Attentional Complements

Dániel Csaba

April 19, 2019

Abstract

I study how attentional constraints affect elasticity and substitution patterns in demand. I identify "attentional complementarity", whereby goods that are substitutes in utility appear as complements in behavior due to limits on attention. Adopting the framework of rational inattention, I identify the channels through which attention influences observed substitution patterns. The in-sample fit of these models can be similar to that of standard discrete choice models, yet counterfactual predictions differ substantially when attentional frictions are accounted for. I provide conditions for identification of the model in standard stochastic choice data. A key obstacle in estimation is the genericity of corner solutions which endogenously gives rise to consideration sets. The estimation strategy I employ is robust to this feature. In an empirical application, I show that elasticities are underestimated by models not accounting for attentional frictions. I conduct an experiment allowing for the detailed analysis of attention strategies and find that subjects allocate their attention in line with the theory.

JEL classification: C51, D83, D90

I am grateful to Andrew Caplin for his advising and support throughout this project and I thank Timothy Christensen and Andrew Schotter for continued discussions and suggestions. Furthermore, I benefited greatly from conversations with Miguel Angel Ballester, Tommaso Bondi, Sylvain Chassang, Guillaume Fréchette, Evan K. Friedman, Alfred Galichon, Laurent Mathevet, Jacopo Perego, Daniel Stackman, and Bálint Szőke, as well as participants of seminars at NYU, and at the Young Economists Symposium (NYU) and the Student Workshop in Experimental Economics Techniques (Columbia) conference. I thank Oded Nov and Theodore Kim for software assistance.

Economics Department, New York University. Email: daniel.csaba@nyu.edu.

1 Introduction

Acquiring information is often an integral part of choosing among alternatives with uncertain prospects. Whether taking the bus or the train is the better choice might depend on the traffic on the roads; the health plan that best suits a patient depends on innate traits and health status. In these examples the state is uncertain but can be learned about. However, there is substantial evidence that even when information is readily available, people fail to fully incorporate it into their decisions (see the thorough surveys of DellaVigna [2009] and Handel and Schwartzstein [2018] for a range of economic examples). Constraints on attention are particularly relevant in environments where agents have to grapple with informationally complex decisions. Given that attention is a scarce resource, the features of the environment people deem worthy to pay attention to are shaped relative to the choice set. It is natural then to ask: what are the implications of endogenous attention allocation on behavior?

In this paper I show that substitution patterns in demand can change considerably when attention is costly. In particular, the endogenous allocation of scarce attention can result in goods that are substitutes in utility to appear as complements in observed behavior. This extends a wide range of recently analyzed behavioral phenomena that can arise due to limits on attention—ranging from reference dependence [Woodford, 2012] to consideration sets [Caplin et al., 2018b]. Analyzing observed choices without taking into account the attentional limitations can lead to false inference affecting far-reaching policy and corporate decisions.

To analyze these problems, I adopt the framework of rational inattention (RI) [Sims, 2003] allowing for a general class of information cost functions. I provide conditions for identification of the general RI model in standard stochastic choice data. This renders counterfactual predictions feasible, and allows for predicting rich behavioral phenomena that is otherwise inconsistent with standard discrete choice models. A distinctive feature of endogenous attention allocation is that some alternatives in the choice set have zero probability of being selected—that is, attentional frictions can give rise to consideration sets. The presence of these corner solutions renders standard out of the box optimization techniques unsuitable for estimating parameters of the RI model. My estimation strategy is robust to this problem. I empirically demonstrate the developed methodology and experimentally test predictions of the model.

More specifically, attentional complementarity can arise through the endogenous allocation of attention. The improvement in the quality of one alternative, through changing what people pay attention to, can make another alternative more likely to be chosen. Continuing with the earlier example, the choice among taking the bus, the train, or a flight can depend on the traffic on the roads and whether or not there is a cheap flight. If a new exclusive bus lane allows buses to avoid the usual traffic, then checking traffic updates is no longer valuable in deciding between the bus and the train. The improvement in taking the bus—namely, that with the exclusive lane buses arrive on time even in case of traffic—frees up attention that can be paid elsewhere. The time spent on checking traffic updates can be spent on checking prices for flights. This will increase the chance of discovering a cheap flight and thus lead to taking the plane more often.

The mechanism through which attention gives rise to complementarity is unrelated to other sources of complementarity in a discrete choice context. For example, correlation in tastes could give rise to complementarity on the aggregate but not on the individual level. Correlated signals, on the other hand, link alternatives that share a common feature and change their expected payoffs jointly. Attentional complements, on the other hand, are driven by the change in the type of information deemed useful and hence acquired, which in turn shapes choice patterns.

