\Pi q = \alpha \bar{x}$ and $\Pi q = \beta \bar{x}$, where $\Pi = \operatorname{diag}(\pi)$. Since Π is a diagonal matrix with strictly positive entries on its diagonal, it is invertible, and we can write $q = \Pi^{-1}(\alpha \bar{x})$ and $r = \Pi^{-1}(\beta \bar{x})$. By the linearity of Π^{-1}, we can write $q = \alpha \Pi^{-1} \bar{x}$ and $r = \beta \Pi^{-1} \bar{x}$, which implies that $q = \frac{\alpha}{\beta} r$.

Therefore, we can write:

$$\sum_{j \in \left\{ \ell \middle| \exists\, \alpha \in \mathbb{R}_{>0} \text{ s.t. } \pi \circ q_{\bullet,\ell} = \alpha \bar{x} \right\}} q_{i,j} = \kappa_{\bar{x}} \left(\sum_{j \in \left\{ \ell \middle| \exists\, \alpha \in \mathbb{R}_{>0} \text{ s.t. } \pi \circ q_{\bullet,\ell} = \alpha \bar{x} \right\}} r_{i,j} \right) \tag{A9}$$

for all i and for some $\kappa_{\bar{x}} > 0$. Since $\bar{\pi}_i > 0 \;\forall\, i$, it must be that $\kappa_{\bar{x}} = 1$ or else (A8) could not hold. Now replace $\bar{\pi}$ in (A8) with an arbitrary prior with full support π. Because (A9) holds with $\kappa_{\bar{x}} = 1$, (A8) holds for arbitrary π. □

Corollary A1. *$\Diamond$ is an equivalence relation.*

Proof. Reflexivity and symmetry are trivially verified.

For transitivity, suppose that $Q \Diamond R$ and $R \Diamond S$. Then $\gamma_Q^\pi = \gamma_R^\pi$ and $\gamma_R^{\pi'} = \gamma_S^{\pi'}$ for some π, π' with full support. But by Proposition A3, $\gamma_S^\pi = \gamma_R^\pi = \gamma_Q^\pi$, which establishes the result. □

Now, since we know that posterior equivalence can be established without making reference

to any specific prior distribution, we can show that $Q \Diamond R$ is equivalent to a set of linear-algebraic conditions on the relationship between Q and R, written as stochastic matrices. Put differently, posterior-equivalence is not just a statistical relationship between stochastic matrices, but also an algebraic one that can be defined without reference to probabilities. To our knowledge, this is a novel characterization of equivalence of information structures.

Proposition A4. *$Q \Diamond R$ iff R can be obtained from Q by a sequence of swapping, summing, and splitting.*

Proof. We begin by proving the 'if' direction. We show that an information structure obtained from another by each of the three column operations in the proposition is posterior-equivalent to the original information structure.

Swapping. (A6) is unaffected by a change in the order of columns.

Summing. Suppose Q' is obtained from Q by summing columns $\bar{j}$ and $\bar{k}$. Let $\bar{y} = \frac{\pi \circ q_{\bullet,\bar{j}}}{\sum_{i=1}^{|\Theta|} \pi_i q_{i,\bar{j}}}$. Then:

$$\begin{aligned}
\Pr_Q^\pi(\bar{y}) &= \sum_{i=1}^{|\Theta|} \pi_i \left(\sum_{j \in \left\{ \ell \middle| \exists\, \alpha \in \mathbb{R}_{>0} \text{ s.t. } \pi \circ q_{\bullet,\ell} = \alpha \bar{y} \right\}} q_{i,j} \right) \\
&= \sum_{i=1}^{|\Theta|} \pi_i \left(\sum_{j \in \left\{ \ell \middle| \exists\, \alpha \in \mathbb{R}_{>0} \text{ s.t. } \pi \circ q_{\bullet,\ell} = \alpha \bar{y} \right\} \setminus \{\bar{j}, \bar{k}\}} q_{i,j} + q_{i,\bar{j}} + q_{i,\bar{k}} \right) \\
&= \sum_{i=1}^{|\Theta|} \pi_i \left(\sum_{j \in \left\{ \ell \middle| \exists\, \alpha \in \mathbb{R}_{>0} \text{ s.t. } \pi \circ q'_{\bullet,\ell} = \alpha \bar{y} \right\} \setminus \{\bar{j}, \bar{k}\}} q'_{i,j} + q'_{i,\bar{j}} \right) \\
&= \sum_{i=1}^{|\Theta|} \pi_i \left(\sum_{j \in \left\{ \ell \middle| \exists\, \alpha \in \mathbb{R}_{>0} \text{ s.t. } \pi \circ q'_{\bullet,\ell} = \alpha \bar{y} \right\}} q'_{i,j} \right) \\
&= \Pr_{Q'}^\pi(\bar{y})
\end{aligned}$$

Moreover, since the columns of Q that are not $\bar{j}$ or $\bar{k}$ are unaffected by summing, it is obvious that $\Pr_Q^\pi(x) = \Pr_{Q'}^\pi(x)\ \forall\, x \neq \bar{y}$ as well.

Splitting. Suppose Q' is obtained from Q by splitting column $\bar{j}$ into columns $\bar{j}$ and $\bar{k}$. Then

$q'_{\bullet,\bar{j}} = \lambda q_{\bullet,\bar{j}}$ and $q'_{\bullet,\bar{k}} = (1-\lambda) q_{\bullet,\bar{j}}$ for some $\lambda \in (0,1)$. Let $\bar{y} = \frac{\pi \circ q_{\bullet,\bar{j}}}{\sum_{i=1}^{|\Theta|} \pi_i q_{i,\bar{j}}}$. Then:

$$
\begin{aligned}
\Pr_Q^\pi(\bar{y}) &= \sum_{i=1}^{|\Theta|} \pi_i \left(\sum_{j \in \left\{ \ell \middle| \exists\, \alpha \in \mathbb{R}_{>0} \text{ s.t. } \pi \circ q_{\bullet,\ell} = \alpha \bar{y} \right\}} q_{i,j} \right) \\
&= \sum_{i=1}^{|\Theta|} \pi_i \left(\sum_{j \in \left\{ \ell \middle| \exists\, \alpha \in \mathbb{R}_{>0} \text{ s.t. } \pi \circ q_{\bullet,\ell} = \alpha \bar{y} \right\} \setminus \{\bar{j}\}} q_{i,j} + q_{i,\bar{j}} \right) \\
&= \sum_{i=1}^{|\Theta|} \pi_i \left(\sum_{j \in \left\{ \ell \middle| \exists\, \alpha \in \mathbb{R}_{>0} \text{ s.t. } \pi \circ q_{\bullet,\ell} = \alpha \bar{y} \right\} \setminus \{\bar{j}\}} q_{i,j} + \lambda q_{i,\bar{j}} + (1-\lambda) q_{i,\bar{j}} \right) \\
&= \sum_{i=1}^{|\Theta|} \pi_i \left(\sum_{j \in \left\{ \ell \middle| \exists\, \alpha \in \mathbb{R}_{>0} \text{ s.t. } \pi \circ q'_{\bullet,\ell} = \alpha \bar{y} \right\} \setminus \{\bar{j},\bar{k}\}} q'_{i,j} + q'_{i,\bar{j}} + q'_{i,\bar{k}} \right) \\
&= \sum_{i=1}^{|\Theta|} \pi_i \left(\sum_{j \in \left\{ \ell \middle| \exists\, \alpha \in \mathbb{R}_{>0} \text{ s.t. } \pi \circ q'_{\bullet,\ell} = \alpha \bar{y} \right\}} q'_{i,j} \right) \\
&= \Pr_{Q'}^\pi(\bar{y})
\end{aligned}
$$

Moreover, since the columns of Q that are not $\bar{j}$ or $\bar{k}$ are unaffected by splitting, it is obvious that $\Pr_Q^\pi(x) = \Pr_{Q'}^\pi(x)\ \forall\, x \neq \bar{y}$ as well.

This shows that $Q' \lozenge Q$ if Q' is obtained from Q by any one of the three column operations. Since $\lozenge$ is transitive, it is therefore true that $Q' \lozenge Q$ if Q' is obtained from Q by a sequence of the three column operations. This concludes the proof of the 'if' direction.

For the 'only if' direction, suppose that $Q \lozenge R$. Select the leftmost column of Q that is not a column of zeroes. Sum to it the next leftmost column that is a multiple of it. Repeat until no more multiples remain. Then repeat this summing process with the next leftmost non-zero column until all non-zero columns have been exhausted. Call the matrix resulting from this sequence of summings Q'. Do the same with R, and call the matrix resulting from this sequence of summings R'.

We must now show that Q' and R' have the same columns. Since $\lozenge$ is an equivalence relation, and as we showed above, column operations preserve the relation, it must be that $Q' \lozenge R'$. Therefore, they induce the same distribution of posteriors. Suppose $\bar{z}$ is a nonzero column of Q' that is not in R'. Then, since both Q' and R' were constructed so that none of their nonzero columns are

multiples of each other, $\frac{\bar{z}}{\|\bar{z}\|_1} \not\in \text{Supp}\left(\gamma^{\pi}_{R'}\right)$, where $\|\cdot\|_1$ denotes the ℓ_1 norm. Thus we have shown by contradiction that each nonzero column of R' must be a multiple of only one column in Q', and vice versa. Now select a column $\bar{y}$ of Q', and consider its multiple $\bar{y}^\star$ in R', where $\bar{y}^\star = \alpha\bar{y}$ for some $\alpha > 0$. Then for some full-support prior π, $\Pr^{\pi}_{Q'}\left(\frac{\bar{y}}{\|\bar{y}\|_1}\right) = \sum_{i=1}^{|\Theta|} \pi_i \bar{y}_i$, and since $\frac{\bar{y}}{\|\bar{y}\|_1} = \frac{\bar{y}^\star}{\|\bar{y}^\star\|_1}$, $\Pr^{\pi}_{R'}\left(\frac{\bar{y}}{\|\bar{y}\|_1}\right) = \sum_{i=1}^{|\Theta|} \pi_i \bar{y}^\star_i = \alpha \sum_{i=1}^{|\Theta|} \pi_i \bar{y}_i$. Since $Q' \lozenge R'$, it must be that $\alpha = 1$. This shows that Q' and R' have the same nonzero columns. Since they have the same dimensions, and they are both stochastic matrices (so that summing over their nonzero columns yields a vector of ones), this must mean that they have the same number of nonzero columns.

Given that Q' and R' have the same columns, it can be shown that one can be obtained from the other by a sequence of swappings. Select the leftmost column of Q' and swap it with the column that has that position in R'. Repeat this process with the next leftmost column until Q' has been transformed into R'. Now, note that the sequence of summings that took R to R' can be reversed to become a sequence of splittings that takes R' to R. Concatenating the sequence of summings that took Q to Q', the sequence of swappings that took Q' to R', and the sequence of splittings that took R' to R gives a sequence of summings, swappings, and splittings that takes Q to R. This concludes the proof. □

A1.6.3 Cost Equivalence

We are now prepared to see whether our framework can predict different behavior than that of CD15. Assuming a finite set of actions, allowing costs to depend on how distributions of posteriors are generated generalizes CD15; put differently, a version of their model with a finite number of actions is equivalent to ours with the following assumption.

Assumption F. ***Cost equivalence.*** *For all priors π, $C(\pi, Q_1) = C(\pi, Q_2)$ whenever $Q_1 \lozenge Q_2$.*

However, as we show below, any behavior that can be rationalized by our model can also be rationalized by CD15; cost equivalence imposes no additional behavioral restrictions, and it is therefore untestable. This result allows us to apply CD15's necessary and sufficient conditions for rational inattention to our framework without imposing any additional conditions.

Proposition A5. *Stochastic choice data are consistent with* (A2) *iff they are consistent with CD15.*[A19]

[A19] Though we have assumed a finite action space in our paper, the proof of Proposition 1 does not rely on this.

To outline the proof, the 'if' direction is obvious, since our model generalizes CD15. To see the 'only if' direction, suppose that $\{(U_i, \theta_i, a_i)\}$ can be rationalized by (A2) with some cost function $C(\pi, Q)$. Define $\mathscr{Q}_{\gamma_Q^\pi}$ to be set of information structures that induce the distribution γ_Q^π over posteriors, and define $\tilde{C}(\pi, Q) := \min_{R \in \mathscr{Q}_{\gamma_Q^\pi}} C(\pi, R)$. It is obvious that $\tilde{C}$ satisfies cost equivalence. Moreover, since a given posterior distribution always induces the same ex-ante gross payoff, the DM should always choose the lowest-cost way of inducing that posterior distribution. Thus, the proof boils down to showing that this minimum is well-defined. Details are below.

Proof. The 'if' direction is obvious, since our model generalizes CD15 with finite action sets.

The 'only if' direction can be seen as follows. Suppose $\{(U_i, \theta_i, a_i)\}$ can be rationalized by (A2) with some cost function $C(\pi, Q)$. Since there are finitely many decision problems, C is pinned down for a finite set of points (i.e. a closed set), and so by the Tietze extension theorem (cf. Rudin, 1974, pg. 422), it may be assumed continuous. Define $\mathscr{Q}_{\gamma_Q^\pi}$ to be set of information structures that induce posterior γ_Q^π, and define $\tilde{C}(\pi, Q) := \min_{R \in \mathscr{Q}_{\gamma_Q^\pi}} C(\pi, R)$, assuming it is well-defined. γ_Q^π always induces the same maximum gross payoff, no matter which information structure in $\mathscr{Q}_{\gamma_Q^\pi}$ generated it. Therefore, since the DM is a payoff maximizer, for each distribution of posteriors she generates, she will always select the lowest-cost method of doing so. This implies that behavior that can be rationalized by C can also be rationalized by $\tilde{C}$, which obviously satisfies cost equivalence.

Now we must verify that $\tilde{C}$ is actually well-defined. Let $b_\pi : \mathscr{Q} \longrightarrow \Delta(\Delta(\Theta))$ be the function that maps an information structure to the distribution of posteriors it induces given prior π. First, we must show that b_π is continuous when $\Delta(\Delta(\Theta))$ is equipped with the weak-* topology, i.e. the topology of weak convergence of measure.

By Bayes' rule, $\text{Supp}(b_\pi(Q)) = \left\{ \left(\frac{\pi_s q_{s,k}}{\sum_{l=1}^n \pi_l q_{l,k}} \right)_{s=1}^n \middle| k \in \{1, \dots |M|\}, \sum_{l=1}^n \pi_l q_{l,k} > 0 \right\}$, and each element $\zeta \in \text{Supp}(b_\pi(Q))$ is induced with probability $\sum_{k \in Q^\zeta} \sum_{l=1}^n \pi_l q_{l,k}$, where Q^ζ is the set of columns of Q that generate the posterior ζ.

Consider a sequence of information structures $Q_1, Q_2, \dots \in \mathscr{Q}$ converging to Q. We must show that $\lim_{j \to \infty} b_\pi(Q_j) = b_\pi(Q)$ (in the sense of weak convergence of measure). By Theorem 25.8 of Billingsley (1995), this is equivalent to showing that $\lim_{j \to \infty} b_\pi(Q_j)(X) = b_\pi(Q)(X)$ for all continuity

Therefore, the use of decision matrices mapping signals to actions can be seen as a notational convenience for the applications contained in this paper rather than a fundamental part of the model.

sets X in the Borel σ-algebra of $\Delta(\Theta)$.[A20]

Since X is a continuity set, $\partial X \cap \text{Supp}(b_\pi(Q)) = \emptyset$. There are two cases. Either $X \cap \text{Supp}(b_\pi(Q)) = \emptyset$ or $\text{int}(X) \cap \text{Supp}(b_\pi(Q)) \neq \emptyset$.

Case 1: $X \cap \text{Supp}(b_\pi(Q)) = \emptyset$. If $\exists\, J \in \mathbb{N}$ such that $b_\pi(Q_j)(X) = 0\ \forall\, j > J$, then clearly $\lim_{j\to\infty} b_\pi(Q_j)(X) = b_\pi(Q)(X) = 0$. If not, then $\forall\, J \in \mathbb{N}$, $\exists\, j > J$ such that $X \cap \text{Supp}(b_\pi(Q_j)) \neq \emptyset$. Suppose, for a contradiction, that $\lim_{j\to\infty} b_\pi(Q_j)(X) \neq 0$. Then $\exists\, \varepsilon > 0$ such that $\forall\, J \in \mathbb{N}$, $\exists\, j > J$ such that $b_\pi(Q_j)(X) > \varepsilon$. Therefore, there must exist a subsequence Q_{j_h} such that $\left(\left(\frac{\pi_s q_{s,k}}{\sum_{l=1}^n \pi_l q_{l,k}}\right)_{s=1}^n\right)_{j_h}$ converges in cl(X) for some k.[A21] If it converges to a point in $\text{int}(X)$, then this contradicts the fact that $b_\pi(Q)(X) = 0$. If it converges to a point in ∂X, then $b_\pi(Q)(\partial X) > 0$, contradicting the fact that X is a continuity set. Thus, $\lim_{j\to\infty} b_\pi(Q_j)(X) = b_\pi(Q)(X)$.

Case 2: $\text{int}(X) \cap \text{Supp}(b_\pi(Q)) \neq \emptyset$. Note that since (Q_j) is a convergent sequence, each entry of the matrices in (Q_j) also defines a convergent sequence. Then each $((z_k)_j) := ((\sum_{l=1}^n \pi_l q_{l,k})_j)$ is a convergent sequence with limit z_k, and each $((y_k)_j) := \left(\left(\left(\frac{\pi_s q_{s,k}}{\sum_{l=1}^n \pi_l q_{l,k}}\right)_{s=1}^n\right)_j\right)$ either converges to some limit y_k (for $z_k > 0$) or else has an undefined limit (when $z_k = 0$).[A22] Since they are continuous functions of the entries of π and (Q_j), and because (Q_j) is convergent, $y_k = \left(\frac{\pi_s q_{s,k}}{\sum_{l=1}^n \pi_l q_{l,k}}\right)$ (when it exists) and $z_k = \sum_{l=1}^n \pi_l q_{l,k}$, where the entries $q_{l,k}$ are taken from Q. Consider the set $K \subseteq \{1, \ldots, M\}$ such that $\{((y_k)_j) | k \in K\}$ is the collection of sequences that converge to points in $\text{int}(X)$. Then, because $\text{int}(X)$ is open, $\forall\, \varepsilon > 0$ and $\forall\, k \in K$, $\exists\, N_k$ such that $\forall\, j > N_k, (y_k)_j \in \text{int}(X)$. Let $\bar{N} = \max_{k\in K} N_k$. Then $\forall\, j > \bar{N}, b_\pi(Q_j)(X) \geq (\sum_{k\in K}(\sum_{l=1}^n \pi_l q_{l,k}))_j$.

We now show that $b_\pi(Q_j)(X) - \sum_{k\in K}(\sum_{l=1}^n \pi_l q_{l,k})_j$ goes to zero as j grows large. Suppose there does not exist $J \in \mathbb{N}$ such that this sequence has the value $0\ \forall\, j > J$. Then there must exist a subsequence (Q_{j_h}) such that $b_\pi(Q_{j_h})(X) - \sum_{k\in K}(\sum_{l=1}^n \pi_l q_{l,k})_{j_h} > 0$ for all j_h. Then for each j_h, there is some $k' \not\in K$ such that $(y_{k'})_{j_h} \in \text{Supp}(b_\pi(Q_{j_h}))$. Because $|M|$ is finite, we may assume that this k' is fixed. If $((y_{k'})_{j_h})$ is convergent, it must converge in $\text{cl}(X)$. If $y_{k'} \in \text{int}(X)$, then this contradicts the fact that $k \not\in K$. If $y_{k'} \in \partial X$, then this contradicts the fact that X is a continuity set. If $((y_{k'})_{j_h})$

[A20] A continuity set is a set X whose boundary ∂X has measure zero.

[A21] We can take k fixed here because even if we construct a subsequence where the sequence of posteriors is constructed by different columns of Q_{j_h} for different sequence elements, we can merely take a subsequence of that subsequence, but with k fixed.

[A22] It is possible that $(y_k)_{j'}$ maybe be undefined for some k and j'. This occurs when $(z_k)_{j'} = 0$. If there are finitely many such j', then we can simply consider a sequence (Q_j) with these j' removed. If there are infinitely many such j', then $(z_k)_j$ must converge to zero. Therefore, WLOG, either (Q_j) is such that $(z_k)_j \neq 0\ \forall\, j, k$ and possibly converges to zero, or $(z_k)_j$ definitely converges to zero.

has no defined limit, then $z_{k'} = 0$. Therefore, $\lim_{h\to\infty} \left[b_\pi(Q_{j_h})(X) - \sum_{k\in K}(\sum_{l=1}^n \pi_l q_{l,k})_{j_h}\right] = 0$.

This establishes the continuity of b_π. Therefore, for a given γ with finite support in $\Delta(\Delta(\Theta))$, $b_\pi^{-1}(\{\gamma\})$ is closed (since singletons are closed). Because $b_\pi^{-1}(\{\gamma\}) \subseteq \mathscr{Q}$ and $\mathscr{Q}$ is a bounded subset of $M_{n\times|M|}(\mathbb{R})$ (which can be identified with $\mathbb{R}^{n|M|}$), by the Heine-Borel theorem $b_\pi^{-1}(\{\gamma\})$ is compact.

In particular $\mathscr{Q}_{\gamma_Q^\pi}$ is compact, and since C is continuous (fixing π), by the Weierstrass theorem, it attains its minimum on $\mathscr{Q}_{\gamma_Q^\pi}$. Therefore, $\tilde{C}$ is well-defined. This concludes the proof. □

A2 Proofs

This appendix contains the proofs omitted from the paper.

A2.1 Proof of Proposition 1

Proof. We begin by proving the "only if" direction. Let $r_1 \geq r_2$ be two possible rewards. Let Q_i be the information structure optimally chosen under reward r_i, $i = 1, 2$. Let D_i^j be the decision matrix chosen under information structure Q_i and reward r_j, $i, j = 1, 2$. Since information structures are "observed" only up to the actions taken, WLOG, we can assume straightforwardness and take $D_i := D_i^i = D_i^{-i}$, $i = 1, 2$.

The NIAC condition gives us:

$$\begin{aligned} & r_1\mathrm{tr}(Q_1 D_1 \Pi) + r_2\mathrm{tr}(Q_2 D_2 \Pi) \geq r_2\mathrm{tr}(Q_1 D_1 \Pi) + r_1\mathrm{tr}(Q_2 D_2 \Pi) \\ \implies & r_1 P^*(r_1) + r_2 P^*(r_2) \geq r_2 P^*(r_1) + r_1 P^*(r_2) \\ \implies & (r_1 - r_2)[P^*(r_1) - P^*(r_2)] \geq 0 \end{aligned} \tag{A10}$$

Since $r_1 \geq r_2$, in order for (A10) to hold, we require that $P^*(r_1) \geq P^*(r_2)$. This proves the "only if" direction.

For the "if" direction, consider a set of reward levels $r_1 \geq r_2 \geq \ldots \geq r_N$ and associated performances $P_1 \geq P_2 \geq \ldots \geq P_N$, where $P_i := P^*(r_i)$. (We can order the performances in this manner since P^* is nondecreasing.)

Consider an assignment of performances to rewards $\left(r_i, P_{\sigma_1(i)}\right)_{i=1}^N$, where σ_1 is a cyclic permu-

tation. Let σ_2 be defined as follows:

$$\sigma_2(i) := \begin{cases} 1, & i = 1 \\ \sigma_1(1), & i = {\sigma_1}^{-1}(1) \\ \sigma_1(i), & \text{otherwise} \end{cases}$$

Now we compute the difference in total gross payoffs between the assignments defined by σ_2 and σ_1.

$$\begin{aligned} &\sum_{j=1}^{N} r_j P_{\sigma_2(i)} - \sum_{j=1}^{N} r_j P_{\sigma_1(i)} \\ &= r_1 P_1 + r_{\sigma_1{}^{-1}(i)} P_{\sigma_1(1)} - \left(r_1 P_{\sigma_1(1)} + r_{\sigma_1{}^{-1}(i)} P_1\right) \\ &= \left(r_1 - r_{\sigma_1{}^{-1}(i)}\right)\left(P_1 - P_{\sigma_1(1)}\right) \\ &\geq 0, \quad \text{since } r_1 \geq r_{\sigma_1{}^{-1}(i)} \text{ and } P_1 \geq P_{\sigma_1(1)} \end{aligned}$$

Now we repeat this process for $j \geq 2$, at each step constructing the permutation σ_{j+1} as follows:

$$\sigma_{j+1}(i) := \begin{cases} j, & i = j \\ \sigma_j(j), & i = {\sigma_j}^{-1}(j) \\ \sigma_j(i), & \text{otherwise} \end{cases}$$

By the preceding argument, the total gross payoffs to the assignment increase (weakly) at each step. Since there are N rewards, this process must finish in $N-1$ steps, ending with $\sigma_N(i) = i$ and the highest possible gross payoff. Since the initial assignment $(r_i, P_{\sigma_1(i)})_{i=1}^N$ was arbitrary, this implies the NIAC condition for our data. □

A2.2 Proof of Proposition 2

Proof. Fix some $x \in A$ and $y \in \Theta$. Then:

$$\begin{aligned} &\Pr(\theta = x | a = x) \geq \Pr(\theta = y | a = x) \\ \iff &r \Pr(\theta = x | a = x) + 0 \cdot \sum_{z \neq x} \Pr(\theta = z | a = x) \geq r \Pr(\theta = y | a = x) + 0 \cdot \sum_{z \neq y} \Pr(\theta = z | a = x) \end{aligned}$$

$$\iff \sum_{z\in\Theta} u(x,z)\Pr(\theta=z|a=x) \geq \sum_{z\in\Theta} u(y,z)\Pr(\theta=z|a=x)$$

$$\iff \sum_{z\in\Theta} u(x,z)\frac{\Pr(a=x|\theta=z)\Pr(\theta=z)}{\cancel{\Pr(a=x)}} \geq \sum_{z\in\Theta} u(y,z)\frac{\Pr(a=x|\theta=z)\Pr(\theta=z)}{\cancel{\Pr(a=x)}}$$

$$\iff u_{k,\bullet}\Pi Q^* d^*_{\bullet,k} \geq u_{l,\bullet}\Pi Q^* d^*_{\bullet,k}, \qquad \text{where } x \text{ and } y \text{ are the } k\text{-th and } l\text{-th elements of } \Theta, \text{ respectively}$$

The last implication holds because the (i,j)-th entry of Q^*D^* is $\Pr(a_j|\theta_i)$. Since all these implications are bidirectional, and x and y were chosen arbitrarily, this completes the proof. □

A2.3 Proof of Proposition 3

Proof. We must verify that the cost function $\acute{C}(\pi, Q)$ induced by $C(P)$ and the associated $\tilde{\mathscr{Q}}$ satisfy the assumptions A, B, C, D, and E.

Assumption A. Since each maximizer is chosen from a finite set, $\tilde{\mathscr{Q}}$ is nonempty.

To verify convexity, let $Q_1, Q_2 \in \tilde{\mathscr{Q}}$ with generic entries $q^1_{i,j}$ and $q^2_{i,j}$, let $\lambda \in (0,1)$, and let $Q_\lambda := \lambda Q_1 + (1-\lambda)Q_2$ with generic entry $q^\lambda_{i,j}$. Consider an off-diagonal entry $q^\lambda_{i,j}$ where $i \neq j$. Then $q^\lambda_{i,j} = \lambda q^1_{i,j} + (1-\lambda)q^2_{i,j} = \lambda\omega_{i,j}(1-q^1_{i,i}) + (1-\lambda)\omega_{i,j}(1-q^2_{i,i}) = \omega_{i,j}(\lambda(1-q^1_{i,i}) + (1-\lambda)(1-q^2_{i,i})) = \omega_{i,j}(1-q^\lambda_{i,i})$. For a diagonal entry $q^\lambda_{j,j}$, we have:

$$\begin{aligned}
&\pi_j q^\ell_{j,j} \geq \pi_i q^\ell_{i,j}, \quad \forall\, i \in \{1,\dots n\}, \ell \in \{1,2\} \\
\implies &\lambda\pi_j q^1_{j,j} \geq \lambda\pi_i q^1_{i,j} \text{ and } (1-\lambda)\pi_j q^2_{j,j} \geq (1-\lambda)\pi_i q^2_{i,j}, \quad \forall\, i \in \{1,\dots n\} \\
\implies &\lambda\pi_j q^1_{j,j} + (1-\lambda)\pi_j q^2_{j,j} \geq \lambda\pi_i q^1_{i,j} + (1-\lambda)\pi_k q^2_{k,j}, \quad \forall\, i,k \in \{1,\dots n\} \\
\implies &\pi_j(\lambda q^1_{j,j} + (1-\lambda)q^2_{j,j}) \geq \pi_i(\lambda q^1_{i,j} + (1-\lambda)q^2_{i,j}), \quad \forall\, i \in \{1,\dots n\} \\
\implies &\pi_j q^\lambda_{j,j} \geq \pi_i q^\lambda_{i,j}, \quad \forall\, i \in \{1,\dots n\}
\end{aligned}$$

This proves the convexity of $\tilde{\mathscr{Q}}$.

To verify closedness, let (Q^k) be a sequence in $\tilde{\mathscr{Q}}$, where an element of the sequence has generic entry $q^k_{i,j}$. For off-diagonal entries $(i \neq j)$, $q^k_{i,j} = \omega_{i,j}(1-q^k_{i,i})\ \forall\, k$ implies $\lim_{k\to\infty} q^k_{i,j} = \omega_{i,j}(1-\lim_{k\to\infty} q^k_{i,i})$. For diagonal entries, $q^k_{j,j} \geq q^k_{i,j}\ \forall\, i \in \{1,\dots,n\}, \forall\, k$ implies $\lim_{k\to\infty} q^k_{j,j} \geq \lim_{k\to\infty} q^k_{i,j}\ \forall\, i \in \{1,\dots,n\}$.

Assumption B. $\tilde{\mathscr{Q}}$ is constructed so that every $Q \in \tilde{\mathscr{Q}}$ is straightforward.

Assumption C. The continuity of $\acute{C}$ follows from the well-behavedness of C.

Assumption D. We now verify the almost-strict convexity of $\acute{C}$, using the same notation as earlier in the proof.

$$
\begin{aligned}
&\lambda \acute{C}(\pi, Q_1) + (1-\lambda)\acute{C}(\pi, Q_2) \\
&= \lambda C\left(\sum_{i=1}^{n} \pi_i q^1_{i,i}\right) + (1-\lambda) C\left(\sum_{i=1}^{n} \pi_i q^2_{i,i}\right) \\
&> C\left(\lambda \sum_{i=1}^{n} \pi_i q^1_{i,i} + (1-\lambda)\sum_{i=1}^{n} \pi_i q^2_{i,i}\right), \quad \text{by the strict convexity of } C \\
&= C\left(\sum_{i=1}^{n} \pi_i (\lambda q^1_{i,i} + (1-\lambda) q^2_{i,i})\right) \\
&= C\left(\sum_{i=1}^{n} \pi_i q^\lambda_{i,i}\right) \\
&= \acute{C}(\pi, Q_\lambda)
\end{aligned}
$$

Assumption E. By Blackwell's theorem (cf. Leshno and Spector, 1992), gross payoffs from using the information structure QR cannot exceed gross payoffs from using Q. By the definition of $\tilde{\mathscr{Q}}$, gross payoffs are performance multiplicatively scaled by the incentive level. Therefore, the performance associated with QR must be no greater than the performance associated with Q. Call these performance levels P^R and P^Q. $P^R \geq \frac{1}{n}$, or else QR would not be in $\tilde{\mathscr{Q}}$; it would have to have a diagonal entry that is not maximal in its column. Therefore, $C(P^R) \leq C(P^Q)$ which implies that $\acute{C}(\pi, QR) \leq \acute{C}(\pi, Q)$.

Since all relevant assumptions hold, Proposition A2 implies the result. □

A2.4 Proof of Proposition 4

Proof. By the convexity of C, the DM's maximand $rP - C(P)$ is concave. Therefore, local maxima are global maxima. The first order condition is $r = C'(P^*)$. If $r \in \left(C'\left(\frac{1}{n}\right), \lim_{x\uparrow 1} C'(x)\right)$, then we can write $P^* = (C')^{-1}(r)$, because the strict convexity of C on $\left(\frac{1}{n}, 1\right)$ implies that C' is strictly increasing and therefore invertible on that interval.

By the differentiability of C, $\frac{1}{n}$ could not be a global minimum unless $C'\left(\frac{1}{n}\right) = 0$, thereby implying that $P^*(r) \geq \frac{1}{n}$.

Finally, if $r \geq \lim_{x\uparrow 1} C'(x)$, then by concavity and differentiability, the maximand is increasing to the left of 1. Therefore, it is maximized for $P^* = 1$. □

A2.5 Proof of Proposition 5

Before proceeding with the proof, we note that that the $\sigma = 1$ case, mutual information, follows from Proposition 1 of Matějka and McKay (2015). However, for the sake of completness, we provide a complete, independent proof for all cases here.

Proof. We require a lemma:

Lemma A1. *Q can be chosen such that a decision matrix of the form $D = \begin{bmatrix} I_n \\ 0 \end{bmatrix}$ is optimal, where I_n denotes the $n \times n$ identity matrix.*

> *Proof.* Since entropy-based costs are finite for all Q, Assumption B holds. Therefore, following the proof of Proposition A2, for any Q, $\exists\, R$ right-stochastic such that QR is straightforward. Therefore, Q dominates QR in the Blackwell order (Blackwell, 1953). Posterior-separable cost functions complete the Blackwell order (cf. Subsection 9.3 of Caplin et al., 2019), and entropy-based costs are posterior-separable (cf. Subsection 8.3 of Caplin et al., 2019). Therefore, $C(\pi, Q) \geq C(\pi, QR)$, and a straightforward signal is optimal, implying the result. □

Because of this result, for notational convenience, we ignore the unused signals in M and assume $M = \Theta$ for the remainder of the proof, so we can consequently write $D = I_n$.

Because $D = I_n$, we can write the decision-maker's problem as:

$$\max_{Q \in \mathcal{Q}^n} \frac{r}{n}\text{tr}(Q) - \alpha(H(\pi) - \mathbb{E}[H(\pi|Q)]) \tag{A11}$$

$$\text{subject to} \quad \sum_{j=1}^{n} q_{i,j} = 1 \;\forall i \qquad (\lambda_i)$$

$$q_{i,j} \geq 0 \;\forall\, i, j \qquad (\mu_{i,j})$$

The first-order conditions for (A11) in the case that $\sigma = 1$ are:

$$\begin{aligned} i = j: &\quad \tfrac{r}{n} - \alpha\left[\ln \tfrac{q_{i,i}}{\sum_{k=1}^n q_{k,i}} + 1 - \tfrac{n-1}{\sum_{k=1}^n q_{k,i}}\right] - \lambda_i + \mu_{i,i} = 0 \\ i \neq j: &\quad -\alpha\left[\ln \tfrac{q_{i,j}}{\sum_{k=1}^n q_{k,j}} + 1 - \tfrac{n-1}{\sum_{k=1}^n q_{k,j}}\right] - \lambda_i + \mu_{i,j} = 0 \end{aligned} \tag{A12}$$

And in the case that $\sigma \neq 1$:

$$\begin{aligned} i = j: &\quad \tfrac{r}{n} + \tfrac{\alpha}{n(\sigma-1)}\left[\left(1 - \sum_{k=1}^n \left(\tfrac{q_{k,i}}{\sum_{l=1}^n q_{l,i}}\right)^\sigma\right) + \sum_{h=1}^n q_{h,i}\left(-\sigma\left(\tfrac{q_{i,i}}{\sum_{l=1}^n q_{l,i}}\right)^{\sigma-1}\left(\tfrac{\sum_{l=1}^n q_{l,i} - q_{i,i}}{\left(\sum_{l=1}^n q_{l,i}\right)^2}\right)\right.\right. \\ &\quad \left.\left. + \sigma \sum_{k\neq i}\left(\tfrac{q_{k,i}}{\sum_{l=1}^n q_{l,i}}\right)^{\sigma-1}\left(\tfrac{q_{k,i}}{\left(\sum_{l=1}^n q_{l,i}\right)^2}\right)\right)\right] - \lambda_i + \mu_{i,i} = 0 \\ i \neq j: &\quad \tfrac{\alpha}{n(\sigma-1)}\left[\left(1 - \sum_{k=1}^n \left(\tfrac{q_{k,j}}{\sum_{l=1}^n q_{l,j}}\right)^\sigma\right) + \sum_{h=1}^n q_{h,j}\left(-\sigma\left(\tfrac{q_{i,j}}{\sum_{l=1}^n q_{l,j}}\right)^{\sigma-1}\left(\tfrac{\sum_{l=1}^n q_{l,j} - q_{i,j}}{\left(\sum_{l=1}^n q_{l,j}\right)^2}\right)\right.\right. \\ &\quad \left.\left. + \sigma \sum_{k\neq i}\left(\tfrac{q_{k,j}}{\sum_{l=1}^n q_{l,j}}\right)^{\sigma-1}\left(\tfrac{q_{k,j}}{\left(\sum_{l=1}^n q_{l,j}\right)^2}\right)\right)\right] - \lambda_i + \mu_{i,j} = 0 \end{aligned} \tag{A13}$$

Before proceeding further, we require two additional lemmas:

Lemma A2. *$\exists\, q \in [0,1]$ such that $q_{i,i} = q\ \forall\, i$.*

Proof. Suppose there were an experiment Q with entries $a_{i,j}$ that solved (A11), with possibly unequal diagonal entries. Let $\tau_k(i) = i + k \mod n$ for $k = 0, \ldots n-1$. Let Q_k be the matrix with entries $q_{\tau_k(i),\tau_k(j)}$. That is, Q_k cycles the rows of Q and cycles the entries in each row so that the set of diagonal entries remains the same. Then, $\mathrm{tr}(Q_k) = \mathrm{tr}(Q)\ \forall\, k$, and since the prior is uniform, $C(\pi, Q_k) = C(\pi, Q)\ \forall\, k$. Consider the convex combination of experiments $Q' := \frac{1}{n}\sum_{k=0}^{n-1} Q_k$. It is obvious that $\mathrm{tr}(Q) = \mathrm{tr}(Q)$, and the diagonal entries $q'_{i,i}$ of Q' are all equal. By the convexity of posterior-separable cost functions, $C(\pi, Q) \geq C(\pi, Q')$. Therefore, the same probability of success can be attained at a (weakly) lower cost with Q' as compared to Q. □

Because the DM's performance is $\frac{1}{n}\mathrm{tr}(Q)$, this implies that her performance is simply given by q.

Lemma A3. *Either the μ constraints are slack, or $Q = I_n$.*

Proof. Suppose there were an experiment Q with entries $q_{i,j}$ that solved (A11),

with some off-diagonal entries possibly not zero. Let:

$$\tau_{k,\ell}(j) = \begin{cases} j, & j = k \\ j + \ell + \mathbf{1}_{\{h:\, h \le k \le h+\ell \text{ or } h \le k+n \le h+\ell\}}(j) \mod n, & j \neq k \end{cases} \tag{A14}$$

Let Q_ℓ be the matrix with entries $q_{i,\tau_{i,\ell}(j)}$. That is, Q_ℓ cycles the off-diagonal entries of each row. Then $\text{tr}(Q_\ell) = \text{tr}(Q)\ \forall\, \ell$, and since the prior is uniform, $C(\pi, Q_\ell) = C(\pi, Q)\ \forall\, \ell$. Consider the convex combination of experiments $Q' \equiv \frac{1}{n}\sum_{\ell=0}^{n-1} Q_\ell$. It is obvious that $\text{tr}(Q') = \text{tr}(Q)$, so that the probability of guessing the correct state is the same under both experiments. Moreover, in each row where Q has a non-zero off-diagonal entry, Q' has no zero off-diagonal entries. By the convexity of posterior-separable cost functions, $C(\pi, Q) \geq C(\pi, Q')$. Therefore, the same probability of success can be attained at a (weakly) lower cost with Q' as compared to Q. Because we showed in Lemma A2 that all diagonal entries could be assumed equal, this shows that either Q' is the identity, or the non-negativity constraint is slack on all off-diagonal entries, and all such entries are equal. □

Consider the case where $\sigma = 1$. Applying Lemma A3, for now we assume that the μ constraints are slack, so that $\mu_{i,j} = 0\ \forall\, i, j$. Making this assumption allows us to rearrange (A12) by subtraction as:

$$\frac{r}{n} = \alpha \left[\left(\ln \frac{q_{i,i}}{\sum_{k=1}^n q_{k,i}} - \frac{n-1}{\sum_{k=1}^n q_{k,i}} \right) - \left(\ln \frac{q_{i,j}}{\sum_{k=1}^n q_{k,j}} - \frac{n-1}{\sum_{k=1}^n q_{k,j}} \right) \right] \quad \forall\, i, j \tag{A15}$$

By Lemma A2, $\exists\, q \in [0,1]$ such that $q_{i,i} = q\ \forall\, i$. Since (A15) applies $\forall\, i, j$, this in turn implies that $\exists\, \tilde{q}$ such that $q_{i,j} = \tilde{q}\ \forall j \neq i$. Because the entries in each row of Q must sum to 1, this implies that $\tilde{q} = \frac{1-q}{n-1}$. Therefore, (A15) can be rewritten as:

$$\frac{r}{\alpha n} = \ln q - \ln \frac{1-q}{n-1} \tag{A16}$$

Rearranging (A16) gives:

$$q = \frac{\exp\left(\frac{r}{\alpha n}\right)}{n - 1 + \exp\left(\frac{r}{\alpha n}\right)} \tag{A17}$$

which is a logistic function of r. Therefore, the FOCs have a solution, and by the concavity of (A11), we need not consider corner solutions. Finally, it is easily observed that $q < 1$ for all r and that $\lim_{r\to\infty} q = 1$.

Now consider the case where $\sigma \neq 1$. Applying Lemmas A2 and A3 and assuming an interior solution, we can rewrite (A13) as:

$$\begin{aligned} i = j: \quad & \tfrac{r}{n} + \tfrac{\alpha}{n(\sigma-1)}\left[1 - q^\sigma - (n-1)\left(\tfrac{1-q}{n-1}\right)^\sigma + \sigma\left((n-1)\left(\tfrac{1-q}{n-1}\right)^\sigma + q^\sigma - q^{\sigma-1}\right)\right] = \lambda \\ i \neq j: \quad & \tfrac{\alpha}{n(\sigma-1)}\left[1 - q^\sigma - (n-1)\left(\tfrac{1-q}{n-1}\right)^\sigma + \sigma\left((n-2)\left(\tfrac{1-q}{n-1}\right)^\sigma + q^\sigma - \left(\tfrac{1-q}{n-1}\right)^{\sigma-1}\left(\tfrac{n-2+q}{n-1}\right)\right)\right] = \lambda \end{aligned} \tag{A18}$$

By subtraction, (A18) can be rearranged as:

$$\frac{r}{n} + \frac{\alpha\sigma}{n(\sigma-1)}\left[\left(\frac{1-q}{n-1}\right)^\sigma + \left(\frac{1-q}{n-1}\right)^{\sigma-1}\left(\frac{n-2+q}{n-1}\right) - q^{\sigma-1}\right] = 0 \tag{A19}$$

This can be further rearranged as:

$$r + \frac{\alpha\sigma}{\sigma-1}\left[\left(\frac{1-q}{n-1}\right)^{\sigma-1} - q^{\sigma-1}\right] = 0 \tag{A20}$$

In general, (A20) does not have a closed-form solution for q. However, we can check for which r the q that solves (A20) is less than 1. For these r, the FOCs are sufficient, by the concavity of (A11). For other r, we must check corner solutions.

Applying the implicit function theorem to (A20), we have:

$$\frac{dq}{dr} = \left[\alpha\sigma\left(\frac{1}{n-1}\left(\frac{1-q}{n-1}\right)^{\sigma-2} + q^{\sigma-2}\right)\right]^{-1} \tag{A21}$$

This is strictly positive for $q \in (0, 1)$. Note also that if $r = 0$, then the solution to (A20) is $q = \frac{1}{n}$, and if $\sigma > 1$ and $r = \frac{\alpha\sigma}{\sigma-1}$, then the solution to (A20) is $q = 1$. Therefore, for $\sigma > 1$, the FOCs are

sufficient for all $r \in \left(0, \frac{\alpha\sigma}{\sigma-1}\right)$. For $\sigma > 1$ and $r \geq \frac{\alpha\sigma}{\sigma-1}$, applying Lemmas A2 and A3, it must be the case that $q = 1$.

When $\sigma \in (0,1)$, $\lim_{q\to\infty} \frac{\alpha\sigma}{1-\sigma}\left[\left(\frac{1-q}{n-1}\right)^{\sigma-1} - q^{\sigma-1}\right] = \infty$, so by the intermediate value theorem, for any $r > 0$, $\exists\, q \in \left[\frac{1}{n}, 1\right]$ that solves (A20). Therefore, for $\sigma \in (0,1)$ the FOCs are always sufficient. To show that there there is a horizontal asymptote at 1, consider an arbitrary $\varepsilon > 0$. We must show that there exists $\bar{r} > 0$ such that for any $r > \bar{r}$, the q that solves (A20) is such that $1 - q < \varepsilon$. Let $\bar{r} = \frac{\alpha\sigma}{\sigma-1}\max\left\{\left(\frac{\varepsilon}{n-1}\right)^{\sigma-1}, 1\right\}$. If $\varepsilon \geq 1$, then clearly $1 - q < \varepsilon$ for every r. If $\varepsilon < 1$, let $q'(r)$ be such that $r = \frac{\alpha\sigma}{\sigma-1}\left(\frac{1-q'(r)}{n-1}\right)^{\sigma-1}$ for $r > \bar{r}$, so that $1 - q'(r) < \varepsilon$. As shown above, the solution to (A20) is strictly increasing in r. Therefore, because $\frac{\alpha\sigma}{\sigma-1}\left(\frac{1-q}{n-1}\right)^{\sigma-1} > \frac{\alpha\sigma}{\sigma-1}\left[\left(\frac{1-q}{n-1}\right)^{\sigma-1} - q^{\sigma-1}\right]$, the q that solves (A20) is larger than $q'(r)$, which implies that $1 - q < \varepsilon$, thereby proving the claim of a horizontal asymptote at 1.

We now turn towards proving the claims made about the concavity/convexity of the performance function in the proposition. For that, we require an expression for the second derivative of the q that solves (A20). Differentiating (A21) with respect to r gives:

$$\begin{aligned}\frac{d^2q}{dr^2} &= -\frac{\sigma-2}{\alpha\sigma}\frac{dq}{dr}\left[\left(\frac{1}{n-1}\right)^{\sigma-1}(1-q)^{\sigma-2} + q^{\sigma-2}\right]^{-2}\left[q^{\sigma-3} - \frac{(1-q)^{\sigma-3}}{(n-1)^{\sigma-1}}\right] \\ &= -\alpha\sigma(\sigma-2)\left(\frac{dq}{dr}\right)^3\left[q^{\sigma-3} - \frac{(1-q)^{\sigma-3}}{(n-1)^{\sigma-1}}\right] \qquad \text{(A22)}\end{aligned}$$

Consider the case where $\sigma \in (0,1)\cup(1,2)$. Since $\frac{dq}{dr}$ is positive, $q^{\sigma-3} > (<)\frac{(1-q)^{\sigma-3}}{(n-1)^{\sigma-1}}$ implies that the performance function is convex (concave). Rearranging, this happens when

$$q < (>)\left[1 + (n-1)^{\frac{\sigma-1}{\sigma-3}}\right]^{-1} \qquad \text{(A23)}$$

Therefore, the performance function is convex for r such that $q < \left[1 + (n-1)^{\frac{\sigma-1}{\sigma-3}}\right]^{-1}$ and concave for r such that $q > \left[1 + (n-1)^{\frac{\sigma-1}{\sigma-3}}\right]^{-1}$. Because the performance function is increasing, this implies a sigmoidal shape, since $\left[1 + (n-1)^{\frac{\sigma-1}{\sigma-3}}\right]^{-1} \in \left(\frac{1}{n}, 1\right)$.