I analyze comparative statics of the resulting individual demand functions with respect to changes in the payoff structure. In the spirit of the Slutsky decomposition of classical Marshallian demand, I decompose the changes in individual demand highlighting the channels through which attention affects choice. Directly, an increase in the payoff of an alternative makes that alternative more attractive and more likely to be chosen relative to others—as the introduction of the bus lane makes taking the bus a more attractive option. Indirectly, it changes the incentives shaping the allocation of attention. Attention can be reallocated to different sources—like checking flights instead of traffic—and the overall invested attention can also change. The latter effect increases the choice probabilities of some alternatives while decreasing them for others—as probabilities have to add up to one.

A key obstacle in the wider use of the RI model in empirical work is its requirement for rich datasets. I provide conditions for identification of the model in standard observational choice data. Since attentional problems depend on all possible states of the world, successful inference requires assumptions that ensure that the observed choices are informative about all of them. To obtain identification, I assume that the econometrician and the consumers possess the same initial information. In particular, they are both aware of the same set of demographic and alternative specific covariates. People then acquire further information about their latent valuation types that determine the match quality with available alternatives. Demographic covariates affect the assignment of valuation types. Given a valuation type, an alternative's attributes determine the associated hedonic utility. Smoothness conditions on functional forms and sufficient conditional variation in the covariates, enables identification of the utility and prior parameters of the RI model.

Estimation of the model poses its own challenges. A key feature of the rational inattention model is that the solution to a problem could imply no chance of selection for certain alternatives. In other words, the model endogenously gives rise to "consideration sets" [Caplin et al., 2018b]. The estimation strategy I employ handles these naturally emerging corner solutions in two respects. First, I use the conditional moment conditions implied by the model instead of the likelihood function in order to avoid issues with chosen alternatives having zero predicted probabilities. Second, since at corner solutions the gradient of the criterion function does not provide useful information for parameter search, the parameters are estimated via sampling from the quasi-Bayesian posterior of the criterion function as in Chernozhukov and Hong [2003] and Chen, Christensen, and Tamer [2018]. Under a flat prior this procedure yields equivalent estimates to directly maximizing the GMM criterion function without relying on information from the gradient for parameter search. I demonstrate my approach on a workhorse data set of travel mode choice.

My estimation strategy and the conditions for identification apply to a wide class of RI models that allow for rich behavioral phenomena. Inference from observational data such that meaningful counterfactual predictions can be carried out is of primary importance, given that many of these patterns can not arise in the standard rational choice model.

I experimentally test the model's predictions on choice probabilities related both to changing costs of attention and changing relative payoffs. The experiment is conducted on Amazon Mechanical Turk involving risky choices. Subjects are provided with an interface through which they can voluntarily reduce the underlying uncertainty. Subjects at the on-line platform have a clear opportunity cost of spending time and effort on the information acquisition task as they could earn money by completing other tasks on MTurk. For the task to reduce uncertainty, I use a geometric counting task first introduced in Caplin, Csaba, Leahy, and Nov [2018a].

I test conditions for the observed behavior to be consistent with the existence of a rationalizing cost function and find that tests are significantly passed once accounting for heterogeneity in subjects. I find that subjects respond to attentional incentives as they adjust their attention strategies based on both the relative cost of information and changes in the payoff structure.

The remainder of the paper is organized as follows. Section 2 discusses related literature. Section 3 introduces the general RI model and maps it to stochastic choice functions highlighting the contrast with random utility models. Section 4 defines attentional complements and analyzes the main forces driving substitution patterns under constrained attention. Section 5 presents a simple example demonstrating complementarity in the context of the standard RI model. Section 6 discusses estimation and inference for the RI model from observational data. Section 7 presents the experiment demonstrating patterns in stochastic choice arising due to information acquisition. Section 8 applies the model and showcases the proposed estimation strategy on a dataset of travel mode choice. Section 9 concludes.

2 Related Literature

The role of endogenous attention and information acquisition have received increasing economic interest in the past decade. Starting with the seminal work of Sims [2003], there have been several influential studies analyzing behavioral patterns that are inconsistent with the standard rational choice model and could arise due to attentional frictions. A detailed review of the burgeoning RI literature can be found in Maćkowiak et al. [2018].

Recent papers that rationalize widely observed behavioral phenomena in the context of RI include: Woodford [2012], who shows that endogenous attention allocation can give rise to decoy effects and reference dependence; Matějka and McKay [2015], who show that the standard rational inattention model does not satisfy independence of irrelevant alternatives and moreover, can violate monotonicity in augmenting the choice set; Caplin et al. [2018b], who show how attentional frictions can give rise to consideration sets; and Steiner et al. [2017], who show that inattention results in inertia in dynamic contexts. I complement these papers by showing that the RI model can considerably alter substitution patterns in demand, changing goods that are substitutes in utility to complements in observed behavior.

I present the RI model in its general framework allowing for a wide variety of information cost functions that are separable in the induced posteriors. Flexible cost functions in the RI framework have been a focus of recent theoretical studies. Caplin et al. [2017] and Denti [2018] give revealed preference characterizations of the posterior separable class, while Hébert and Woodford [2017] show that all posterior separable cost functions can be derived from a sequential information sampling problem. Importantly, posterior separable cost functions can accommodate neighborhood based costs, these allow nearby states to be more difficult to distinguish than ones that are far apart [Hébert and Woodford, 2017, Morris and Yang, 2016]. Both my theoretical and empirical analyses are applicable to all posterior separable cost functions.