Now consider the case where $\sigma \in (2,3)$. This flips the sign of (A22) from the $\sigma \in (0,1)\cup(1,2)$ case. Therefore, the performance function is concave for r such that $q < \left[1 + (n-1)^{\frac{\sigma-1}{\sigma-3}}\right]^{-1}$ and convex for r such that $q \in \left(\left[1 + (n-1)^{\frac{\sigma-1}{\sigma-3}}\right]^{-1}, 1\right)$. Because the performance function is

increasing, this implies an inverse-S shape, since $\left[1+(n-1)^{\frac{\sigma-1}{\sigma-3}}\right]^{-1} \in \left(\frac{1}{n},1\right)$.

Now consider the case where $\sigma > 3$. Since $\sigma - 3 > 0$, this flips the condition (A23) from the $\sigma \in (2,3)$ case, so that $q > \left[1+(n-1)^{\frac{\sigma-1}{\sigma-3}}\right]^{-1}$ implies concavity. But $\sigma > 3$ also implies that $\left[1+(n-1)^{\frac{\sigma-1}{\sigma-3}}\right]^{-1} > \frac{1}{n}$, so that the performance function is everywhere concave.

We now turn our focus to the special cases not previously covered. When $\sigma = 2$, $\frac{\alpha\sigma}{\sigma-1} = 2\alpha$, and (A20) can be written as:

$$r + 2\alpha\left[\frac{1-q}{n-1} - q\right] = 0 \tag{A24}$$

Rearranging, this gives:

$$q = \frac{n-1}{2\alpha n}r + \frac{1}{n} \tag{A25}$$

This is clearly an affine function of r, and it matches the claim about the performance function when $\sigma = 2$ in the proposition.

When $\sigma = 3$, $\frac{\alpha\sigma}{\sigma-1} = \frac{3\alpha}{2}$, and (A20) can be written as:

$$r + \frac{3\alpha}{2}\left[\frac{(1-q)^2}{(n-1)^2} - q^2\right] = 0 \tag{A26}$$

Rearranging, this gives:

$$3\alpha n(n-2)q^2 + 6\alpha q - (3\alpha + 2r(n-1)^2) = 0 \tag{A27}$$

Applying the quadratic formula, taking the root associated with the plus sign to ensure increasing performance, and performing tedious algebraic manipulations, it can be concluded that:

$$q = \frac{1}{n(n-2)}\left[\frac{(n-1)\sqrt{9\alpha^2 + 6\alpha n(n-2)r}}{3\alpha} - 1\right] \tag{A28}$$

This is clearly a square-root function of r, and it matches the claim about the performance function when $\sigma = 3$ in the proposition.

□

A2.6 Proof of Proposition 6

Proof. Suppose that the DM has been given a uniform guess task and has received a signal $\hat{m}$. Her belief that the state of the world is θ is, by Bayes' rule:

$$\Pr(\theta|\hat{m}) = \frac{\frac{1}{n}\frac{1}{s}\phi\left(\frac{\hat{m}-\theta}{s}\right)}{\sum_{i=1}^{n}\frac{1}{n}\frac{1}{s}\phi\left(\frac{\hat{m}-\theta_i}{s}\right)} = \frac{\frac{1}{s}\phi\left(\frac{\hat{m}-\theta}{s}\right)}{\sum_{i=1}^{n}\frac{1}{s}\phi\left(\frac{\hat{m}-\theta_i}{\sigma}\right)} \tag{A29}$$

where $\phi(\cdot)$ is the standard normal density. Notice that the denominator in (A29) depends only on $\hat{m}$; it is the same for all θ. Therefore, if the DM is trying to determine the most likely state given her signal, she only needs to compare the numerators of (A29) for each possible θ; in other words, she only needs to find the state that maximizes the conditional probability density of her signal.

Since the normal probability density function is symmetric around its mean, which is also its mode, the conditional probability density of her signal is maximized at θ_1 if $\hat{m} \leq \frac{1}{2}(\theta_1 + \theta_2)$, at θ_i if $\hat{m} \in \left[\frac{1}{2}(\theta_{i-1} + \theta_i), \frac{1}{2}(\theta_i + \theta_{i+1})\right]$ for $i \in \{2, 3, \ldots, n-1\}$, and at θ_n if $\hat{m} \geq \frac{1}{2}(\theta_{n-1} + \theta_n)$.

Because consecutive states are equidistant, if the DM guesses optimally given her signal, her probabilities of guessing state i given true state j are:

$$\Pr(a = \theta_i|\theta = \theta_j) = \begin{cases} \Phi\left(\zeta\eta(3-2j)\right), & i = 1 \\ \Phi\left(\zeta\eta(2(i-j)+1)\right) - \Phi\left(\zeta\eta(2(i-j)-1)\right), & i \in \{2, 3, \ldots, n-1\} \\ 1 - \Phi\left(\zeta\eta(2(n-j)-1)\right), & i = n \end{cases} \tag{A30}$$

where Φ is the cumulative distribution function of the standard normal distribution. This implies that the DM's problem is, as in (12):

$$\max_{\zeta \in [0,\infty)} \frac{r}{n}\left[2\Phi\left(\zeta\eta\right) + (n-2)\left(2\Phi\left(\zeta\eta\right) - 1\right)\right] - K(\zeta)$$

We can rewrite this as:

$$\max_{\zeta \in [0,\infty)} \frac{r}{n}\left[(2n-2)\Phi\left(\zeta\eta\right) - (n-2)\right] - K(\zeta) \tag{A31}$$

The first-order condition is:

$$F(r,\zeta) \equiv \frac{(2n-2)r\eta}{n}\phi(\zeta\eta) - K'(\zeta) = 0 \tag{A32}$$

In order to ensure that the first-order condition is sufficient, we verify the second-order condition:

$$-\frac{(2n-2)r\eta^3}{n}\zeta\phi(\zeta\eta) - K''(\zeta) < 0, \quad \text{since } \zeta \text{ is positive}$$

The DM's performance function is:

$$P^*(r) = \frac{1}{n}[(2n-2)\Phi(\zeta(r)\eta) - (n-2)] \tag{A33}$$

In order to show that $P^*(r)$ is strictly concave, we compute:

$$\begin{aligned}\frac{d^2P^*}{dr^2} =& \frac{dP^*}{dr}\left[\frac{(2n-2)\eta}{n}\phi(\zeta\eta)\frac{d\zeta}{dr}\right] \\ =& -\frac{(2n-2)\eta^3}{n}\zeta\phi(\zeta\eta)\frac{d\zeta}{dr} + \frac{(2n-2)\eta}{n}\phi(\zeta\eta)\frac{d^2\zeta}{dr^2}\end{aligned} \tag{A34}$$

In order to determine the sign of (A34), we must compute $\frac{d\zeta}{dr}$ and $\frac{d^2\zeta}{dr^2}$. By the implicit function theorem:

$$\begin{aligned}\frac{d\zeta}{dr} =& \frac{-\frac{\partial F}{\partial r}}{\frac{\partial F}{\partial \zeta}} \\ =& \frac{\frac{(2n-2)\eta}{n}\phi(\zeta\eta)}{\frac{(2n-2)r\eta^3}{n}\zeta\phi(\zeta\eta) + K''(\zeta)} \\ >& 0\end{aligned} \tag{A35}$$

Differentiating (A35) with respect to r gives:

$$\begin{aligned}\frac{d^2\zeta}{dr^2} = \left[\frac{(2n-2)r\eta^3}{n}\zeta\phi(\zeta\eta) + K''(\zeta)\right]^{-2}\Bigg\{&-\frac{(2n-2)\eta^3}{n}\zeta\phi(\zeta\eta)\frac{d\zeta}{dr}\left(\frac{(2n-2)r\eta^3}{n}\zeta\phi(\zeta\eta) + K''(\zeta)\right) \\ &- \Bigg[\left(\frac{(2n-2)\eta^3}{n}\zeta\phi(\zeta\eta) + \frac{(2n-2)r\eta^3}{n}\frac{d\zeta}{dr}\phi(\zeta\eta) - \frac{(2n-2)r\eta^5}{n}\zeta^2\frac{d\zeta}{dr}\phi(\zeta\eta) + K'''(\zeta)\right) \\ &\times\left(\frac{(2n-2)\eta}{n}\phi(\zeta\eta)\right)\Bigg]\Bigg\}\end{aligned}$$

$$= \left[\frac{(2n-2)r\eta^3}{n}\zeta\phi(\zeta\eta) + K''(\zeta)\right]^{-2}\left\{-\frac{(2n-2)\eta^3}{n}\zeta\phi(\zeta\eta)\frac{d\zeta}{dr}K''(\zeta)\right.$$
$$- \left[\left(\frac{(2n-2)\eta^3}{n}\zeta\phi(\zeta\eta) + \frac{(2n-2)r\eta^3}{n}\frac{d\zeta}{dr}\phi(\zeta\eta) + K'''(\zeta)\right)\right.$$
$$\left.\left.\times\left(\frac{(2n-2)\eta}{n}\phi(\zeta\eta)\right)\right]\right\} \tag{A36}$$
$$< 0$$

Substituting (A35) and (A36) back into (A34) gives us that $\frac{d^2P^*}{dr^2} < 0$, since $\frac{d\zeta}{dr} > 0$ and $\frac{d^2\zeta}{dr^2} < 0$. This concludes the proof. □

A2.7 Proof of Proposition 7

Proof. We solve the DM's non-concave maximization problem by reducing it to a finite number of concave maximization problems. In order to do so, we require a lemma.

Lemma A4. *For any $r > 0$, $P^*(r) \not\in [0, d_1] \cup [d_2, d_3]$.*

> *Proof.* Because $\lim_{x \downarrow d_1} C'(P) = 0$, and $C(P)$ is strictly increasing on d_1, d_2, there exists $\varepsilon > 0$ such that if $P \in (d_1, d_1 + \varepsilon)$, then $rP - C(P) > rd_1 - C(d_1)$. Moreover, since $C(d_1) = C(\grave{P})$ for all $\grave{P} \in [0, d_1)$, all $\grave{P} \in [0, d_1]$ are suboptimal. The same argument applies to $[d_2, d_3]$, *mutatis mutandis.* □

Therefore, the optimal P^* for each r lies in $(d_1, d_2) \cup (d_3, 1]$, and we can search for the optimal P^* separately in (d_1, d_2) and $(d_3, 1]$ and then take the maximum of the two. In each of these intervals, the DM's maximization problem is concave, and so the first-order conditions are sufficient if they can be satisfied on those intervals.

The first-order conditions in (d_1, d_2) and $(d_3, 1]$ yield $P_L^* := \frac{r}{2c_1} + d_1$ and $P_H^* := \frac{r}{2c_2} + d_3$, respectively. Assuming these conditions can be satisfied on their respective intervals, the net utilities associated with those performance levels are $\frac{r^2}{4c_1} + d_1 r$ and $\frac{r^2}{4c_2} + d_3 r - c_1(d_2 - d_1)^2$, respectively. Therefore, $P_H^* \succeq P_L^*$ iff:

$$\frac{r^2}{4c_2} + d_3 r - c_1(d_2 - d_1)^2 \geq \frac{r^2}{4c_1} + d_1 r$$
$$\iff \frac{1}{4}\left(\frac{c_1 - c_2}{c_1 c_2}\right) r^2 + (d_3 - d_1)r - c_1(d_2 - d_1)^2 \geq 0 \tag{A37}$$

When $c_1 = c_2$, then (A37) can be rearranged as:

$$r \geq \frac{c_1(d_2 - d_1)^2}{d_3 - d_1} = \delta \tag{A38}$$

When $c_1 \neq c_2$, we find the roots of the quadratic expression in (A37) by applying the quadratic formula:

$$\frac{2c_1c_2}{c_1 - c_2}\left[\pm\sqrt{(d_3 - d_1)^2 + \frac{(c_2 - c_1)(d_2 - d_1)^2}{c_2}} - (d_3 - d_1)\right] \tag{A39}$$

Denote by δ^+ the root with the positive square root and δ^- the root with the negative square root. When $c_1 > c_2$, δ^+ is positive, and δ^- is negative. Therefore, since $r > 0$, we can conclude that $\delta = \delta^+$.

When $c_1 < c_2$, both δ^+ and δ^- are positive, and (A37) is satisfied when $r \in [\delta^+, \delta^-]$.[A23] However, by Proposition 1, NIAC would be violated if $P^*(r) \in (d_1, d_2)$ were optimal for $r \geq \delta^-$, so again we can conclude that $\delta = \delta^+$.

Now, note that $P_L^* = 1$ iff $r = 2c_1(d_2 - d_1) := \bar{r}_L$ and $P_H^* = 1$ iff $r = 2c_2(1 - d_3) = \bar{r}_H$. Tedious algebraic manipulations show that $\delta < \bar{r}_L$ for any parameters satisfying the restrictions in the definition of C in (14), and $\delta < \bar{r}_H$ with the additional restriction on d_3 provided in the proposition. This ensures that the DM selects $P^*(r) \in (d_1, d_2)$ for $r < \delta$, $P^*(r) \in (\delta_3, 1)$ for $r \in [\delta, \bar{r}_H)$ and $P^*(r) = 1$ for $r \geq \bar{r}_H$. □

A3 Laboratory Experiment Instructions

This appendix contains the instructions that were read out loud to subjects in our laboratory experiment for the $10 prize treatment, as well as the slides that were displayed to them as the instructions were read out. Instructions and slides were similarly delivered for the $20 treatment, *mutatis mutandis*.

[A23] Note that $\delta^- > \delta^+$ when $c_1 < c_2$.

A3.1 Oral Instructions

Text in square brackets was not read aloud and was used to remind the person reading the instructions of what needed to be done. Text in angle brackets was not read aloud and was used to indicate which slides should be displayed while the instructions were being read.

⟨Slide 2⟩ Welcome to the Columbia Experimental Laboratory for the Social Sciences (CELSS)! Your participation in this experiment is much appreciated. During this session, we require your complete, undivided attention. As such, we ask that you remain quiet for the duration of the session, refrain from opening other applications on your computer, refrain from talking or passing notes to other participants, and put away all of your possessions, including your cell phones, which must be turned off. Do not touch the computer terminals until the session begins.

Before we begin, please read and sign both copies of the consent form located at your terminal. Please hand one signed copy to us, and place the second under your chair; you may take that copy with you when you have completed the experiment. [COLLECT CONSENT FORMS AND ENSURE THAT THEY ARE ALL SIGNED AND DATED]

⟨Slide 3⟩ You will be paid in cash for your participation in this experiment. Payment will occur in private once you have completed the experiment. This payment will depend on your own decisions and on chance; different participants may earn different amounts. During the session, please do not communicate with other subjects, and please do not write anything down unless we tell you to.

⟨Slide 4⟩ The currency in this experiment is called "points." In this experiment, you will be asked to complete a series of tasks. Each task has a potential reward, in points, for a correct answer. You will be asked to complete two types of task in this experiment, which we will refer to as the "dots task" and the "angle task." You will either complete all the dots tasks or all the angle tasks first. You will be asked to complete both types of task 100 times each, once for each of 100 different reward levels for a correct answer. The reward level will take values between 1 point and 100 points with reward increments of 1 point. The order in which you will see the tasks corresponding to each reward level will be random.

⟨Slide 5⟩ We will now describe the two types of tasks. At the start of each task, the reward level will be displayed in large characters for three seconds [SHOW SCREENSHOT OF REWARD LEVEL], after which it will be replaced with an image. The reward will continue to be displayed in small characters next to the image. ⟨Slide 6⟩ In the "dots task," the image that you will be shown is a pattern of dots. [SHOW SCREENSHOT OF DOTS] You will be asked to determine the number of dots on the screen. The number of dots will be between 38 and 42, inclusive, and each of those five numbers will be equally likely. When you are ready to answer, select the option corresponding to your guess, and then click the submit button. The number of points you could earn from correctly determining the number of dots is indicated near the top-right of the screen. [POINT TO REWARD ON SCREENSHOT] There is no time limit for your response.

⟨Slide 7⟩ In the "angle" task, the image that you will be shown consists of two intersecting blue lines of random length. [SHOW SCREENSHOT OF ANGLE] You will be asked to determine the angle between these two lines. [SHOW ANGLES] The angle will be ⟨Slide 8⟩ 35 degrees, ⟨Slide 9⟩ 40 degrees, ⟨Slide 10⟩ 45 degrees, ⟨Slide 11⟩ 50 degrees, or ⟨Slide 12⟩ 55 degrees, with each of the five angles equally likely. ⟨Slide 13⟩ Keep in mind that 0 degrees is the angle between two lines in the exact same position, and 90 degrees is the angle between two adjacent lines of a rectangle. The reward you could earn from correctly determining the angle is indicated near the top-right of the screen as before. [POINT TO REWARD ON SCREENSHOT] There is no time limit for your response.

⟨Slide 14⟩ After you have completed all the tasks, the computer will randomly select one "dots" task and one "angle" task. For each of these two tasks that you answered correctly, you will receive the corresponding point value.

Your payment for the experiment will be determined as follows. You will be given a $10 participation fee for completing the experiment. In addition to this fee, you will have the opportunity to earn up to two additional $10 prizes. The number of points you earned from each of the selected tasks determines the probability that the computer will award you these prizes. ⟨Slide 15⟩ For example, say the selected "dots" task had

a reward level of 84, and the selected "angle" task had a reward level of 33. If you answered the selected dots task correctly, this would give you 84 points and therefore an 84% probability of being awarded the first $10 prize. If you answered the selected "angle" task incorrectly, this would give you zero points and therefore a 0% probability of being awarded the second $10 prize.

When you have completed all the tasks, you will be given a brief questionnaire. This questionnaire will not affect your payment. ⟨Slide 16⟩ After that, you will be shown a results screen that looks like this. [RESULTS SCREEN] This screen will show you what tasks were selected for payment, whether you answered them correctly, and whether you were awarded the corresponding prizes. At that point, please raise your hand, and we will give you a receipt form [SHOW FORM] for you to fill out. ⟨Slide 17⟩ Please write your terminal number, located at the top-right of your carrel [POINT TO NUMBER ON CARREL], on the line marked "Computer ID." If you were awarded both $10 prizes, please write $20 for "Experimental Earnings" and $30 for "Total." If you were awarded one of the two prizes, please write $10 for "Experimental Earnings" and $20 for "Total." If you were awarded neither of the prizes, please write $0 for "Experimental Earnings" and $10 for "Total." Once you have finished filling out the receipt form, please hand it to one of the experimenters for verification. We will then give you your earnings, and you may leave the lab.

Before we proceed, are there any questions? [WAIT FOR QUESTIONS]

We will now begin the experiment. ⟨Slide 18⟩

A3.2 Slides

Slides were displayed according to the transitions indicated in the instructions given in the preceding subsection of the appendix.

Lab Experiment

Experimenters: Ambuj Dewan and Nathaniel Neligh

May 31, 2016

Introduction

- Welcome to CELSS!
- Please remain seated and turn off your cell phones.
- Please read and sign both copies of the consent form located at your terminal.

Description and Instructions

- Payment will occur at the end of the session.
  - Your payment will depend only on your own decisions and chance, not the decisions of others.
- Experiment takes place entirely on computer screens.

Description and Instructions

- Will complete a series of tasks for points.
- Two types of task: **dots** and **angle**.
- 100 tasks of each type.
- Each task has a reward level.

Description and Instructions

Reward Level

81
Points

This is dots task number 2 out of 100.

A correct answer to this question is worth **81** points.

How many dots are in the picture?

38

39

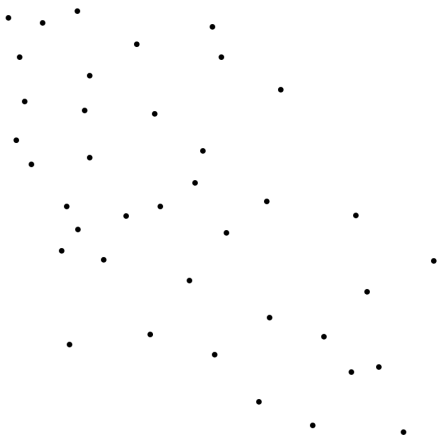
40

41

42

Description and Instructions

Dots Task



This is dots task number 2 out of 100.

A correct answer to this question is worth **81** points.

How many dots are in the picture?

38

39

40

41

42

Submit

Description and Instructions

Angle Task

This is angle task number 2 out of 100.

A correct answer to this question is worth **69** points.

What is the angle between the two blue lines?

35°

40°

45°

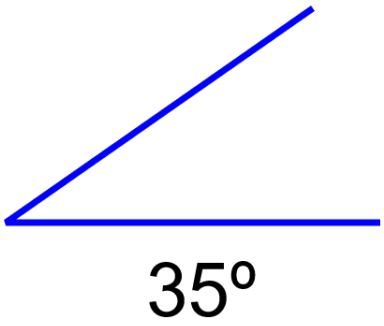
50°

55°

Submit

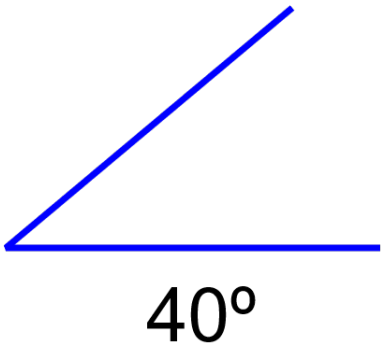
Description and Instructions

Angle Task – 35°



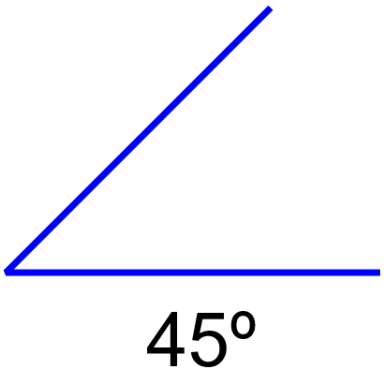
Description and Instructions

Angle Task – 40°



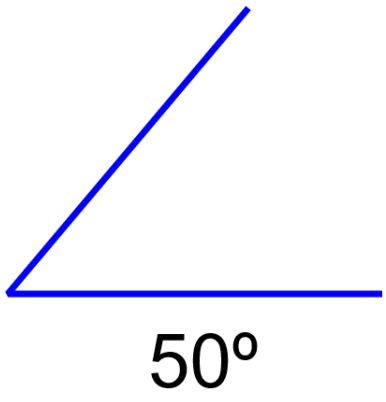
Description and Instructions

Angle Task – 45°



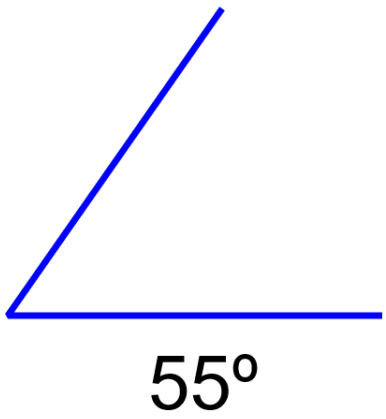
Description and Instructions

Angle Task – 50°



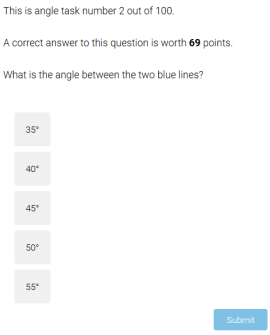
Description and Instructions

Angle Task – 55°



Description and Instructions

Angle Task



Description and Instructions

Payment

- Can earn \$10, \$20, or \$30.
- The computer randomly selects one “dots” and one “angle” task.
- You receive the corresponding number of points for a correct answer.
- The number of points for each task is the probability that the computer will award you a \$10 prize.

Description and Instructions

Payment Example

Task	Dots	Angle
Selected task value	84 points	33 points
Answered correctly?	Yes	No
Prize probability	84%	0%

Description and Instructions

Results Screen

The computer selected dots task number 9. It had a reward level of 79. The prize for this task was $10.

You answered this task correctly, so your probability of being awarded the prize was 79%.

You won the prize.

The computer selected angle task number 6. It had a reward level of 15. The prize for this task was $10.

You answered this task correctly, so your probability of being awarded the prize was 15%.

You did not win the prize.

The participation fee was $10.

Your final payoff was $20.

The experiment is now over. Please raise your hand and wait for an experimenter to come assist you.

Description and Instructions

Receipt Form

Payment Receipt

Date: _____

Computer ID: _____

Experimental Earnings: _____

Show-up fee: $10 _____

Total: _____

Signature: _____

EXPERIMENT IN PROGRESS

Table A1: Categorization of subjects by odd incentives

Category	**Of All Subjects**	**Of R.I. Subjects**	**Of Resp. Subjects**
All subjects	81 (100%)	—	—
R.I. subjects	60 (74.1%)	60 (100%)	—
Resp. subjects	///	32 (45.1%)	32 (100%)
W.B. subjects	///	///	6 (18.8%)

Note: "R.I." = rationally inattentive; "Resp." = responsive; "W.B." = well-behaved, i.e. subjects whose behavior is consistent with continuous, convex cost functions. — denotes that the column category is a subset of the row category, and /// denotes that the row category is defined only on a subset of the column category.

Table A2: Model Selection for Responsive Subjects, Odd Incentives

Model	Constant (1)	Binary (2)	Logistic (7)	SIC(8)	Concave (9)
Number of Subjects	1 (3.1%)	5 (15.6%)	17 (53.1%)	1 (3.1%)	8 (25.0%)

A4 Robustness Checks and Statistical Power Tests

A4.1 Half-Sample Analysis

As explained in Section 4 of the paper, presenting all the odd incentives followed by all the even incentives to each subject ensures roughly the same variation in incentives in both halves of the experiment. This allows us to perform the analyses of Sections 5 and 6 separately on both the odd incentives and even incentives as a robustness check to account for changes in subjects' behavior that may arise from fatigue or learning. Results are summarized in Tables A1 to A4.

We also examine the consistency of categorization between the full sample and the half-samples. Results for rationality, responsiveness, and well-behavedness are reported in the Venn diagrams of Figures A1, A2, and A3, respectively. Note that for rationality and responsiveness, a plurality of

Table A3: Categorization of subjects by even incentives

Category	**Of All Subjects**	**Of R.I. Subjects**	**Of Resp. Subjects**
All subjects	81 (100%)	—	—
R.I. subjects	71 (87.7%)	60 (100%)	—
Resp. subjects	///	33 (55.0%)	33 (100%)
W.B. subjects	///	///	9 (27.2%)

Note: "R.I." = rationally inattentive; "Resp." = responsive; "W.B." = well-behaved, i.e. subjects whose behavior is consistent with continuous, convex cost functions. — denotes that the column category is a subset of the row category, and /// denotes that the row category is defined only on a subset of the column category.

Table A4: Model Selection for Responsive Subjects, Even Incentives

Model	Binary (2)	Affine (3)	Logistic (7)	SIC (8)	Concave (9)
Number of Subjects	8 (24.2%)	1 (3.0%)	19 (57.6%)	1 (3.0%)	4 (12.1%)

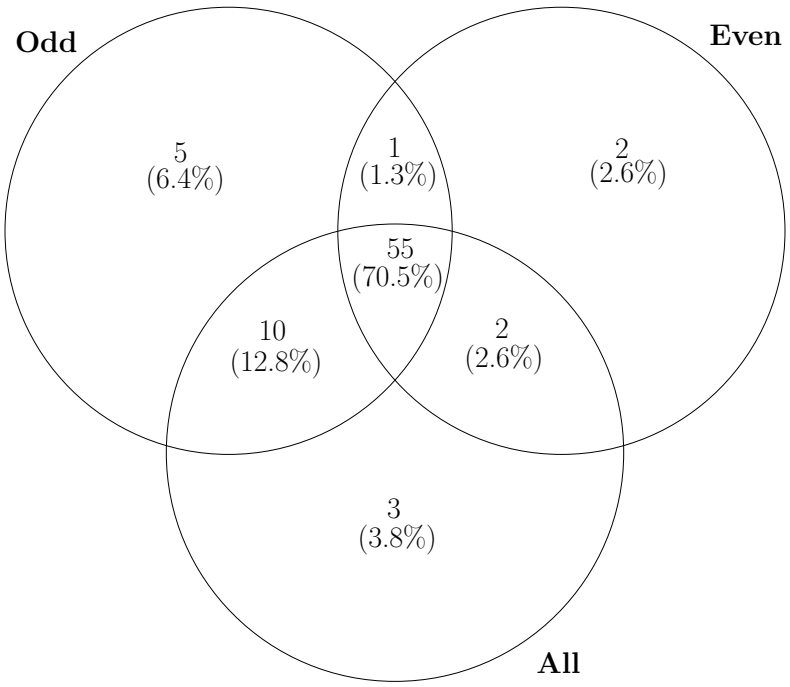


Figure A1: Venn diagram of number of subjects classified as rational in each of the samples. Percentages reported as proportion out of subjects classified as rational in at least one of the samples (78).

subjects are in the three-way intersection of the Venn diagram (the majority of subjects in the case of the former). The results for well-behavedness should be interpreted with caution, since the power of the test is fairly low to begin with (see the next subsection), and removing half the data would only make the power worse.

In Table A5, we report the correlations between AIC estimates between samples, looking at each of the subjects who were classified as responsive in at least one sample, the subjects who were classified as responsive in the full sample of incentives, and the subjects who were classified as responsive in all three samples (viz. all incentives, odd incentives, and even incentives). Correlations are fairly high when looking at the subjects who were classified as responsive in at least one sample or the subjects that are classified as responsive in the full sample of incentives. Correlations are higher between the full sample and the even incentives (presented to the subjects second) than between the full sample and the odd incentives (presented to the subjects first). This implies that

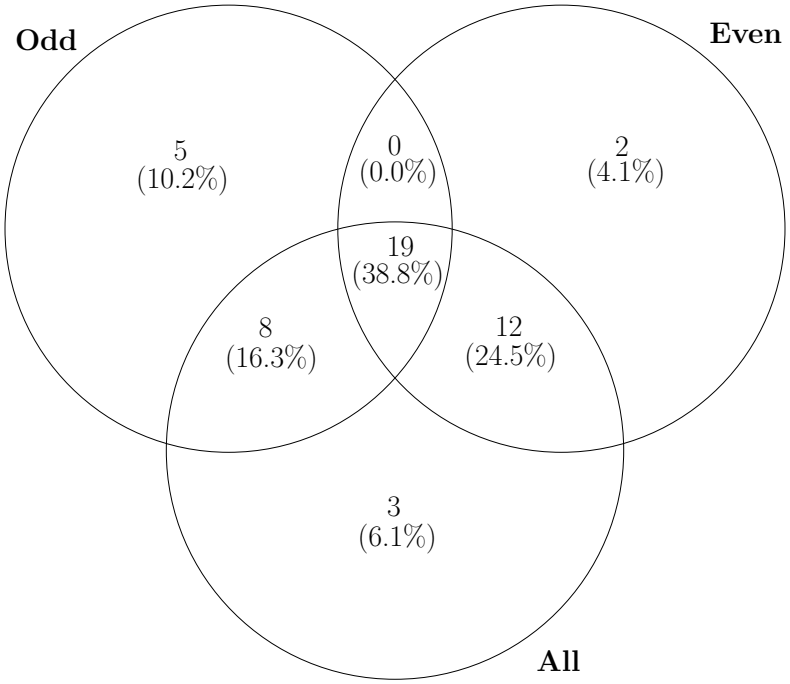


Figure A2: Venn diagram of number of subjects classified as responsive in each of the samples. Percentages reported as proportion out of subjects classified as responsive in at least one of the samples (49).

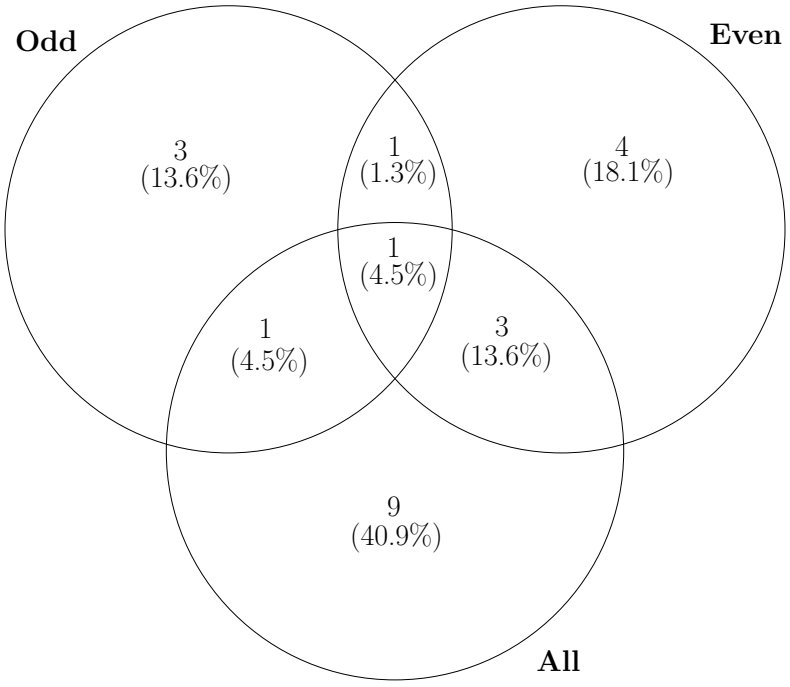


Figure A3: Venn diagram of number of subjects classified as well-behaved in each of the samples. Percentages reported as proportion out of subjects classified as well-behaved in at least one of the samples (22).

our model selection exercise is more reliable using data from the second half of the experiment, indicating that subject behavior "stabilized" as the experiment proceeded.

Finally, we report on the consistency of model selection between samples in Tables A6, A7, and A8. Note that for the most part, a responsive subject categorized according to one kind of performance function in one sample is more likely to maintain said categorization in another sample to than to switch to some other given categorization.

Table A5: Correlations of AIC estimates for each model between samples

All incentives and odd incentives			
	Responsive in:		
Model	At least one sample	Full sample	All three samples
1 (Constant)	0.656	0.675	0.587
2 (Binary)	0.641	0.677	0.375
3 (Affine w/ break)	0.700	0.716	0.732
4 (Affine)	0.701	0.722	0.727
5 (Quadratic)	0.304	0.334	0.604
6 (Cubic)	0.106	0.083	0.202
7 (Logistic)	0.793	0.824	0.809
8 (SIC)	0.808	0.840	0.753
9 (Concave)	0.820	0.846	0.800
All incentives and even incentives			
	Responsive in:		
Model	At least one sample	Full sample	All three samples
1 (Constant)	0.870	0.907	0.739
2 (Binary)	0.851	0.879	0.522
3 (Affine w/ break)	0.869	0.903	0.855
4 (Affine)	0.856	0.898	0.842
5 (Quadratic)	0.745	0.784	0.465
6 (Cubic)	0.644	0.714	0.193
7 (Logistic)	0.723	0.780	0.865
8 (SIC)	0.805	0.849	0.821
9 (Concave)	0.855	0.834	0.834
Odd incentives and even incentives			
	Responsive in:		
Model	At least one sample	Full sample	All three samples
1 (Constant)	0.368	0.392	−0.014
2 (Binary)	0.574	0.589	0.161
3 (Affine w/ break)	0.512	0.530	0.372
4 (Affine)	0.506	0.530	0.355
5 (Quadratic)	0.152	0.156	0.030
6 (Cubic)	0.069	0.057	−0.318
7 (Logistic)	0.397	0.405	0.444
8 (SIC)	0.529	0.539	0.284
9 (Concave)	0.492	0.480	0.366

Table A6: Two-way table of lowest-AIC model, subjects that are responsive both in the full sample and for odd incentives (27 subjects)

Odd \All	2 (Binary)	7 (Logistic)	8 (SIC)	9 (Concave)
1 (Constant)	1 (100.0%/14.3%/3.7%)	0 (0.0%/0.0%/0.0%)	0 (0.0%/0.0%/0.0%)	0 (14.3%/14.3%/3.7%)
2 (Binary)	**3 (60.0%/42.9%/11.1%)**	0 (0.0%/0.0%/0.0%)	2(40.0%/50.0%/7.4%)	0 (0.0%/0.0%/0.0%)
7 (Logistic)	2 (14.3%/28.6%/7.4%)	**10 (71.4%/90.9%/37.0%)**	2 (14.3%/50.0%/7.4%)	0 (0.0%/0.0%/0.0%)
9 (Concave)	1 (14.3%/14.3%/3.7%)	1 (14.3%/9.1%/3.7%)	0 (0.0%/0.0%/0.0%)	**5 (71.4%/100.0%/18.5%)**

Data in table's cells are: Number of subjects (Percentage of row/Percentage of column/Percentage of responsive subjects in both samples).

Table A7: Two-way table of lowest-AIC model, subjects that are responsive both in the full sample and for even incentives (31 subjects)

Even \All	2 (Binary)	7 (Logistic)	8 (SIC)	9 (Concave)
2 (Binary)	**3 (37.5%/42.9%/9.7%)**	3 (37.5%/21.4%/9.7%)	1 (12.5%/16.7%/3.2%)	1 (12.5%/25.0%/3.2%)
3 (Affine w/ break)	1 (100.0%/14.3%/3.2%)	0 (0.0%/0.0%/0.0%)	0 (0.0%/0.0%/0.0%)	0 (0.0%/0.0%/0.0%)
7 (Logistic)	2 (11.8%/28.6%/6.5%)	**9 (52.9%/64.3%/29.0%)**	5 (29.4%/83.3%/16.1%)	1 (5.9%/25.0%/3.2%)
8 (SIC)	0 (0.0%/0.0%/0.0%)	1 (100.0%/7.1%/3.2%)	**0 (0.0%/0.0%/0.0%)**	0 (0.0%/0.0%/0.0%)
9 (Concave)	1 (25.0%/14.3%/3.2%)	1 (25.0%/7.1%/3.2%)	0 (0.0%/0.0%/0.0%)	**2 (50.0%/50.0%/6.5%)**

Data in table's cells are: Number of subjects (Percentage of row/Percentage of column/Percentage of responsive subjects in both samples).

Table A8: Two-way table of lowest-AIC model, subjects that are responsive for both odd and even incentives (19 subjects)

Even \Odd	2 (Binary)	7 (Logistic)	9 (Concave)
2 (Binary)	**1 (25.0%/33.3%/5.3%)**	3 (75.0%/27.3%/15.8%)	0 (0.0%/0.0%/0.0%)
3 (Affine w/ break)	1 (100.0%/33.3%/5.3%)	0 (0.0%/0.0%/0.0%)	0 (0.0%/0.0%/0.0%)
7 (Logistic)	1 (10.0%/33.3%/5.3%)	**7 (70.0%/63.6%/36.8%)**	2 (20.0%/40.0%/10.5%)
8 (SIC)	0 (0.0%/0.0%/0.0%)	1 (100.0%/9.1%/5.3%)	0 (0.0%/0.0%/0.0%)
9 (Concave)	0 (0.0%/0.0%/0.0%)	0 (0.0%/0.0%/0.0%)	**3 (100.0%/60.0%/15.8%)**

Data in table's cells are: Number of subjects (Percentage of row/Percentage of column/Percentage of responsive subjects in both samples).

A4.2 Incentive Structure and Simulation Results

Our experiment used a fine-grained incentive structure to study subject behavior. Subjects were faced with each integer incentive level from 1 to 100. While this can give a better sense of how behavior responds to incentives than a coarser incentive structure, it comes at the expense of replication of a task for a given incentive level, thereby sacrificing power for statistical tests. In this subsection, we conduct simulations to test the power of our statistical test that classifies subjects as well-behaved or not and the reliability of our classifications of subjects by performance function. We simulate data for the fine incentive structure used in our experiment, as well as a coarse incentive structure that uses ten replications of each incentive level that is a multiple of 10 (10 each of 10, 20, 30, etc.) in order to highlight the benefits and drawbacks of our approach.

A4.2.1 Discontinuity Test

Here we present power tests for the discontinuity test introduced in Subsection 5.4. Using both the fine and the coarse incentive structures, binary data were simulated using the logistic equation $P_t = 0.2 + \frac{\varphi}{1+\exp(-\psi(r_t-\xi))}$, for various values of φ, ψ, and ξ. (Recall that continuity is the alternative hypothesis in this test.) φ controls how high the curve rises from 0.2, ψ controls the slope of the rise, and ξ is location of the center of the rise. 100 samples were taken for each (φ, ψ, ξ) tuple, and the proportion of samples for which the break was detected was calculated. The results are summarized in Table A9. The coarse incentive structure generally yields higher power than the fine one.

A4.2.2 Classification Simulations

We also wish to see how reliable our subject classifications are for both fine and coarse incentives. To that end, we simulate 100 subjects' data for various parameter values for each of logistic performance (mutual information costs), concave performance (normal signal costs), binary performance (fixed costs), and SIC performance (Tsallis costs), and calculate how many subjects are classified into each of these performance categories by AIC. Results are reported in Tables A10, A11, A12, and A13 for both fine and coarse incentives. Note that coarse incentives outperform fine ones when the true model is SIC, but when the true model is binary, fine incentives and coarse incentives lead to roughly the same rate of correct classification, and for logistic and concave performance, fine

Table A9: Power of discontinuity test (fine incentive structure on left, coarse incentive structure on right)

$\xi = 0.05$			
$\varphi \setminus \psi$	25	50	75
0.2	0.50	0.27	0.21
0.4	0.48	0.58	0.28
0.6	0.53	0.60	0.43
0.8	0.14	0.49	0.43

$\xi = 0.1$			
$\varphi \setminus \psi$	25	50	75
0.2	0.32	0.41	0.07
0.4	0.51	0.62	0.21
0.6	0.16	0.32	0.24
0.8	0.00	0.27	0.14

$\xi = 0.3$			
$\varphi \setminus \psi$	25	50	75
0.2	0.34	0.44	0.11
0.4	0.24	0.62	0.06
0.6	0.00	0.13	0.02
0.8	0.00	0.01	0.00

$\xi = 1$			
$\varphi \setminus \psi$	25	50	75
0.2	0.35	0.46	0.05
0.4	0.21	0.70	0.01
0.6	0.00	0.06	0.02
0.8	0.00	0.00	0.00

$\xi = 4$			
$\varphi \setminus \psi$	25	50	75
0.2	0.32	0.53	0.09
0.4	0.16	0.68	0.07
0.6	0.00	0.10	0.02
0.8	0.00	0.00	0.01

$\xi = 0.05$			
$\varphi \setminus \psi$	25	50	75
0.2	0.74	0.57	0.38
0.4	0.78	0.79	0.65
0.6	0.74	0.84	0.75
0.8	0.17	0.80	0.64

$\xi = 0.1$			
$\varphi \setminus \psi$	25	50	75
0.2	0.76	0.63	0.33
0.4	0.65	0.88	0.43
0.6	0.43	0.76	0.38
0.8	0.00	0.66	0.19

$\xi = 0.3$			
$\varphi \setminus \psi$	25	50	75
0.2	0.73	0.70	0.18
0.4	0.53	0.90	0.20
0.6	0.08	0.51	0.17
0.8	0.00	0.07	0.03

$\xi = 1$			
$\varphi \setminus \psi$	25	50	75
0.2	0.65	0.74	0.18
0.4	0.33	0.95	0.11
0.6	0.05	0.39	0.01
0.8	0.00	0.07	0.00

$\xi = 4$			
$\varphi \setminus \psi$	25	50	75
0.2	0.64	0.75	0.11
0.4	0.29	0.88	0.07
0.6	0.02	0.48	0.00
0.8	0.00	0.08	0.00

incentives lead to vastly higher rates of correct classification.

Table A10: Classification simulation results, mutual information (logistic performance) as true model (fine incentive structure on left, coarse incentive structure on right; α as in Model 7 of Table 5.

α	Logistic	Concave	Binary	SIC
10	0.03	0.00	0.96	0.01
15	0.34	0.01	0.59	0.06
30	0.87	0.02	0.04	0.07
45	0.93	0.05	0.02	0.00
60	0.94	0.00	0.02	0.04

α	Logistic	Concave	Binary	SIC
10	0.00	0.00	0.98	0.02
15	0.00	0.00	0.69	0.31
30	0.09	0.00	0.04	0.87
45	0.30	0.01	0.01	0.68
60	0.46	0.01	0.01	0.52

Table A11: Classification simulation results, costs linear in precision of normal signal (concave performance) as true model (fine incentive structure on left, coarse incentive structure on right; α as in Model 9 of Table 5.

α	Logistic	Concave	Binary	SIC
1	0.01	0.34	0.61	0.04
2	0.04	0.79	0.10	0.07
4.5	0.04	0.95	0.00	0.01
7	0.03	0.96	0.00	0.01
10	0.02	0.98	0.00	0.00

α	Logistic	Concave	Binary	SIC
1	0.00	0.15	0.76	0.04
2	0.02	0.15	0.12	0.73
4.5	0.00	0.21	0.00	0.79
7	0.07	0.48	0.00	0.45
10	0.12	0.64	0.00	0.24

A4.2.3 Summary

A succinct summary of this appendix subsection would be: coarse incentives are better for testing the well-behavedness of cost functions; fine incentives are better for estimating and classifying subjects according to cost functions. Since we conduct both types of analyses in this paper, this makes the choice of incentive structure somewhat arbitrary. In order to ensure the reliability of our model selection exercise, we opted for fine-grained variation in incentives.

A4.3 Probability-Weighted Incentives

One of the important aspects of our incentivization scheme is that the incentive level for each trial is the probability of winning a \$10 or \$20 prize for a correct answer, depending on the treatment. This ensures that under the assumptions of expected utility theory, subjects' utilities are linear in incentives. However, there is experimental evidence to suggest that decision-makers do not evaluate probabilities linearly (e.g. Tversky and Kahneman, 1992; Barron and Erev, 2003). For instance,

Table A12: Classification simulation results, fixed costs (binary performance) as true model (fine incentive structure on left, coarse incentive structure on right; β_1 and δ as in Model 2 of Table 5, $\beta_0 = 0.2$)

$\delta = 25$				
β_1	Logistic	Concave	Binary	SIC
0.2	0.84	0.04	0.02	0.12
0.4	0.34	0.10	0.15	0.41
0.6	0.06	0.09	0.77	0.08
0.8	0.00	0.00	1.00	0.00

$\delta = 50$				
β_1	Logistic	Concave	Binary	SIC
0.2	0.89	0.01	0.19	0.00
0.4	0.69	0.03	0.28	0.00
0.6	0.14	0.00	0.86	0.00
0.8	0.00	0.00	1.00	0.00

$\delta = 75$				
β_1	Logistic	Concave	Binary	SIC
0.2	0.71	0.00	0.29	0.00
0.4	0.45	0.02	0.53	0.00
0.6	0.08	0.00	0.92	0.00
0.8	0.00	0.00	1.00	0.00

$\delta = 25$				
β_1	Logistic	Concave	Binary	SIC
0.2	0.48	0.03	0.00	0.49
0.4	0.11	0.00	0.09	0.80
0.6	0.00	0.01	0.85	0.14
0.8	0.00	0.00	1.00	0.00

$\delta = 50$				
β_1	Logistic	Concave	Binary	SIC
0.2	0.38	0.00	0.09	0.53
0.4	0.26	0.00	0.25	0.49
0.6	0.00	0.00	0.84	0.07
0.8	0.00	0.00	1.00	0.00

$\delta = 75$				
β_1	Logistic	Concave	Binary	SIC
0.2	0.55	0.00	0.17	0.28
0.4	0.40	0.00	0.40	0.20
0.6	0.08	0.00	0.91	0.01
0.8	0.00	0.00	1.00	0.00

they may overweight small probabilities. Another possibility is that they treat probability points similarly to how they treat certain monetary rewards and are risk-averse or risk-seeking over these incentives. In this subsection, we account for these possibilities in our model selection exercise.[A24]

In the interest of ensuring the numerical stability of our estimates, we limited ourselves to two, single-parameter weighting functions. The first is the single-parameter version of Prelec's (1998) weighting function.

$$w_{\mathrm{PR}}(r) := \exp\left(-\left(\ln\left(\frac{-r}{100}\right)\right)^{\gamma}\right) \tag{A40}$$

The second treats probability weighting like a CRRA utility function of incentives:

$$w_{\mathrm{RS}}(r) := \left(\frac{r}{100}\right)^{\gamma} \tag{A41}$$

Note that both weighting functions divide incentives by 100 so that they can be properly treated

[A24]We do not perform corresponding analyses for our tests of cost function properties; continuous, monotonic probability weighting should not affect the results of those analyses.

Table A13: Classification simulation results, Tsallis costs (SIC performance) as true model (fine incentive structure on left, coarse incentive structure on right; α and σ as in Model 8 of Table 5.

$\alpha = 75$				
σ	Logistic	Concave	Binary	SIC
0.5	0.91	0.00	0.05	0.04
1.5	0.91	0.03	0.02	0.04
2	0.81	0.04	0.00	0.15
2.5	0.77	0.06	0.00	0.17
4	0.23	0.09	0.32	0.36

$\alpha = 75$				
σ	Logistic	Concave	Binary	SIC
0.5	0.41	0.00	0.01	0.58
1.5	0.32	0.02	0.01	0.65
2	0.15	0.01	0.00	0.84
2.5	0.06	0.00	0.00	0.94
4	0.00	0.01	0.40	0.59

$\alpha = 200$				
σ	Logistic	Concave	Binary	SIC
0.5	0.77	0.00	0.23	0.00
1.5	0.90	0.01	0.08	0.01
2	0.88	0.00	0.10	0.02
2.5	0.88	0.01	0.00	0.11
4	0.48	0.04	0.00	0.48

$\alpha = 200$				
σ	Logistic	Concave	Binary	SIC
0.5	0.51	0.01	0.21	0.27
1.5	0.54	0.00	0.07	0.39
2	0.58	0.00	0.03	0.39
2.5	0.53	0.00	0.00	0.47
4	0.13	0.01	0.00	0.86

$\alpha = 2000$				
σ	Logistic	Concave	Binary	SIC
0.5	0.59	0.01	0.39	0.01
1.5	0.51	0.00	0.49	0.00
2	0.70	0.01	0.28	0.01
2.5	0.75	0.00	0.25	0.00
4	0.93	0.00	0.04	0.03

$\alpha = 2000$				
σ	Logistic	Concave	Binary	SIC
0.5	0.50	0.00	0.32	0.18
1.5	0.53	0.00	0.31	0.16
2	0.49	0.00	0.25	0.26
2.5	0.49	0.00	0.27	0.24
4	0.44	0.01	0.02	0.53

as probabilities.

First we estimated each of the models of Section 6 with both probability weighting functions by maximum likelihood, excluding Models 1 (constant performance) and 2 (binary performance), because they are non-identified under probability weighting. We then found the best fit for each responsive subject for each of the two probability weighting schemes. We also found the best fit for each responsive subject selecting among the eight models with no probability weighting, the six models with probability weights given by (A40), and the six models with probability weights given by (A41). Results are summarized in Tables A14 to A17.

From Table A16, note that roughly half of the responsive subjects continue to be best fit by a model with linear probability weights. The remaining 19 responsive subjects (45.2%) are roughly evenly split between Prelec and CRRA probability weighting. Moreover, the number of subjects best fit by binary, logistic/SIC[A25], and concave performance — 6 (14.3%), 25 (59.5%), and 9 (21.4%), respectively — are nearly the same as in the model selection exercise of Section 6, where linear probability weights were assumed — 10 (23.8%), 26 (61.9%), and 6 (14.3%), respectively.

Table A17 presents results evaluating the consistency of model selection between the analysis of Section 6 and this appendix subsection. Note that most responsive subjects, 31 (73.8%), maintain the same best-fit performance function, even when probability weights are allowed to be non-linear. Moreover, the maximal entry in each column is the one for the corresponding row, indicating that not being reclassified is the most common outcome for each best-fit performance function when switching from only allowing linear weights to allowing Prelec and CRRA weights. This lends support to the validity of the model selection exercise in the main body of the paper.

[A25]Recall that logistic performance is a special case of SIC performance, since Shannon entropy is a special case of Tsallis entropy.

Table A14: Model selection for responsive subjects, Prelec probability weights

Model	Affine (3)	Quadratic (5)	Cubic (6)	Logistic (7)	SIC (8)	Concave (9)
Number of Subjects	2 (4.8%)	1 (2.4%)	1 (2.4%)	27 (64.3%)	8 (19.0%)	3 (7.1%)

Table A15: Model selection for responsive subjects, CRRA probability weights

Model	Affine (3)	Cubic (6)	Logistic (7)	SIC (8)	Concave (9)
Number of Subjects	5 (11.9%)	1 (4.2%)	23 (54.8%)	2 (4.8%)	11 (26.2%)

Table A16: Model selection for responsive subjects, all weights

Performance \Weight	Linear	Prelec	CRRA	**Total**
Binary (2)	6 (14.3%)	0 (0.0%)	0 (0.0%)	6 (14.3%)
Affine (3)	0 (0.0%)	1 (2.4%)	0 (0.0%)	1 (2.4%)
Cubic (6)	0 (0.0%)	0 (0.0%)	1 (2.4%)	1 (2.4%)
Logistic (7)	12 (28.6%)	8 (19.0%)	2 (4.8%)	22 (52.4%)
SIC (8)	3 (7.1%)	0 (0.0%)	0 (0.0%)	3 (7.1%)
Concave (9)	2 (4.8%)	2 (4.8%)	5 (11.9%)	9 (21.4%)
Total	23 (54.8%)	11 (21.4%)	8 (19.0%)	

Note: Numbers in parentheses are percentages of responsive subjects.

Table A17: Two-way table of model selection, all weights versus linear weights

All Weights \Linear	Binary (2)	Logistic (7)	SIC (8)	Concave (9)
Binary (2)	6 (100.0%/60.0%/14.3%)	0 (0.0%/0.0%/0.0%)	0 (0.0%/0.0%/0.0%)	0 (0.0%/0.0%/0.0%)
Affine (3)	0 (0.0%/0.0%/0.0%)	1 (100.0%/3.4%/2.4%)	0 (0.0%/0.0%/0.0%)	0 (0.0%/0.0%/0.0%)
Cubic (6)	0 (0.0%/0.0%/0.0%)	0 (0.0%/0.0%/0.0%)	1 (100.0%/14.3%/2.4%)	0 (0.0%/0.0%/0.0%)
Logistic (7)	3 (13.6%/30.0%/7.1%)	16 (72.7%/84.2%/38.1%)	3 (13.6%/42.9%/7.1%)	0 (0.0%/0.0%/0.0%)
SIC (8)	0 (0.0%/0.0%/0.0%)	0 (0.0%/0.0%/0.0%)	3 (12.0%/30.0%/7.1%)	0 (0.0%/0.0%/0.0%)
Concave (9)	1 (11.1%/10.0%/2.4%)	2 (22.2%/10.5%/4.8%)	0 (0.0%/0.0%/0.0%)	6 (66.7%/100.0%/14.3%)

Data in table's cells are: Number of subjects (Percentage of row/Percentage of column/Percentage of responsive subjects).

A4.4 The Cheremukhin et al. (2015) Generalization of Mutual Information

Cheremukhin et al. (2015) generalize the Shannon entropy-based cost function by allowing for convex transformations of mutual information. Letting H represent mutual information and assuming that perfectly uninformative information structures are free, they define the function by its derivative:[A26]

$$K'(H) = \frac{\bar{\theta}\pi}{\text{arccot}(\rho(H - \bar{\kappa}))} \tag{A42}$$

where $\bar{\theta}$, ρ, and $\bar{\kappa}$ are non-negative parameters. $\bar{\theta}$ is a multiplicative factor that regulates the marginal cost of information, and ρ regulates the curvature of the derivative. K' is relatively flat for $H < \bar{\kappa}$, so a higher ρ indicates a sharper increase in the marginal cost of information going from $H < \bar{\kappa}$ to $H > \bar{\kappa}$. Therefore, for large ρ, $\bar{\kappa}$ can be interpreted as a capacity constraint on the DM's ability to acquire and/or process information.

However, it should be noted that a near-constant marginal cost of information for $H < \bar{\kappa}$ is not the same thing as a near-constant marginal cost of performance; information is non-linear in performance, as can be seen from (11). Essentially, when ρ is high, the near-constant marginal cost of information for $H < \bar{\kappa}$ makes the cost of performance approximately mutual information for $P \leq P_{\bar{\kappa}}$, where $P_{\bar{\kappa}}$ is defined such that $\ln(5) + P_{\bar{\kappa}} \ln(P_{\bar{\kappa}}) + (1 - P_{\bar{\kappa}}) \ln\left(\frac{1-P_{\bar{\kappa}}}{4}\right) = \bar{\kappa}$. Because the marginal cost increases sharply above $P_{\bar{\kappa}}$, it is effectively the highest level of performance that the DM can achieve. Therefore, the DM's performance is approximately logistic up until the r that induces $P_{\bar{\kappa}}$, after which it is almost completely flat. Marginal cost curves for different values of $\bar{\theta}$, $\bar{\kappa}$, and ρ are displayed in the left panels of Figures A4, A5, and A6.

By applying the chain rule, we can rewrite (A42) in terms of performance in uniform guess tasks as:

$$C'(P) = \frac{\bar{\theta}\pi(\ln(P) - \ln(1 - P) + \ln(4))}{\text{arccot}\left(\rho\left(\ln(5) + P\ln(P) + (1 - P)\ln\left(\frac{1-P}{4}\right) - \bar{\kappa}\right)\right)} \tag{A43}$$

Since $\text{arccot}(0) = \frac{\pi}{2}$, this model nests mutual information for $\rho = 0$, taking $\alpha = 2\bar{\theta}$. In general,

[A26] In the definition of the corresponding function in Equation (7) of Cheremukhin et al. (2015), ρ is the reciprocal of what it is here. However, the convention we adopt here is consistent with Figure 2 and Footnote 10 of Cheremukhin et al. (2015), as well as computer code provided by the authors.

Table A18: Model Selection for Responsive Subjects, Including Cheremukhin et al. (2015) Cost Functions

Model	Binary (2)	Logistic (7)	SIC (8)	Concave (9)	CRL
Number of Subjects	10 (23.8%)	18 (42.9%)	6 (14.3%)	5 (11.9%)	3 (7.1%)

(A43) cannot be inverted to obtain a closed form for the performance function. However, the performance function can be graphed, as it is for various values of $\bar{\theta}$, $\bar{\kappa}$, and ρ in the right panels of Figures A4, A5, and A6.

We estimate two versions of (A43), one with ρ restricted to a value of 600, as in Cheremukhin et al. (2015), and one where ρ is allowed to vary freely. We refer to the corresponding performance functions as "capacity-restricted logistic" (CRL) and "generalized logistic" (GL), respectively.

Repeating the model selection exercise of Section 6 with these two additional cost functions does not substantially alter our results (see Table A18. According to the AIC criterion, no subjects are best fit by the model with flexibly estimated ρ, and only three subjects are best fit by the model with restricted ρ. Since the exact choice of ρ was not theoretically motivated aside from being large enough to make the marginal cost of information nearly vertical to the right of $\bar{\kappa}$, we regard the ρ-restricted model of Cheremukhin et al. (2015) to be a reasonable way of describing these three subjects' data that ensures the differentiability of the cost function, but not in a way that strongly distinguishes predicted behavior from the other cost functions we estimate.

Estimating the flexible-ρ version of (A43) allows us to see if setting a very high ρ is a reasonable assumption to make for all subjects, as Cheremukhin et al. (2015) do. Figure A7 is a histogram of $\log_{10}(1+\hat{\rho})$ for responsive subjects. The estimates in the leftmost bin are actual zeroes, indicating a constant marginal cost of mutual information. Therefore, 13 responsive subjects would have their performance estimated to be logistic even under the flexible model. Looking at the second bin, for an additional 17 subjects, the estimated curvature parameter lies between 0 and 9. Therefore, the majority of responsive subjects do not have curvature parameters in the ranges suggested by Cheremukhin et al.'s (2015) results; only five have a $\hat{\rho}$ that exceeds 99. Our interpretation of these results is that value-based decision-making and effortful perceptual tasks are not governed by the same informational processes.

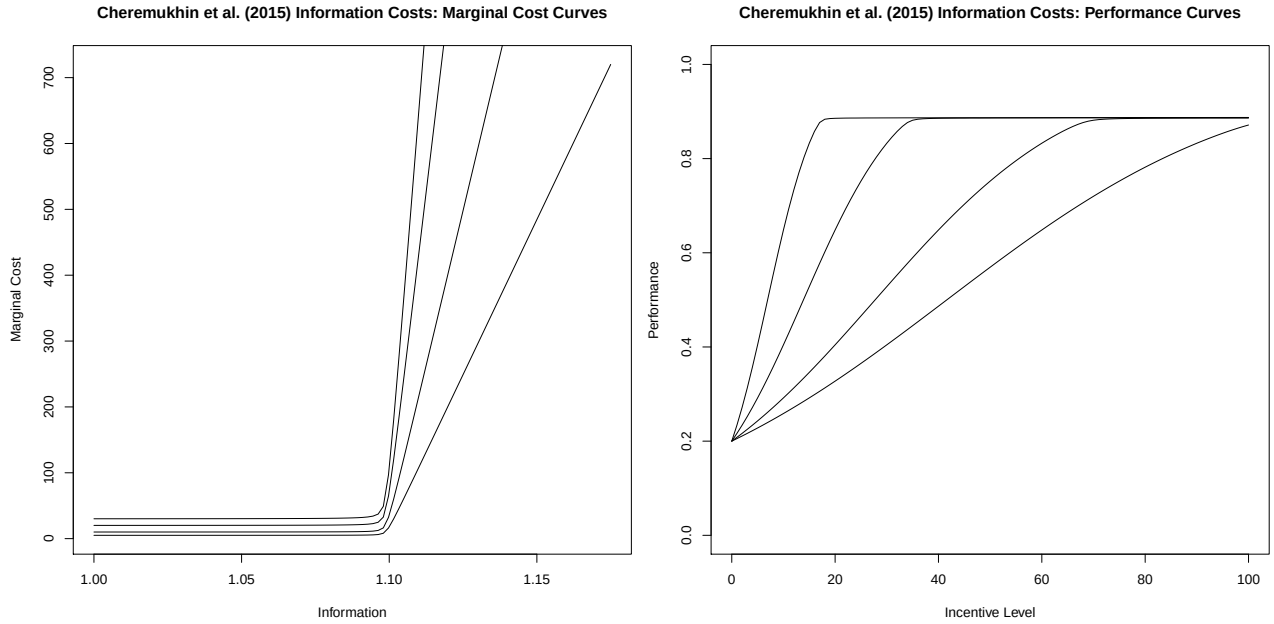


Figure A4: Cheremukhin et al. (2015) costs, $n = 5$, $\bar{\theta} \in \{5, 10, 20, 30\}$, $\rho = 600$, $\bar{\kappa} = \ln(3)$. The left panel shows marginal costs with respect to information for increasing values of $\bar{\theta}$ going counter-clockwise from the bottom-right. The right panel shows performance curves for increasing values of $\bar{\theta}$ going from the top-left to the bottom-right.

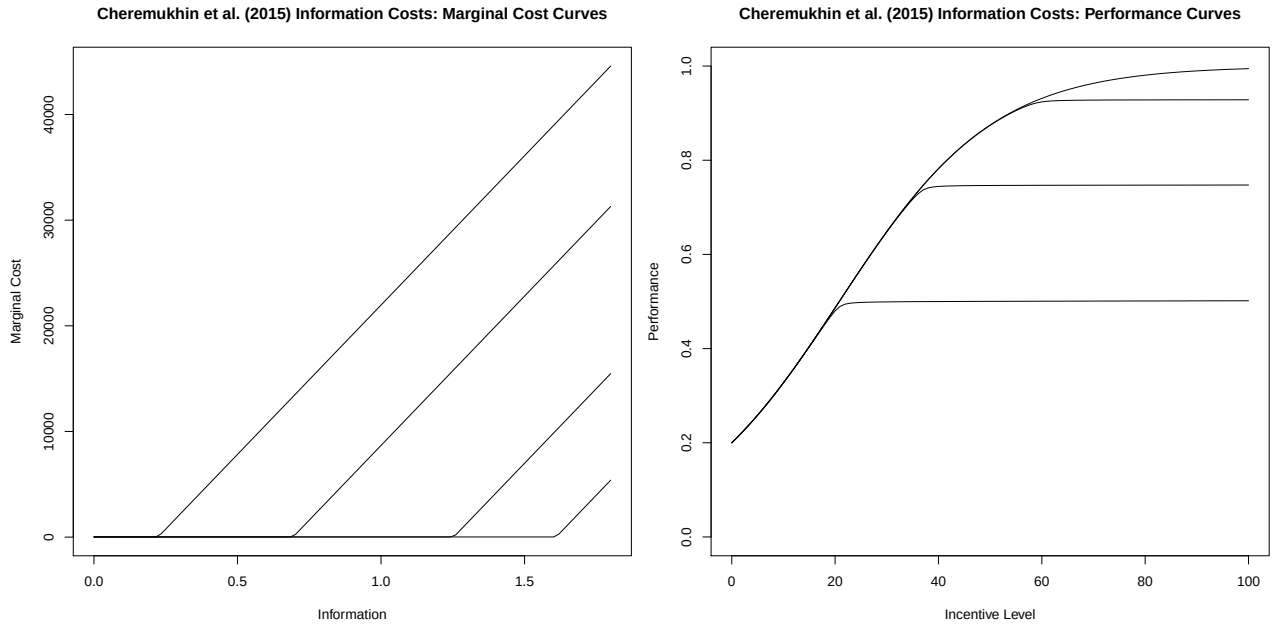


Figure A5: Cheremukhin et al. (2015) costs, $n = 5$, $\bar{\theta} = 15$, $\rho = 600$, $\bar{\kappa} \in \{\ln(1.25), \ln(2), \ln(3.5), \ln(5)\}$. The left panel shows marginal costs with respect to information for increasing values of $\bar{\kappa}$ going from the top-left to the bottom-right. The right panel shows performance curves for increasing values of $\bar{\kappa}$ going from the bottom to the top.

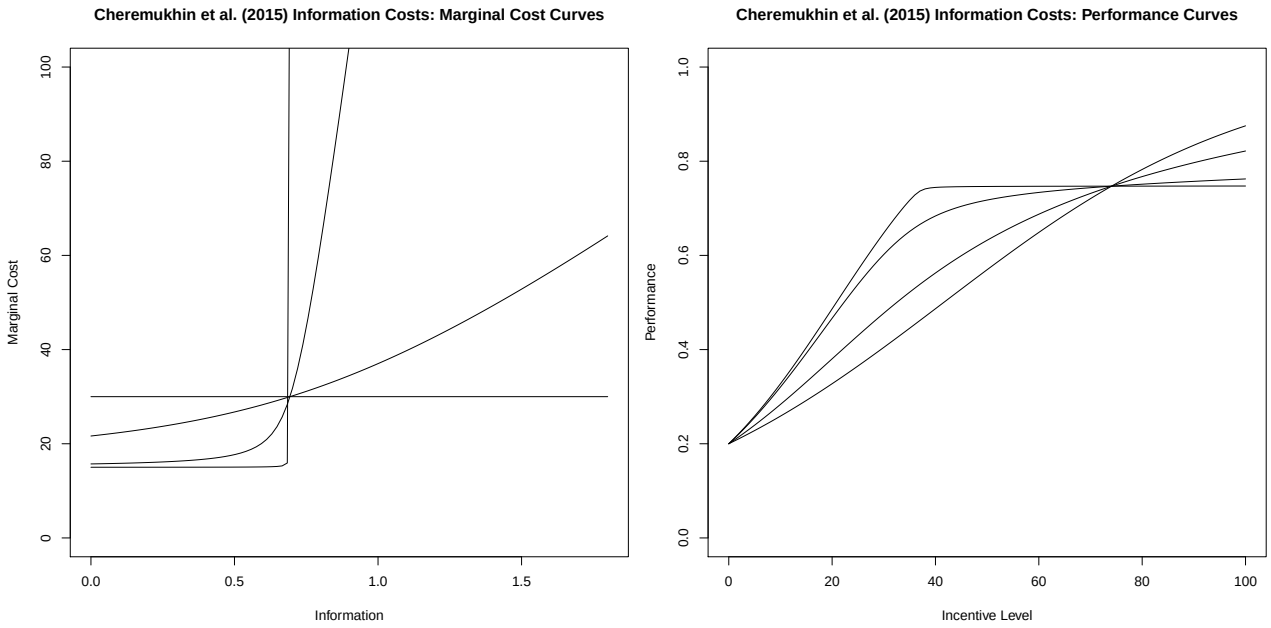


Figure A6: Cheremukhin et al. (2015) costs, $n = 5$, $\bar{\theta} = 15$, $\rho \in \{0, 1, 10, 600\}$, $\bar{\kappa} = \ln(2)$. The left panel shows marginal costs with respect to information for increasing values of ρ going counter-clockwise from the right. The right panel shows performance curves for increasing values of ρ going from the bottom to the top, between the points of intersection.

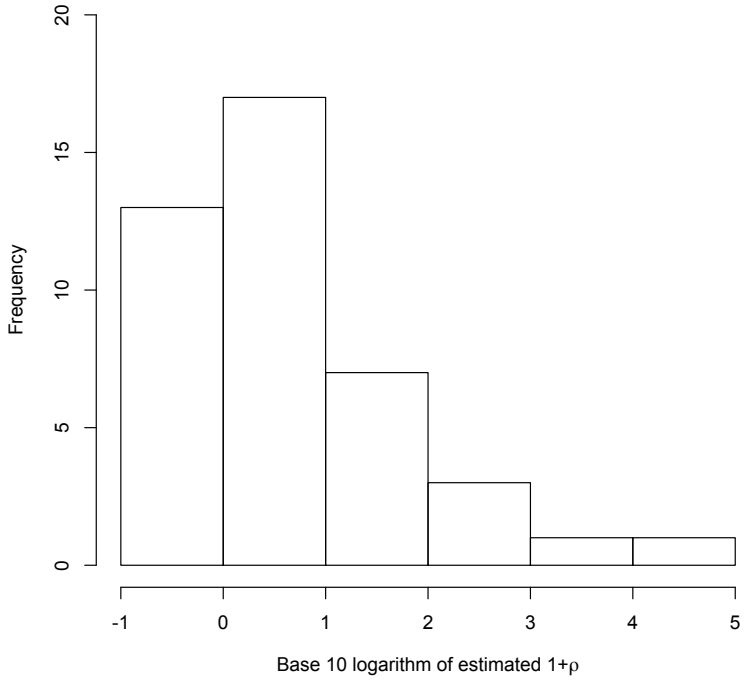


Figure A7: Histogram of $\log_{10}(1 + \hat{\rho})$ for responsive subjects

A4.5 Bayesian Hierarchical Modeling

Because we are looking at the population as a mixture of types, it seems natural to approach the data using a Bayesian hierarchical model. We analyze the data by constructing a hierarchical Bayesian prior and then determining each subject's posterior probability of belonging to each type using Markov Chain Monte Carlo methods. In our analysis, "types" refers to the form of performance function a subject has, as per the results in Section 6.

The prior distribution has two components: a distribution over the population proportion of types; and a distribution over performance function parameters for each type. For the distribution over population proportion of types we choose a Dirichlet distribution with all parameters set to one. The resulting distribution is uniform over the simplex, and it is considered to be the standard prior over proportions of types for hierarchical Bayesian analysis.

The parameter prior distributions are constructed in a manner that requires some explanation. The standard method for generating the prior distribution over parameters within each group generally involves assuming that the prior over parameters comes from a specific family (like the normal or beta) and then fitting the hyperparameters of the distribution using maximum likelihood methods. Unfortunately, this approach is not practical in our setting, because we want our prior to only include non-decreasing performance functions. The parameter spaces for non-decreasing affine with break, cubic, and quadratic performance functions have very difficult structures, and so no standard distribution could be used to cover these spaces.

We instead use a somewhat *ad hoc* method based on the sequential drawing of model parameters. There are methods for drawing parameters of non-decreasing performance functions of all the classes we consider by drawing all of the parameters in a specific order from parameter distributions whose range depends on the previously drawn parameters. When the range of the parameter distribution is bounded below, we draw from a transformed gamma distribution, and when it is bounded both above and below, we draw from a transformed beta distribution. (Full details are available from the authors on request.) This gives us a method for randomly drawing performance functions.

To calibrate the distributions we draw from, we invert the process, converting the estimated parameters for each individual and model type (see Section 6) into draws from the transformed distributions and then inverting the transformation to get a draw from a standard distribution of the correct type. These draws for each distribution are then collected, and we find model parameters

that match the mean and variance of the observed distribution. These model parameters are used to assign a model prior.

Consider the generation of the prior over binary performance functions. The binary model has three parameters: β_1, a low performance level; β_2, a high performance level; and δ, the break point where it switches from one to the other. We can construct a non-decreasing binary performance function by drawing β_1 from one beta distribution (A) on the $[0.2, 1]$ interval, drawing δ from a beta distribution (B) on the $[5, 95]$ interval, and then drawing β_2 from a beta (C) on the $[\beta_1, 1]$ interval. We draw the function by sampling from beta distributions and then applying the affine transformation mapping the $[0, 1]$ interval into the correct interval.

To generate the parameters for each of these beta distributions, we look at the parameters estimated for the binary model for different subjects. For example, to get the parameters for (C), we convert the estimated $\hat{\beta}_2$ into draws from a beta by dividing by $1-\hat{\beta}_1$ and subtracting $\hat{\beta}_1$. Then, we look at the mean and standard deviation of these standard draws for all the subjects examined and use the resulting moments to fit distribution (C). Note that means and standard deviation fully pin down a beta distribution or gamma distribution. We then do the same with $\hat{\delta}$ to find the parameters of (B) and $\hat{\beta}_1$ to find the parameters of (A).

Note that we do not use all subjects to fit all models, because it makes little sense to use data from a subject who has a distinctly binary performance function to calibrate the prior for individuals with a logistic performance function (linear Shannon costs). We instead use the N subjects best fit by a particular model,using likelihood as a measure of fit, to calibrate the priors for that model. We vary N and report the results below.

Once the priors have been assigned, the updating process is fairly simple. We employ a very simple component-wise Metropolis-Hastings algorithm using the prior as our proposal distribution (cf. Chapter 11 of Gelman et al., 2003). We construct a function f which given a subject i, set of model parameters $\mathfrak{m}$, and probability of each model of type ψ will return the likelihood of that subject's observed data given those parameters model probabilities. Note that $\mathfrak{m}$ includes model parameters for every model being considered and each subject. Call the parameters for subject i's models $\mathfrak{m}^i$. We then run the following algorithm for each subject i.

1. Pick a random model proportion and set of model parameters for each subject from the proposal distribution. Call them the current model proportion ψ_c and set of model parameters

$\mathfrak{m}_c$. Get the resulting likelihood $f(i, \psi_c, \mathfrak{m}_c)$, and call it the comparison likelihood l_c.

2. Get a new random model proportion ψ_n. Check the new likelihood $l_n = f(i, \psi_n, \mathfrak{m}_\psi)$.

3. Accept the new proportions with probability $\min\left\{1, \frac{l_n}{l_c}\right\}$. If the new proportions are accepted, store $\psi_n, \mathfrak{m}_c$ in the trace and set $\psi_c = \psi_n$, $l_c = l_n$

4. Get a new random set of model parameters for subject 1, $\mathfrak{m}_n^1$ from the prior. Define $\mathfrak{m}_n$ as a set of parameters that is identical to $\mathfrak{m}_c$ but we replace $\mathfrak{m}_c^1$ with $\mathfrak{m}_n^1$

5. Check the new likelihood $l_n = f(i, \psi_c, \mathfrak{m}_n)$.

6. Accept the new proportions with probability $\min\left\{1, \frac{l_n}{l_c}\right\}$. If the new proportions are accepted, store $\psi_c, \mathfrak{m}_n$ in the trace and set $\mathfrak{m}_c = \mathfrak{m}_n$, $l_c = l_n$

7. Repeat steps 4–6 for all responsive subjects

8. Repeat steps 2–7 10000 times

We then throw away the early values in the trace. The remaining values provide an approximation of the posterior for the subject. Note that we do not have to modify or weight the trace values to get the posterior in the case, because we use the prior distribution as our proposal distribution (Chib and Greenberg, 1995).

Unfortunately, the resulting distribution over population proportions is difficult to reasonably visualize. Instead, in Table A19 we report the mean posterior probability for each model for various values of N. As the table shows, the results are somewhat sensitive to the choice how the parameter priors are fit. However, there are some consistent findings. In particular, the logistic (Shannon costs), SIC (Tsallis costs), and cubic (costs on the order of $P^{\frac{4}{3}}$) models all perform well across all values of N. The strong performance of the logistic and SIC models is generally unsurprising. The high performance of the cubic model is more unusual and may be at least in part due to the somewhat restrictive method we use to draw the monotone cubic performance function. The binary and affine break models all perform well for some N, but both are sensitive to assumptions about how the parameter priors are fit. This likely relates to the fact that the likelihoods for these models are very sensitive to the break location parameter. Constant, linear, quadratic, and concave (normal signal costs) models never perform well. The general failure of the

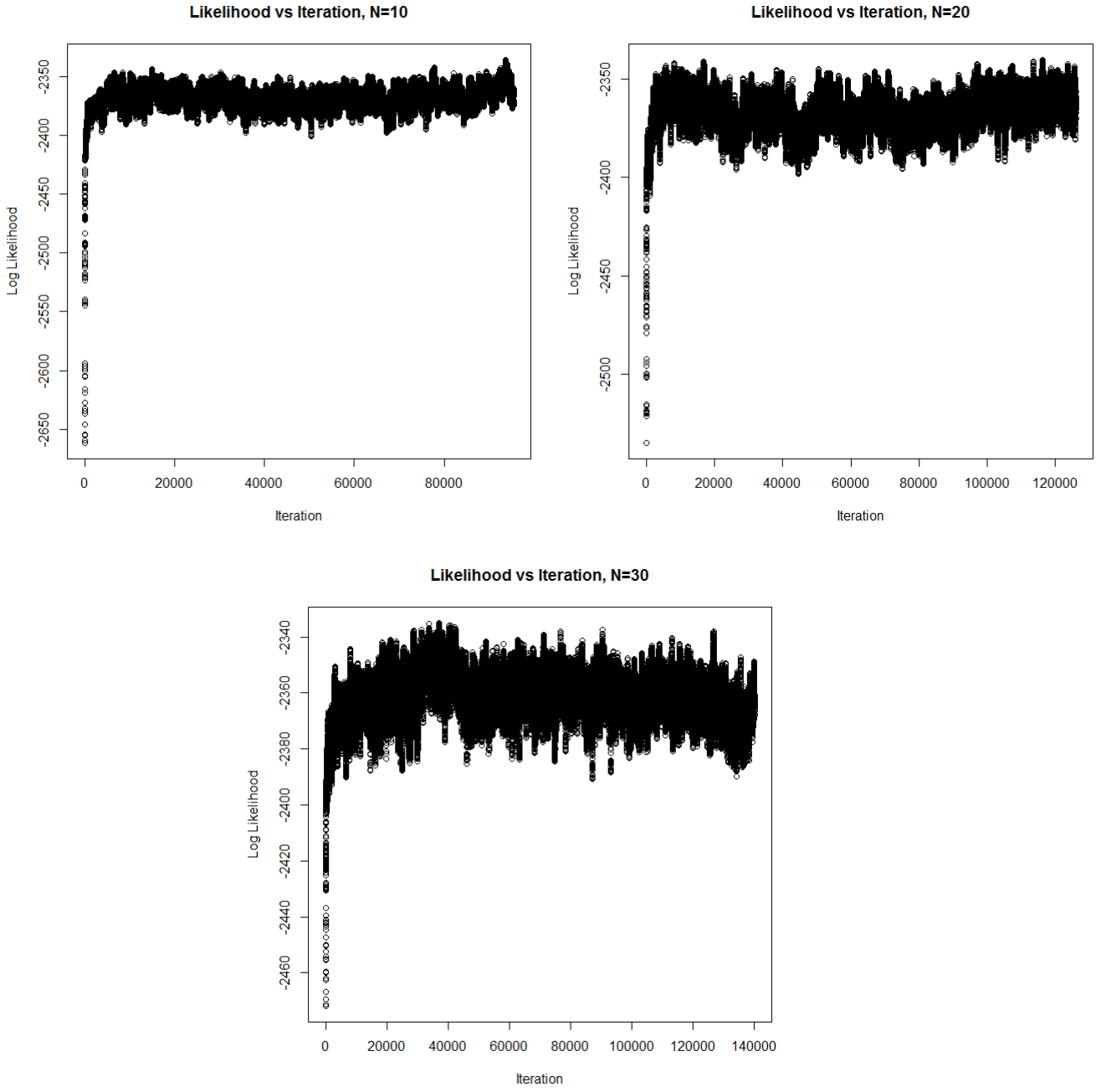


Figure A8: Evolution of likelihood over iterations of the Metropolis-Hastings algorithm

Table A19: Average posterior probabilities for different values of N

	N		
	10	20	30
Constant	0.016	0.015	0.015
Binary	0.077	0.049	0.108
Affine with Break	0.200	0.052	0.135
Affine	0.038	0.082	0.069
Quadratic	0.047	0.047	0.017
Cubic	0.216	0.285	0.180
Logistic	0.121	0.178	0.253
SIC	0.259	0.265	0.195
Concave	0.022	0.021	0.023

normal model is interesting, and may be related to the high sensitivity in that model to the cost parameter.

It should be noted that these estimates may not be perfectly reliable for high N. As we can see from the likelihood graphs in Figure A8, while the process does converge fairly quickly in all cases, continuing regions of high and low likelihood suggest a multimodal posterior. Metropolis-Hastings algorithms can sometimes take a very long time to fully cover these types of posteriors, since transitioning between modes can take many iterations. However, consistency of results accross multiple runs of the code leads us to believe that this problem does not have a substantial impact on mean model likelihoods.

References

Greg Barron and Ido Erev. Small feedback-based decisions and their limited correspondence to description-based decisions. *Journal of Behavioral Decision Making*, 16(3):215–233, 2003.

Patrick Billingsley. *Probability and Measure*. Wiley, 1995.

David Blackwell. Equivalent comparisons of experiments. *Annals of Mathematical Statistics*, 24 (2):265–272, 1953.

Andrew Caplin and Mark Dean. Revealed preference, rational inattention, and costly information acquisition. *American Economic Review*, 105(7):2183–2203, 2015.

Andrew Caplin and Daniel Martin. A testable theory of imperfect perception. *The Economic Journal*, 125(582):184–202, 2015.

Andrew Caplin, Mark Dean, and John Leahy. Rationally inattentive behavior: Characterizing and generalizing Shannon entropy. *Working paper*, 2019.

Anton Cheremukhin, Anna Popova, and Antonella Tutino. A theory of discrete choice with information costs. *Journal of Economic Behavior & Organization*, 113:34–50, 2015.

Siddhartha Chib and Edward Greenberg. Understanding the Metropolis-Hastings algorithm. *The American Statistician*, 49(4):327–335, 1995.

Andrew Gelman, John B. Carlin, Hal S. Stern, and Donald B. Rubin. *Bayesian Data Analysis, Second Edition*. Taylor & Francis, 2003.

Emir Kamenica and Matthew Gentzkow. Bayesian persuasion: Web appendix. *American Economic Review*, 2010. URL https://assets.aeaweb.org/assets/production/articles-attachments/aer/data/oct2011/20090933_app.pdf.

Emir Kamenica and Matthew Gentzkow. Bayesian persuasion. *American Economic Review*, 101 (6):2590–2615, 2011.

Moshe Leshno and Yishay Spector. An elementary proof of Blackwell's theorem. *Mathematical Social Sciences*, 25(1):95–98, 1992.

Filip Matějka and Alisdair McKay. Rational inattention to discrete choices: A new foundation for the multinomial logit model. *American Economic Review*, 105(1):272–298, 2015.

C.B. McGuire. Comparisons of information structures. In C.B. McGuire and Roy Radner, editors, *Decision and Organization: A Volume in Honor of Jacob Marschak*, chapter 5, pages 101–130. North-Holland Publishing Company, Amsterdam, 1972.

Luciano Pomatto, Philipp Strack, and Omer Tamuz. The cost of information. *Working paper*, 2019.

Drazen Prelec. The probability weighting function. *Econometrica*, 66(3):497–528, 1998.

Walter Rudin. *Real and Complex Analysis*. McGraw-Hill, 1974.

Knut Sydsæter, Peter Hammond, Atle Seierstad, and Arne Strøm. *Further Mathematics for Economic Analysis*. Financial Times Prentice Hall, 2008.

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

Supplementary Appendices for "Estimating Information Cost Functions in Models of Rational Inattention"

Not For Online Publication

Ambuj Dewan and Nathaniel Neligh

October 1, 2019

S1 Perceptual Distance

S1.1 Preliminaries

In uniform guess tasks, where the action space is identified with the state space, the "dissimilarity" between two actions is the same as the "dissimilarity" between the corresponding states. Perceptual distance refers to the notion that distant states are easier to distinguish from each other than nearby ones. For example, if $\Theta = \{1, 2, 3, 4, 5\}$, and the true state is $\theta = 2$, then the DM may be more likely to answer 1 (which is 1 away from 2) than she is to answer 5 (which is 3 away from 2). This is especially plausible if the states in Θ represent physical, measurable quantities. To give a more concrete example, when shopping for televisions, one is much more likely to misperceive a 27-inch screen as a 23-inch screen than as a 40-inch screen. We formalize this notion below.

Definition 1. Let ρ be a metric on Θ. Then in this task, the DM evinces *perceptual distance* iff $\forall\, x, y, z \in \Theta, \rho(x, y) > \rho(x, z) \implies \Pr(a = y|\theta = x) < \Pr(a = z|\theta = x)$.

In other words, the DM evinces perceptual distance if for each possible true state, she is more likely to choose an answer (i.e. an action) close to the true state than one farther away from it.

Though one can define a metric on a given set in many different ways, it makes sense to take ρ to be a "natural" metric on Θ. For instance, if Θ is a subset of the real line as in the example above, then absolute value, $\rho(x, y) = |x - y|$, may be a sensible metric to use. Since the state space in

our experiment is such a subset, absolute value is the metric we use in analyzing our experimental results.[S1]

S1.2 Cost Functions

S1.2.1 Entropy-Based Cost Functions

Recall from the proof of Proposition 5 in Appendix Subsection A2.5 that in a uniform guess task, the probability of guessing any given incorrect state is independent of the state when the cost function is entropy-based. What that means is that in our experiment, a subject who whose cost function is mutual information (implying logistic performance) should not evince perceptual distance. For example, if the true number of dots is 39, reporting 42 should be just as likely as reporting 38 for an individual with a mutual information cost function, even though the difference between 42 and 39 is 3, whereas the difference between 38 and 39 is 1. As we explain below, not all subjects with logistic performance fail to evince the perception of the distance. We can reconcile these observations with a mutual-information-like cost function that implies logistic performance but depends directly on performance as in Appendix Subsection A1.5, with Ω chosen to match the observed distribution of mistakes. See, for instance, (11) in Subsection 3.2 for the relevant functional form.

S1.2.2 Normal signals

In Subsection 3.3, we assumed that adjacent states were equally spaced. The state space in our experiment also has this property. This assumption of equidistant states allows us to draw some conclusions about whether a DM who receives normal signals necessarily evinces the perception of distance. The answer, in general, is no. This is because the lowest possible state θ_1 is guessed for any signal $\hat{m} \leq \frac{1}{2}(\theta_1 + \theta_2)$.[S2] If the costs of precision are very high, so that the DM selects a very low signal precision, then her distribution of signals may have fat enough tails that for some true state, guessing the lowest state is likelier than guessing the next outermost state, i.e. $\Pr\left(\hat{m} \leq \frac{1}{2}(\theta_1 + \theta_2) \middle| \theta = \theta_j\right) > \Pr\left(\hat{m} \in \left[\frac{1}{2}(\theta_1 + \theta_2), \frac{1}{2}(\theta_2 + \theta_3)\right] \middle| \theta = \theta_j\right)$ for some $j \geq 2$.

However, while we cannot conclude that a DM with normal signals necessarily evinces the perception of distance over the entire state space, we can say that she does if we restrict our focus

[S1] This would also hold for any strictly monotonically increasing transformation of ρ that preserves its metric properties on Θ.

[S2] A symmetric argument applies to the highest possible state.

to guesses of inner states (i.e. states θ_2 to θ_{n-1}).

Proposition S1. *In a uniform guess task with equidistant states, a DM with normal signals evinces the perception of distance for guesses of inner states; that is to say,* $\forall\, x \in \Theta$ *and* $\forall\, y, z \in \Theta \setminus \{\theta_1, \theta_n\}$, $|x - y| > |x - z| \Longrightarrow \Pr(a = y|\theta = x) < \Pr(a = z|\theta = x)$.

Proof. To proceed, we need a lemma:

Lemma S1. *Let* β, ζ, *and* η *be strictly positive. Then* $\Phi((\xi + \beta)\zeta\eta) - \Phi(\xi\zeta\eta)$ *is strictly decreasing in* ξ *for positive* ξ *and strictly increasing for in* ξ *for negative* ξ.

This lemma is easily proven by differentiating to obtain $\zeta\eta[\phi((\xi + \beta)\zeta\eta) - \phi(\xi\zeta\eta)]$. Since the normal density is decreasing on the positive real line and increasing on the negative real line, this derivative is negative for positive ξ and positive for negative ξ.

For guesses of inner states that are not the true state, the result follows from setting $\xi = 2k + 1$ and $\beta = 2$ for $k \neq -1$ and comparing it to the expression in Lemma S1 when $\xi = 2k + 3$. This shows that guessing an inner state that is not the true state is likelier than guessing the inner state that is immediately farther from it. Applying this logic iteratively and exploiting the symmetry of the normal distribution to compare guesses of inner states on opposite sides of the true state gives the result.

In order to show that guessing the true state is likelier than guessing any other inner state, assume that the true state is not θ_{n-1} or θ_n, so that state immediately above the true state is also an inner state. (An obvious symmetric argument applies in case the true state is θ_{n-1} or θ_n.) Lemma S1 implies that:

$$\begin{aligned} &\Phi(\zeta\eta) - \Phi(0) > \Phi(2\zeta\eta) - \Phi(\zeta\eta) \quad \text{and} \quad \Phi(\zeta\eta) - \Phi(0) > \Phi(3\zeta\eta) - \Phi(2\zeta\eta) \\ \Longrightarrow\ &2[\Phi(\zeta\eta) - \Phi(0)] > \Phi(3\zeta\eta) - \Phi(\zeta\eta) \\ \Longrightarrow\ &\Phi(\zeta\eta) - \Phi(-\zeta\eta) > \Phi(3\zeta\eta) - \Phi(\zeta\eta) \end{aligned}$$

Since the probability of guessing the true state is at least $\Phi(\zeta\eta) - \Phi(-\zeta\eta)$ (the true state could be the lowest state), combining this implication with the result for inner states that are not the true state proves the result. □

S1.3 Results

For each subject and trial t, we compute the error distance $\rho(a_t, \theta_t) = |a_t - \theta_t|$. In our experiment this distance in an integer in $\{1, 2, 3, 4\}$. In order to test for perceptual distance, for each responsive subject we compute the distribution of error distances that would be predicted if the subject were to be equally likely to make any mistake in each state, given the empirically observed distribution of true states and the subject's overall accuracy rate. We then compare the empirically observed distribution of error distances to this distribution using a chi-squared test.

At the 5% level, we find that 26 out of 42 responsive lab subjects (61.9%) have a distribution of mistakes that evinces perceptual distance. Of course, the notion that perceptual distance matters for error distance distributions is not limited to responsive subjects; mutual information implies responsiveness (i.e. a strictly increasing performance function), so subjects who are not responsive have already rejected mutual information for other reasons. But as a test of the general notion that each possible mistake is equally likely given an true state of nature, it is worth running these tests on the entire pool of rationally inattentive subjects. At the 5% level, we find that 45 out of 70 rationally inattentive lab subjects (64.3%) reject the null hypothesis of not perceiving distance.

S2 Demographics, Aggregate Results, and Categorization

In our experiment, demographic data were collected in a brief post-experiment questionnaire (but before feedback was given). In this appendix, we provide a summary of these data and aggregate results of our subjects. We then determine the extent to which demographic covariates predict the categorization of subjects as rationally inattentive and responsive, as well as what their best-fitting performance function is.

S2.1 Demographic Data

Table S1 lists basic demographic data for the laboratory subjects. The pool is fairly gender-balanced;[S3] the null of perfect gender balance cannot be rejected (two-sided test of proportions, $p = 0.146$). The pool is also highly educated; over 55% of laboratory subjects have completed

[S3] Subjects were given the option to list their gender as "other/non-binary." No subjects used this option, though one subject declined to disclose their gender.

Table S1: Laboratory Demographics

Number of subjects	$n = 81$
Gender ($n = 80$)	41.3% male; 58.8% female
Age ($n = 80$)	Average: 23.00; St. dev.: 4.17
Highest level of education achieved ($n = 81$)	
Some post-secondary	44.4%
Completed bachelor's degree	29.6%
Completed graduate or professional degree	25.9%
Area of study ($n = 80$)	
Economics, psychology, or neuroscience	24.7%

a post-secondary degree. In general, demographic characteristics are not strong determinants of subjects' behavior in this experiment.

S2.2 Aggregate Analysis

Table S2 displays a regression of correctness on incentive level. The regression in column 2 includes demographic covariates, including age (in years) and dummies for maleness, holding at least a bachelor's degree, studying economics, psychology, or neuroscience, participating in the $20 prize treatment, and being shown the "dots" tasks before the "angle" tasks. It also controls for the order in which tasks were completed.

It is apparent that in the aggregate, performance is higher at higher incentive levels. In particular, on average each increase of 1 point in incentive level results in a 0.3% increase in the probability of answering correctly.

For the most part, demographic covariates have no significant effect on performance. Moreover, there is no significant effect of doing the "dots" tasks before the "angle" tasks. However, performance does decline slightly over time, indicating that subjects may experience some fatigue.[S4]

S2.3 Rational Inattentiveness

In Figure S1, we present a histogram of the p-values of the monotonicity test of Doveh et al. (2002) used to determine whether subjects adhere to the NIAC condition. (Recall from Proposition 1 that

[S4]The effect of task number on performance vanishes if we only consider the second half of the data, i.e. the last 50 tasks for each subject. (Recall that the first fifty tasks contained the odd-numbered incentives, and the last fifty tasks contained the even-numbered incentives, so each half of the data contains the same range variation in incentives as the whole data set.) This is consistent with some portion of the subjects choosing to exert effort early in the experiment before succumbing to fatigue. We examine the consistency of results between both halves of the data in the next appendix section.

Table S2: Regressions of correctness on incentive level and demographic covariates

	(1)	(2)
Incentive Level	0.003^{***}	0.003^{***}
	(0.0004)	(0.0004)
Age		-0.0001
		(0.006)
Male		0.004
		(0.056)
Bachelor's		-0.062
		(0.058)
Econ/Psych/Neuro		-0.097^{*}
		(0.054)
$20 Prize		0.023
		(0.049)
Dots First		0.049
		(0.052)
Task Number		-0.001^{***}
		(0.0003)
Constant	0.425^{***}	0.498^{***}
	(0.032)	(0.140)
Observations	7900	7900
R^2	0.03799	0.05635

Note: *p<0.1; **p<0.05; ***p<0.01
Standard errors clustered on subject.

this is equivalent to testing for positive monotonicity.) We implemented the test with a significance level of 5%, so we reject NIAC only for the subjects whose p-values fall in the leftmost cell of the histogram. Note that roughly half of subjects have p-values below 0.5, whereas the other half have p-values greater than 0.95. Examination of the data confirm that the p-values for the latter half of subjects are all indeed 1. This is not surprising: these are the subjects for whom a positive derivative restriction on their estimated performance is non-binding; their estimated coefficients are the same regardless of whether the derivative restriction is imposed.

We now turn our focus to the NIAS tests. Recall from Subsubsection 5.2.2 that for each subject, we run as many hypothesis tests as different actions they took. For most subjects, this means running 5 tests, but if, for instance, a subject never selected 39, then we would only run 4 tests for them. If a subject rejects the null of posterior maximality at the true state for at least one action, then we classify them as rejecting NIAS. Therefore, we are interested in the *minimum* p-value for each subject. We present a histogram of these p-values in Figure S2. Note that the distribution of p-values is unimodal, spiking at the right tail of the distribution.

To determine the extent to which demographics predict a subject's classification as rationally inattentive, we run a logit regression of an indicator for rational inattentiveness on demographic covariates. These covariates are age, an indicator for being male, an indicator for having attained at least a bachelor's degree, an indicator for studying economics, psychology, or neuroscience, an indicator for participating in the \$20 prize treatment, and an indicator for having done the dots tasks first. We display the results of this regression in column 1 of Table S3.

Demographic covariates do not seem to be predictive of rational inattentiveness in this particular subject pool. Neither do experimental variables, such as the higher prize and completing the dots tasks first. This suggests that for a given set of tasks, rational inattentiveness is an innate characteristic that is not well captured by demographics, and moreover, it may be difficult to manipulate experimentally.

S2.4 Responsiveness

We test for responsiveness with three tests, each of which generates a p-value. As explained in Subsection 5.3, we classify subjects as responsive if they reject the null hypothesis for at least one of these three tests. In other words, a subject is classified as responsive if the *minimum*

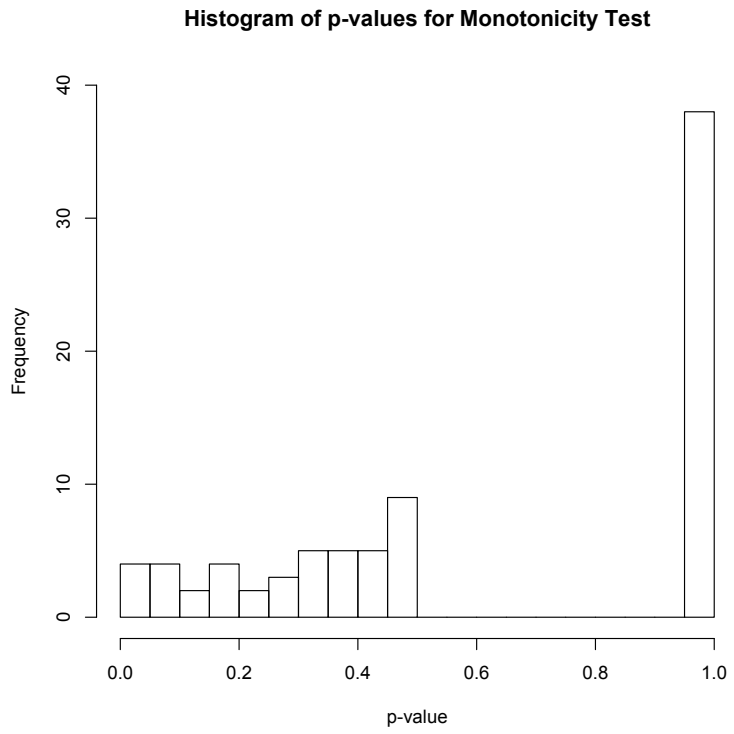


Figure S1: Histogram of p-values for the Doveh et al. (2002) test

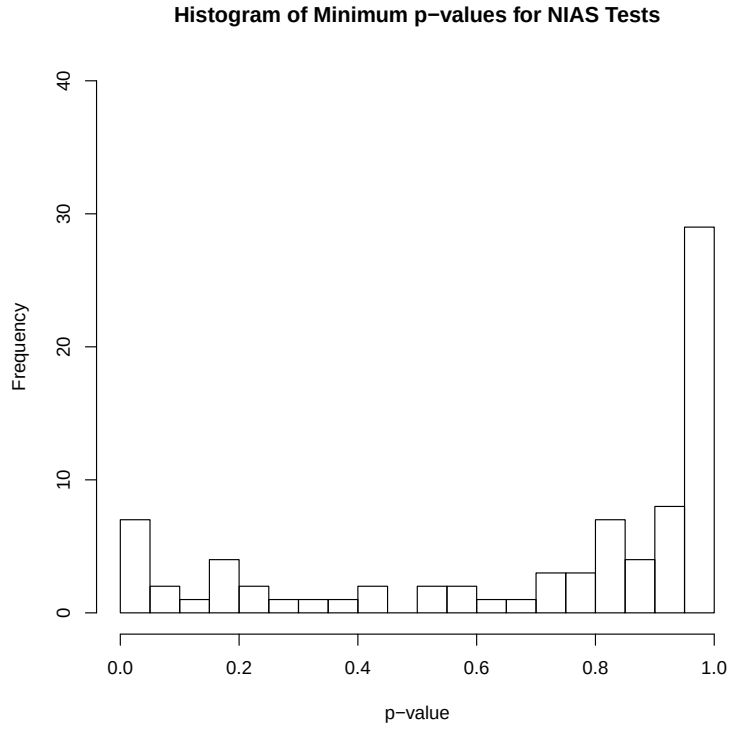


Figure S2: Histogram of minimum p-values for the NIAS bootstrap tests

of these p-values is below 0.05. In Figure S3, we present a histogram of minimum p-values for the responsiveness tests. This histogram only includes subjects that were classified as rationally attentive. Note that the distribution appears to be unimodal, spiking at below 0.05.

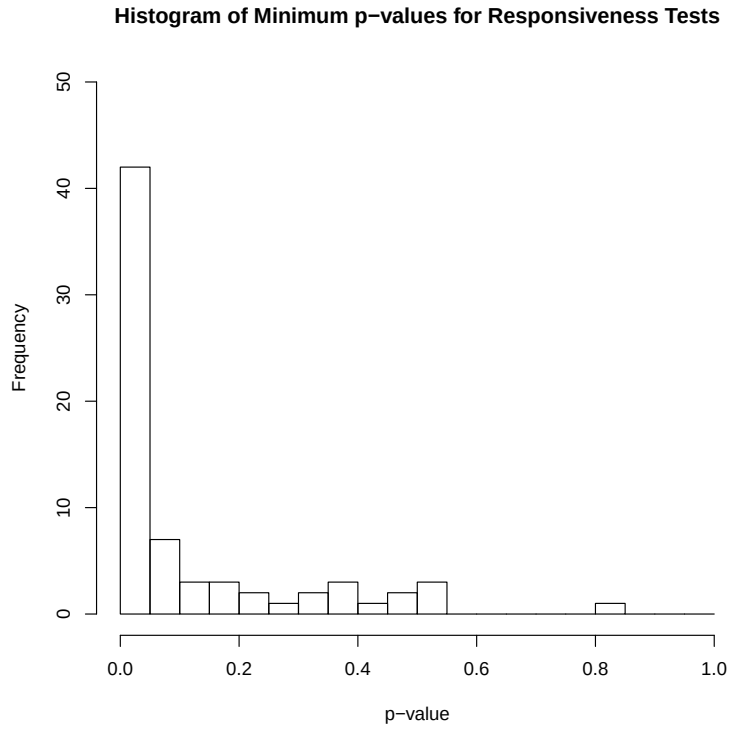


Figure S3: Histogram of minimum p-values for the responsiveness tests

To determine the extent to which demographics predict responsiveness, we run a logit regression of an indicator for responsiveness on demographic covariates for the subjects who fail to reject rational inattentiveness. We display the results of this regression in column 2 of Table S3.

As is the case with rational inattentiveness, demographic covariates are not significant predictors of responsiveness.

S2.5 Cost Functions

To determine the extent to which demographics predict model selection, we run a multinomial logit regression of the best-fitting model on the same set of demographic covariates as in previous subsections, with logistic performance (Model 7, mutual-information costs) as the baseline. This regression shows us the extent to which these demographic factors affect the likelihood of selecting

Table S3: Demographics and Categorization: Logit Regressions

	Rational Inattentiveness	Responsiveness
	(1)	(2)
Age	−0.0002	−0.015
	(0.084)	(0.064)
Male	−0.475	−0.691
	(0.718)	(0.588)
Bachelor's	0.963	0.054
	(0.804)	(0.623)
Econ/Psych/Neuro	0.749	1.261*
	(0.886)	(0.704)
$20 Prize	0.128	0.263
	(0.714)	(0.557)
Dots First	−0.602	−0.482
	(0.752)	(0.562)
Constant	1.726	0.911
	(1.939)	(1.488)
AIC	74.110	99.128
Note:		*p<0.1; **p<0.05; ***p<0.01

a model that implies a non-convexity or discontinuity in the cost function over one that is consistent with convexity. We display the results of this regression in Table S4.

As is the case with previous demographic regressions, demographic factors are not significant predictors. This seems to indicate that not only is rational inattentiveness not well captured by demographics, so is the nature of one's cost function for information in a given task.

Table S4: Model Selection and Demographics

	Binary (2)	SIC (8)	Concave (9)
Age	−0.288	0.155	−0.322
	(0.224)	(0.141)	(0.250)
Male	−0.204	−0.768	0.492
	(0.984)	(1.270)	(1.094)
Bachelor's	−0.245	−0.699	0.387
	(1.181)	(1.305)	(1.431)
Econ/Psych/Neuro	−0.056	−0.902	−0.457
	(0.968)	(1.454)	(1.136)
$20 Prize	−0.005	−0.743	−1.458
	(0.900)	(1.296)	(1.040)
Dots First	−0.153	2.230	−0.267
	(0.943)	(1.365)	(1.169)
Constant	6.076	−5.042	6.739
	(4.794)	(3.578)	(5.245)
AIC		133.323	
Note:			*p<0.1; **p<0.05; ***p<0.01

S3 Dynamic Evidence Accumulation and Reaction Times

In the main body of the paper, we focused on static models of information acquisition. However, information is typically not obtained instantaneously, but rather gathered over a period of time. In fact, dynamic models of evidence accumulation have a long tradition in mathematical psychology. In drift-diffusion models (DDMs) (e.g. Ratcliff, 1978; Diederich, 1997), evidence is modeled as a stochastic process that evolves according to a diffusion process (Smith, 2000), such as Brownian motion. The decision-maker stops gathering evidence and makes a decision when this process hits some (possibly time-dependent) boundary. This boundary is often exogenously given, as in

Ratcliff (1978), implying an exogenous information approach. However, under some conditions, the boundary can be derived as the result of an optimal stopping problem (e.g. Fudenberg et al., 2018; Tajima et al., 2016), implying an endogenous information approach. Other endogenous information approaches consider the optimal selection of the intensity of evidence accumulation when the stopping rule is exogenously given (e.g. Woodford, 2014) or the optimal selection of both evidence accumulation intensity and stopping rule (e.g. Moscarini and Smith, 2001). The vast majority of these dynamic evidence accumulation models restrict their focus to situations where the decision-maker must choose between two options, though Moscarini and Smith extend their model to consider situations with multiple discrete choice alternatives.

There is a large literature that uses choice and reaction-time data to compare models of evidence accumulation. For example, Woodford (2014) presents a model of dynamic evidence accumulation with mutual-information costs and uses Krajbich et al.'s (2010) data to compare the fit of his endogenous information model to a DDM with Brownian motion, and as mentioned in the main body of the paper, Ratcliff and Smith (2004) use data from several experiments to compare the fits of four different dynamic evidence accumulation models.

In our experiment, in addition to data on subject responses, we also collected data on how much time subjects spent on each task. We call this the *reaction time*. A full analysis of how our data fit dynamic rational inattention models is beyond the scope of this appendix and is indeed the subject of our ongoing work. Our goal in the remainder of this appendix section is to present evidence that information acquisition has salient dynamic features.

S3.1 Time and Attention

In the main body of the paper, we remained agnostic about the exact nature of what attention comprises, and by corollary, we remained agnostic about the exact source of information costs. One possibility is that attention can be decomposed into a quantity component — time spent on a task — and a quality component — how much effort is exerted during that time. Here, we provide some suggestive evidence that attention indeed has a quantity component.

Tables S5 and S6 display linear regressions of reaction time on incentive level and correctness on incentive level, respectively, aggregating over the subject pool. The coefficients on the dependent variables in both regressions are positive and significant. In the case of the first regression, this

Table S5: Linear regression of reaction time on incentive level

	Reaction Time
Incentive Level	0.178***
	(0.017)
Constant	14.159***
	(1.378)
Observations	8100
R^2	0.059

Note: *p<0.1; **p<0.05; ***p<0.01
Standard errors clustered on subject.

Table S6: Linear regression of correctness on reaction time

	Correctness
Reaction Time	0.007***
	(0.001)
Constant	0.427***
	(0.036)
Observations	8100
R^2	0.096

Note: *p<0.1; **p<0.05; ***p<0.01
Standard errors clustered on subject.

indicates that subjects respond to higher incentives by increasing the quantity of attention paid to the task at hand. In the case of the second regression, this indicates that increasing the quantity of attention results in higher performance; this is the speed-accuracy trade-off commonly noted in the literature on perceptual psychology (e.g. Schouten and Bekker, 1967).

S3.2 Dual-Process Mechanisms

As we showed in Section 6, choice data for approximately one-third of responsive subjects are best fit by binary performance functions. This suggests that these subjects employ two different strategies for determining the number of dots on the screen — one for low incentives, and one for high incentives. In this subsection, we provide further suggestive evidence for this hypothesis.

Figure S4 shows the histogram of reaction time on every task for the subject population. The distribution of reaction times is clearly bimodal. There are at least two possible, non-mutually exclusive explanations for this. One is that some portion of the subjects simply do not exert any effort on the task and make a response at the earliest opportunity, while others exert effort in acquiring information. Another is that subjects have binary performance functions, choosing not to spend time acquiring information for some incentive levels but choosing to do so for others.

The fact that a significant portion of subjects are best fit by binary performance functions provides an explanation for the pattern observed in Figure S4. Some subjects make snap decisions when confronted with low incentives but take the time to acquire information at higher incentive levels. This can be seen more clearly in Figure S5, which shows the histogram of reaction time on every task for responsive subjects only. Observe that this histogram is also clearly bimodal.

To interrogate this question further, we run the dip test of Hartigan and Hartigan (1985) on each subject's reaction times to determine which ones have multimodal reaction time distributions. We can reject the null of unimodality at the 5% level for 26 out of 42 responsive subjects (61.9%). This is more than the number of responsive subjects whose data are best fit by binary performance functions, meaning that some subjects with logistic, SIC, or concave performance functions do not have unimodal reaction time distributions. This suggests that rather than continuously adjusting their quantity of attention as incentive levels increase, some subjects randomize between paying a high quantity and a low quantity of attention, and the probability of paying a high quantity of attention increases as incentive levels increase, resulting in a sort of "fuzzy" threshold at which

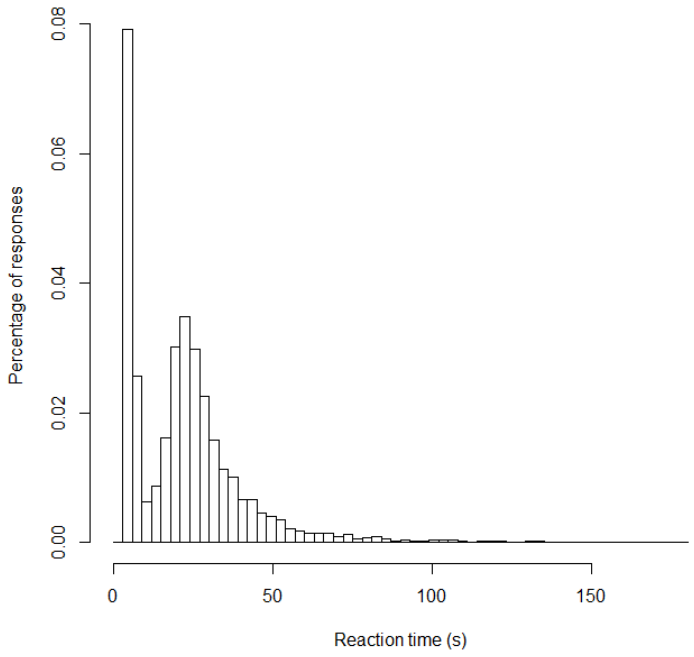


Figure S4: Histogram of reaction times for all subjects

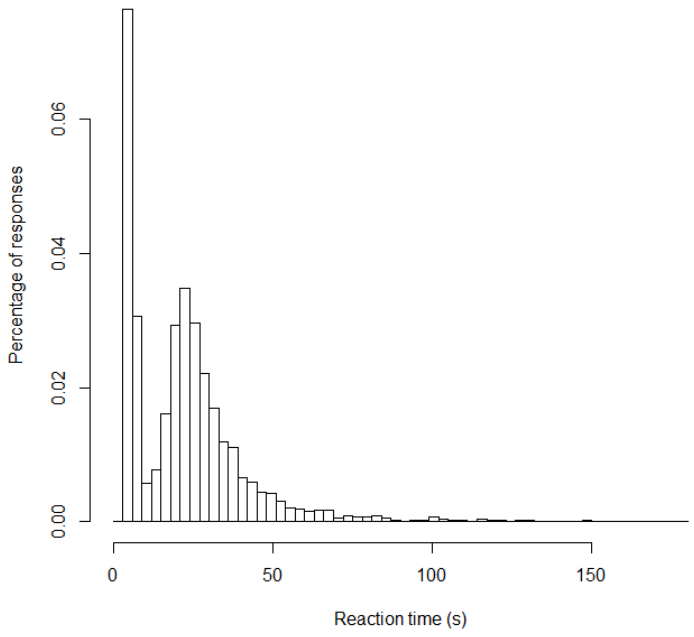


Figure S5: Histogram of reaction times for responsive subjects

they begin paying attention.

A clear analogy can be drawn between these results and concepts from psychophysics (cf. Chapter 12 of Frisby and Stone, 2010). By observing the probability of responding to or detecting a stimulus as its intensity is increased, researchers can trace out a "psychometric function." Classical psychophysics predicts that this curve is binary: the stimulus is detected with certainty above a certain threshold intensity and is undetected otherwise. Contrarily, modern psychophysics accounts for the inherent stochasticity of the human perceptual apparatus and predicts a smoothly increasing, sigmoidal psychometric curve: as the intensity of the stimulus is increased, the probability of detecting it increases continuously; there is a wide range of stimulus intensities at which the stimulus is ex-ante neither detected nor undetected with certainty. In our experiment, the incentive level is analogous to stimulus intensity, and performance is analogous to the probability of signal detection.

Since there are both subjects with binary performance and subjects with sigmoidal performance who have bimodal reaction time distributions, one possible explanation is that both types have an incentive threshold at which they begin exerting effort or paying attention. The binary types are certain about the location of this threshold, and thus, they behave according to the predictions of classical psychophysics. The logistic and SIC types with estimated $\hat{\sigma} \in (0, 2)$ also have a threshold, but they are less certain about where that threshold is, and the further away they are from that threshold, the more likely they are to behave in line with the predictions of classical psychophysics. This produces a sigmoidal performance curve.

On the whole, this evidence suggests that for a large portion of the subject pool (61.9%), there are two information-acquisition processes that they can employ in this task. Still, there is a significant portion of the pool (38.1%) that is apparently able to adjust their quantity of attention continuously. As was the case with previous categorizations of subjects, there is significant heterogeneity.

S4 Angle Task

In addition to the "dots" tasks discussed in the main body of the paper, laboratory subjects also completed 100 "angle" tasks. For each of these tasks, subjects were shown a pair of intersecting line

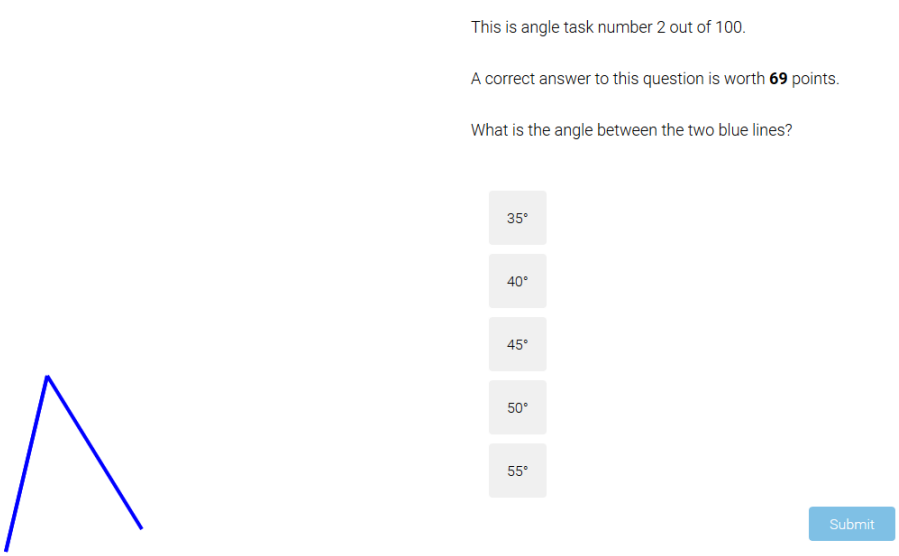


Figure S6: Angle display for a task

segments of random length[S5] and orientation and were told to identify the angle between them. This angle could have been 35°, 40°, 45°, 50°, or 55°, with each being equally likely. Subjects were rewarded for a correct answer and received no reward for an incorrect answer. Therefore, the "angle" tasks were uniform guess tasks of the same format as the "dots" tasks. Figure S6 shows what this screen looked like to the subjects.

Table S7 presents linear regressions of correctness on incentive level and demographic covariates for the entire laboratory subject pool. As was the case with the "dots" task, demographics are not significant predictors of correctness. However, neither is incentive level. This evidence indicates that this is not a task in which subjects generally respond to incentives.

S5 Online Experiment

In this appendix, we describe and present results from the online experiments mentioned in the main body of the paper.

Subjects were recruited using the Amazon Mechanical Turk (MTurk) platform[S6] and partic-

[S5] Giving the arms of the angle random length ensured that subjects could not simply measure the distance between the endpoints of the arms to estimate the size of the angle.

[S6] In recent years, many experiments and surveys have been conducted on MTurk. Research has shown that results from MTurk samples are similar to convenience samples typically used by researchers (e.g. student samples) and are more representative of the U.S. population, though they also differ markedly in some psychological and political

Table S7: Linear regression of correctness on incentive level and demographic covariates in the "angle" tasks

	(1)	(2)
Incentive Level	0.0001	0.0001
	(0.0002)	(0.0002)
Age		−0.001
		(0.002)
Male		−0.007
		(0.014)
Bachelor's		−0.001
		(0.017)
Econ/Psych/Neuro		0.028*
		(0.017)
$20 Prize		0.017
		(0.013)
Dots First		−0.015
		(0.014)
Task Number		−0.00001
		(0.0002)
Constant	0.444***	0.461***
	(0.014)	(0.036)
Observations	7900	7900
R^2	1.16×10^{-5}	0.0009818

Note: *p<0.1; **p<0.05; ***p<0.01
Standard errors clustered on subject.

Table S8: Online Demographics

Number of subjects	$n = 118$
Gender ($n = 117$)	52.5% male; 47.5% female
Age ($n = 118$)	Average: 32.48; St. dev.: 8.88
Highest level of education achieved ($n = 118$)	
Some post-secondary	43.2%
Completed bachelor's degree	50.0%
Completed graduate or professional degree	6.8%

ipated in the experiment on the Qualtrics platform. A total of 118 subjects completed the experiment. Subjects completed 200 tasks, each of the "dots" type. Roughly half the subjects (57 subjects) were given a participation fee of $3 US and potential monetary prizes of $3, while the other half (61 subjects) were given a participation fee of $5 US and potential monetary prizes of $5 US.

S5.1 Demographics

Table S8 lists basic demographic data for the online subjects. The pool is fairly gender-balanced,[S7] though it is slightly more male than female, and highly educated; over 55% of the pool has a post-secondary degree.

The online pool is signficantly different from the laboratory pool in some ways. In particular, the online pool is significantly older (one-tailed t-test of unpaired samples, $p < 0.001$) and has a significantly greater proportion of subjects with bachelor's degrees but no advanced degrees (one-sided test of equality of proportions, $p = 0.003$).

S5.2 Rational Inattentiveness

S5.2.1 No Improving Attention Cycles

We test against weak positive monotonicity using the method of Doveh et al. (2002). At the 5% level, we fail to reject positive monotonicity for 103 out of 118 online subjects (87.3%).[S8]

characteristics. See, for example, Berinsky et al. (2012) and Goodman et al. (2013).

[S7] One online subject declined to disclose their gender.

[S8] The optimization in the computation of the restricted regression for online subject 93 failed to converge, and so we did not perform the test for them. That subject has a success rate in the tasks of 99% (i.e. they identify the true state of nature correctly in 198 out of 200 tasks), and so we include them in the 103 online subjects who fail to reject positive monotonicity.

S5.2.2 No Improving Action Switches

We test for NIAS using the bootstrap procedure outlined in Section 5. 82 out of 118 online subjects (69.5%) fail to reject NIAS.

Overall, this gives us 72 out of 118 online subjects (61.0%) whom we classify as rationally inattentive. This is a significantly smaller portion than in the laboratory pool (one-sided test of proportions, $p < 0.001$).

S5.3 Responsiveness to Incentives

We test for responsiveness using the full-sample and split-sample tests outlined in Section 5. At the 5% significance level 28 out of 72 online subjects (38.8%) who fail to reject rationality are responsive to incentives. This is a significantly smaller portion than in the laboratory pool (one-sided test of proportions, $p = 0.009$).

S5.4 Model Selection

We follow the same model selection procedures as in Section 5. As with the laboratory subjects, the only models that best fit the subjects are binary response and logistic response. 2 out of 28 responsive subjects (7.1%) are best fit by constant performance, 8 out of 28 responsive subjects (28.6%) are best fit by binary performance, 17 out of 28 responsive subjects (60.7%) are best fit by logistic performance, and 1 out of 28 responsive subjects (3.6%) are best fit by the concave performance function implied by normal signals. No subjects are better fit by the SIC generalization of logistic performance than by logistic performance itself. Ignoring the subjects who are best fit by constant response, and collapsing logistic and SIC performance into a single category, these are similar to the proportions found in the laboratory. This seems to indicate that once the subset of responsive subjects is identified, the incidence of different types of cost functions within it is stable across demographic contexts.

S6 Application to the Delegation of Investment

The characteristics of the decision-maker's cost function can obviously have effects on her own decisions. But as we show in this appendix section, these characteristics can also have effects on

economically-relevant outcomes when there is strategic interaction.

In order to demonstrate this notion, let us consider a situation in which an investor is deciding which of n options to invest in, and he cannot split his investment across options. Suppose that only one of these options can be a winner, in which case an investment in it will pay a net return of x. Losing opportunities pay a net return of zero. This setup has the relevant features of a situation where the success of an investment depends on the outcome of a contest. Many economic situations, such as competing to be granted development rights by the government for a plot of land, take the form of contests. Another salient example is a patent race, where various firms compete to be the first to patent an invention, such as a drug or a piece of technology.

Suppose that the investor wishes to delegate researching these options to an expert. This is a common occurrence in reality; people frequently solicit the services of financial advisors, presumably because it is prohibitively difficult or costly for laypeople to research investment opportunities themselves, while financial advisors who are trained to seek and interpret financial information can research these opportunities at a much lower cost.

We can analyze this situation in a simple principal-agent framework, where the investor is the principal and the expert is the agent.[S9] The agent acquires information about the available investment opportunities at a cost and selects one of the options on the principal's behalf. Suppose that the principal employs the agent with a contract that pays r if the agent correctly selects the winner and zero otherwise.[S10] Furthermore, suppose that *a priori*, each option is equally likely to be the winner. Then, the agent's problem can be represented as a uniform guess task, with the reward for a correct answer being r. Consequently, the principal's problem is

$$\max_{r\in[0,x]} (x-r)P^*(r) \tag{S1}$$

where $P^*(r)$ is the agent's performance function.

As we established in Proposition 1, if the agent is rationally inattentive, then her performance function is (weakly) increasing. Thus, the principal faces a trade-off between incentivizing the agent

[S9] We use male pronouns for the principal and female pronouns for the agent.

[S10] This type of contract is optimal for the principal if we assume that (a) there is a limited-liability constraint so that the agent cannot earn a negative payoff in any state of the world, which implies that the principal cannot "sell the firm" to the agent; and (b) the agent's cost of an uninformative information structure is zero. As Caplin and Dean (2015) demonstrate, the latter assumption is without loss of generality; it is not a testable restriction on information cost functions.

to acquire better information and giving up a larger portion of his net return upon success. The exact nature of this trade-off depends on the potential net return x and the agent's information cost function. In the following subsections, we analyze the properties of the principal's optimal payment strategy r^* under three of the cost function models fit by our data: [S11] fixed costs; mutual information; and normally-distributed signals.[S12]

S6.1 Fixed Costs

Suppose the agent has a fixed cost κ for acquiring information. If she pays the cost, then she learns the winner with certainty. If not, then she learns nothing about the identity of the winner. Thus, she chooses to acquire information if $r - \kappa \geq \frac{r}{n}$, i.e. when $r \geq \frac{\kappa n}{n-1}$.

Therefore, if $x < \frac{\kappa n}{n-1}$, then the reward required to incentivize the agent to acquire information is higher than the potential net return, so the principal is better off not hiring the agent at all and simply picking an option at random. If instead $x \geq \frac{\kappa n}{n-1}$, then the principal could incentivize information acquisition by paying as little as $r = \frac{\kappa n}{n-1}$. To ensure that this payment is not so high than the principal could do better on his own, he requires that $\frac{x}{n} \leq x - \frac{\kappa n}{n-1}$, which holds if and only if $x \geq \frac{\kappa n^2}{(n-1)^2}$. But since $\frac{\kappa n}{n-1} < \frac{\kappa n^2}{(n-1)^2}$, the principal will not hire the agent unless $x \geq \frac{\kappa n^2}{(n-1)^2}$.

To summarize: if $x < \frac{\kappa n^2}{(n-1)^2}$, then the principal does not hire the agent and selects an option at random. If $x \geq \frac{\kappa n^2}{(n-1)^2}$, then the principal hires the agent and gives her a payment of $\frac{\kappa n}{n-1}$, and the agent picks the winner with certainty. This implies a discontinuity in the principal's payment as a function of the potential net return x. Figure S7 shows what this payment scheme looks like for $\kappa = 40$.

[S11] We exclude Tsallis entropy costs from our analysis since the corresponding performance function does not in general have a closed form (see Subsection 3.2 of the main paper), and its properties vary substantially with the σ parameter. However, numerical simulations seem to indicate that the principal's profit function is strictly quasiconcave in the reward r paid to the agent (see the proof of Proposition S2 in Subsection S6.2 for why this is important), and it can be confirmed that this is the case when $\sigma = 2$ (i.e. when costs are quadratic).

[S12] Some caution is required in applying the assumption of normally-distributed signals, because it implies that the options have some existing ranking, and it is not clear what it means for the options to be "equidistant" from each other. In any case, if the normal-signals and Tsallis models are excluded from consideration, then in our data, the best-fitting model for each subject is either binary (fixed costs) or logistic (mutual information). (Results available from the authors on request.)

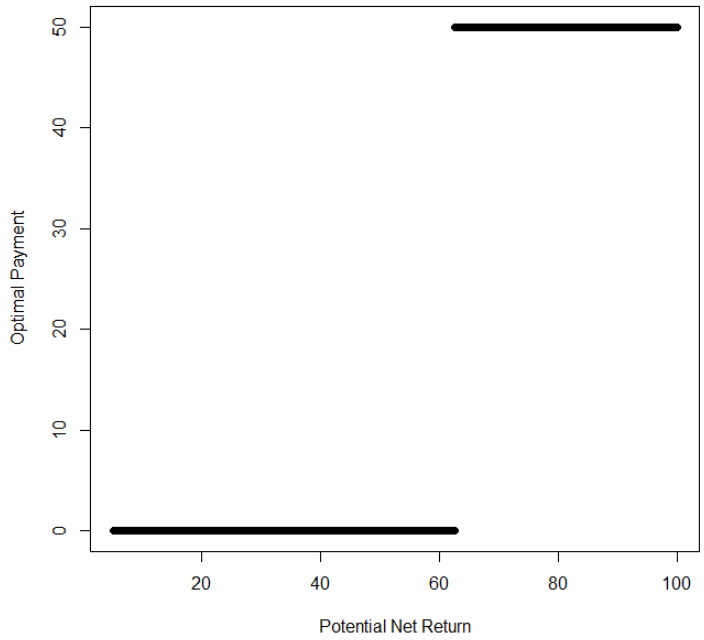


Figure S7: Optimal payment as a function of potential net return, fixed costs

S6.2 Mutual Information

Suppose the agent has a mutual-information cost function with cost parameter α. Then, since her performance function is logistic (see Proposition 5), the principal chooses r to maximize:

$$\frac{x - r}{(n-1)\exp\left(-\frac{r}{\alpha}\right) + 1} \tag{S2}$$

If this maximand is strictly quasiconcave, then this problem has a unique solution for each x, and the maximum theorem guarantees that the principal's optimal choice of r^* is continuous in x. This turns out to be the case.

Proposition S2. *If the agent has a mutual information cost function, then the principal's optimal payment strategy $r^*(x)$ is continuous.*

Proof. The principal's maximand is:

$$\frac{x - r}{(n-1)\exp\left(-\frac{r}{\alpha}\right) + 1} \tag{S3}$$

As argued above, if this maximand is strictly quasiconcave in r, then this problem has a unique solution for each x, and since it is continuous in both x and r, the maximum theorem guarantees that the principal's optimal payment strategy $r^*(x)$ is continuous. Therefore, it simply remains to be shown that the maximand is strictly quasiconcave. We begin by differentiating it with respect to r:

$$\frac{\left(\frac{x-r-\alpha}{\alpha}\right)(n-1)\exp\left(-\frac{r}{\alpha}\right)-1}{\left((n-1)\exp\left(-\frac{r}{\alpha}\right)+1\right)^2} \tag{S4}$$

Since the denominator in (S4) is always strictly positive, the sign of (S4) depends only on the sign of the numerator. The numerator is strictly positive (negative) when:

$$\begin{aligned}&\left(\frac{x-r-\alpha}{\alpha}\right)(n-1)\exp\left(-\frac{r}{\alpha}\right) > (<) 1\\ \iff &(n-1)\left(\frac{x-r-\alpha}{\alpha}\right) > (<)\exp\left(\frac{r}{\alpha}\right)\end{aligned} \tag{S5}$$

The LHS of (S5) is strictly decreasing, and diverges to positive infinity as r is taken to negative infinity and to negative infinity as r is taken to positive infinity. The RHS of (S5) is strictly increasing, and it approaches zero as r is taken to negative infinity and diverges to positive infinity as r is taken to positive infinity. Therefore, by the intermediate value theorem, the LHS and RHS must intersect, and they do so only at a single r.

Therefore, (S3) exhibits a region of strict increase up until the point where $(n-1)\left(\frac{x-r-\alpha}{\alpha}\right) = \exp\left(\frac{r}{\alpha}\right)$, after which it is strictly decreasing. Thus, (S3) is strictly quasiconcave. □

To provide an example, suppose $n = 5$, $\alpha = 10$, and $x \in [5, 100]$. (A graph of the principal's maximand (S2) is shown in Figure S8.) For these parameters, $r^*(x)$ is continuous and increasing, as shown in Figure S9.

S6.3 Normally-Distributed Signals

Suppose that the options are ranked and equidistant on some scale. For example, in the case of bidding for development rights, the projects could be ranked by the estimated length of time

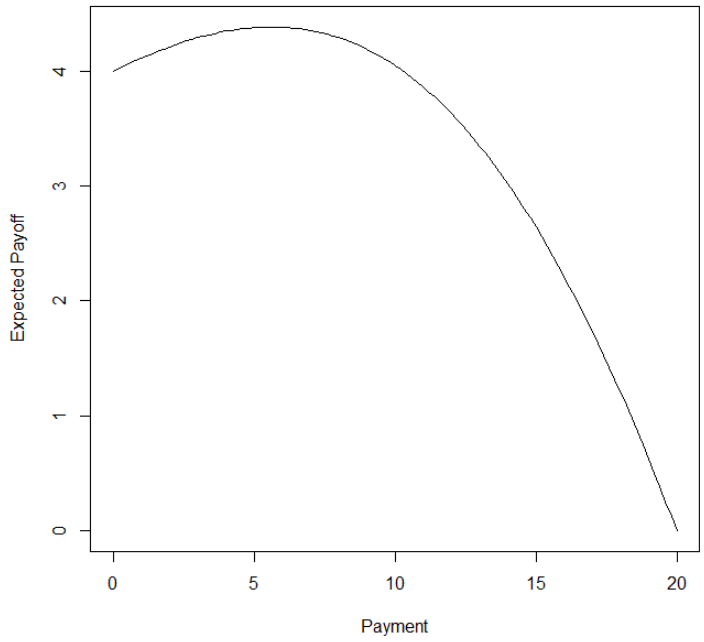


Figure S8: Principal's expected payoff as a function of payment for $x = 20$, mutual information costs

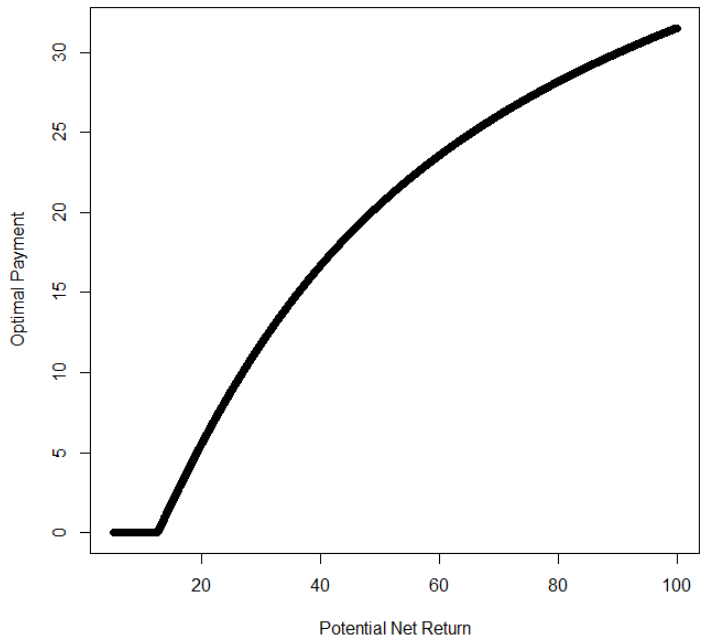


Figure S9: Optimal payment as a function of potential net return, mutual information costs

until project completion.[S13] In this case, if the agent's cost function satisfies the conditions of Proposition 6, then it can be shown that the principal's optimal choice of r^* is continuous in x.

Proposition S3. *If the options are ranked and equidistant, and the agent has a convex, increasing cost of precision of normal signals with non-negative third derivative, then the principal's optimal payment strategy $r^*(x)$ is continuous.*

Proof.

$$\begin{aligned}
&\frac{d^2}{dr^2}[(x-r)P^*(r)] \\
=&\frac{d}{dr}[-P(r)+(x-r)\frac{d}{dr}P^*(r)] \\
=&-2\frac{d}{dr}P^*(r)+(x-r)\frac{d^2}{dr^2}P^*(r) \qquad \text{(S6)}
\end{aligned}$$

(S6) is negative, since $P^*(r)$ is strictly increasing and strictly concave, and $x > r$, so the principal's ex-ante expected payoff is strictly concave in r. Therefore, there is a unique r^* for each x, and by the maximum theorem, $r^*(x)$ is continuous. □

Figure S10 shows what this payment scheme looks like if costs are linear in the precision of normally-distributed signals, with a marginal cost of precision of 7.5.

S6.4 Welfare and Robustness

The properties of an agent's information cost function also have implications for the robustness of the model's predictions, particularly for the principal's welfare. If the principal is slightly — even infinitesimally — misinformed about the parameters of an agent's cost function, then this can have major impacts on his welfare if the agent's cost function is discontinuous.

Consider an agent with a fixed-cost information cost function, with cost parameter κ. Suppose that the principal believes that the agent's cost parameter is $\kappa' := \kappa - \varepsilon$, where $\varepsilon \in (0, \kappa)$. If the principal had a correct assessment of the agent's information costs, then he would pay her $\frac{\kappa n}{n-1}$ for a success, causing the agent to acquire information, and earning $x - \frac{\kappa n}{n-1}$ in expectation. However, since he misperceives her fixed cost for information acquisition as κ', he instead offers $\frac{(\kappa-\varepsilon)n}{n-1}$. This

[S13] Shorter completion times mean that the development will be more quickly available for public use, but may also signal poor craftsmanship.

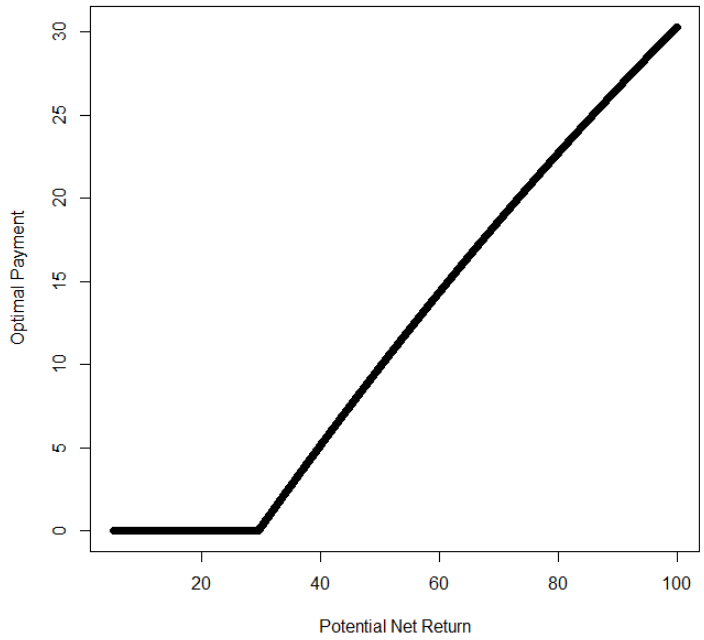


Figure S10: Optimal payment as a function of potential net return, normally-distributed signals

is not enough to incentivize the agent to acquire information, resulting in expected earnings of $\frac{x}{n} - \frac{\kappa-\varepsilon}{n-1}$. The expected welfare loss to the principal from this mistake is therefore $\frac{(n-1)x}{n} - \kappa + \frac{\varepsilon}{n-1}$; an arbitrarily small error produces a welfare loss on the order of $\frac{(n-1)x}{n} - \kappa$, which can be very large if x is very large.

By contrast, this does not occur if the agent has a cost function that generates a continuous performance function. In that case, the continuity of the performance function $P^*(r)$ implies that the principal's welfare $(x-r)P^*(r)$ is also continuous in r; this is because an agent with continuous performance does not drastically adjust her behavior in response to small changes in incentives. Therefore, by continuity, small mistakes on the principal's part in assessing the agent's cost function parameters only produce small welfare losses.

Thus we have shown that the model's welfare predictions are not robust to small (downward) perturbations in the principal's assessment of the agent's costs when the agent has a discontinuous cost function. Practically speaking, this means principals must exercise extra caution in a world where agents have fixed-cost information cost functions, perhaps by intentionally overpaying agents or by carefully studying them before hiring. Even a small error in designing the payment scheme

could produce catastrophic welfare losses; a near-optimal contract does not necessarily produce a near-optimal outcome for the principal. This problem does not present itself when the agent has a cost function that generates continuous performance.

References

Adam J. Berinsky, Gregory A. Huber, and Gabriel S. Lenz. Evaluating online labor markets for experimental research: Amazon.com's Mechanical Turk. *Political Analysis*, 20(3):351–368, 2012.

Andrew Caplin and Mark Dean. Revealed preference, rational inattention, and costly information acquisition. *American Economic Review*, 105(7):2183–2203, 2015.

Adele Diederich. Dynamic stochastic models for decision making under time constraints. *Journal of Mathematical Psychology*, 41(3):260–274, 1997.

E. Doveh, A. Shapiro, and P.D. Feigin. Testing of monotonicity in parametric regression models. *Journal of Statistical Planning and Inference*, 107(1–2):289–306, 2002.

John P. Frisby and James V. Stone. *Seeing: The Computational Approach to Biological Vision*. MIT Press, 2010.

Drew Fudenberg, Philipp Strack, and Tomasz Strzalecki. Speed, accuracy, and the optimal timing of choices. *American Economic Review*, 108(12):3651–84, 2018.

Joseph K. Goodman, Cynthia E. Cryder, and Amar Cheema. Data collection in a flat world: The strengths and weaknesses of Mechanical Turk samples. *Journal of Behavioral Decision Making*, 26(3):213–224, 2013.

J.A. Hartigan and P.M. Hartigan. The dip test of unimodality. *Annals of Statistics*, 13(1):70–84, 1985.

Ian Krajbich, Carrie Armel, and Antonio Rangel. Visual fixations and the computation and comparison of value in simple choice. *Nature Neuroscience*, 13(10):1292–1298, 2010.

Giuseppe Moscarini and Lones Smith. The optimal level of experimentation. *Econometrica*, 69(6): 1629–1644, 2001.

Roger Ratcliff. A theory of memory retrieval. *Psychological Review*, 85(2):59–108, 1978.

Roger Ratcliff and Philip L. Smith. A comparison of sequential sampling models for two-choice reaction time. *Psychological Review*, 111(2):333–367, 2004.

J.F. Schouten and J.A.M. Bekker. Reaction time and accuracy. *Acta Psychologica*, 27:143–153, 1967.

Philip L. Smith. Stochastic dynamic models of response time and accuracy: A foundational primer. *Journal of Mathematical Psychology*, 44(3):408–463, 2000.

Satohiro Tajima, Jan Drugowitsch, and Alexandre Pouget. Optimal policy for value-based decision-making. *Nature Communications*, 7:1–12, 2016.

Michael Woodford. Stochastic choice: An optimizing neuroeconomic model. *American Economic Review*, 104(5):495–500, 2014.