The relationship of RI models to standard discrete choice models was first pointed out by Matějka and McKay [2015] relating the RI model with Shannon entropy cost to standard multinomial logit models. Fosgerau et al. [2017] generalize these results to a broader class of entropy functions and relate them to additive random utility models. However, these equivalence results only hold for a fixed choice set: comparative statics and counterfactual predictions are starkly different under the RI and the random utility models. This has important consequences in many economic applications, ranging from predicting

responses to policy changes or changes in product development. In the second half of the paper I discuss inference for the RI model from observational data such that meaningful counterfactual predictions can be carried out.

In the handful of studies discussing the inference problem for the RI model in observational data, the uncertainty the decision makers acquire information about is their type describing the match quality between consumer and available alternatives. Caplin et al. [2015] take a steady state approach, where agents can infer the distribution of types from the market shares. Joo [2018] also assumes that the agents learn about their types and uses separate prior and utility shifters to estimate parameters. His approach to estimation is an indirect one, estimating unconditional choice probabilities and then constructing a distribution over types such that the corresponding choice probabilities can be rationalized by a standard RI model. Importantly, the validity of this approach depends on the unconditional choice probabilities being interior for each individual, that is, the formation of consideration sets is not allowed. My approach, on the other hand, estimates the RI model directly, tailoring the procedure explicitly to the possibility of consideration set formation. A different, reduced form approach to estimating consideration sets is put forward in Abaluck and Adams [2018] who identify consideration sets using the implied asymmetries in the Slutsky matrix.

Parallel to the theoretical work, there is a growing experimental literature on attention. Starting with Cheremukhin et al. [2011] there have been a number of papers [Caplin and Dean, 2015, Dewan and Neligh, 2017, Dean and Neligh, 2017] analyzing and testing detailed implications of the RI model. The overall finding is that subjects behave in line with the general theory of RI in situations where information can be acquired at a cost. However, behavioral patterns implied specifically by the Shannon entropy cost do not always hold. In particular, the assumption of symmetry across states is often rejected in the data. In other work, Ambuehl et al. [2018] analyze selection effects in the context of transactions with uncertain consequences. They show that subjects with higher informational costs respond more strongly to explicit incentives—that are independent of the uncertainty—to participate in the transaction. My experimental test of patterns in demand when costly information is available complements these studies and confirms the basic predictions of the RI model.

Besides the RI framework there have been several other successful approaches explaining well-documented behavioral patterns related to how people—voluntarily or involuntarily—allocate their attention. Some of the most influential ones are the model of focusing by Kőszegi and Szeidl [2012], the salience model of Bordalo et al. [2012, 2013], and the sparse-max model of Gabaix [2014].

Finally, recent studies have analyzed situations where complementarities are likely to arise. Gentzkow [2007] develops a model that allows for both substitutability and complementarity stemming from preferences and applies it to media consumption. His primary focus is on situations where consumers can select multiple goods and substitution patterns are driven by individual preferences and not attentional constraints. Allen and Rehbeck [2017] analyze complementarities in perturbed utility models without making explicit mention of the role of attention. Manzini et al. [2018] develop a model free approach to analyze complementarities in stochastic choice solely based on the co-movements of choice probabilities. Berry et al. [2014] provide an overview of other structural models that can account for complementary choice for other reasons—quantity, quality or dynamic complements among others. In this respect, the present paper provides a structural model of stochastic choice that can exhibit complementarities at the individual level due to limits on attention as opposed to complementarities arising due to pure complementary preferences, correlation in tastes at the population level or correlated signals affecting the utility

of several alternatives.

3 The RI Framework

3.1 The Decision Problem

Consider a standard decision problem under uncertainty. The environment is described by a finite set of states with generic element $\omega \in \Omega$. The set of outcomes is denoted by $\mathcal{Z}$ and actions induce state-contingent outcomes, $a \in \mathcal{Z}^{\Omega}$. In a decision problem the decision maker (DM) faces nonempty choice sets, $\mathcal{A} := \{ A \subset \mathcal{Z}^{\Omega} : 0 \neq |A| < \infty \}$. The DM's preferences are described by a utility function over outcomes, $u \colon \mathcal{Z} \mapsto \mathbb{R}$. When convenient denote the composition of utility function and action as $u_a := u \circ a \colon \Omega \mapsto \mathbb{R}$. The set of beliefs over the states is denoted by $\Gamma := \Delta(\Omega)$.

The DM is a standard Bayesian expected utility maximizer. Given a belief $\gamma \in \Gamma$ and a decision problem $A \in \mathcal{A}$ the best response function[1] defines an action as,

$$b(\gamma, A) := \arg\max_{a \in A} \langle u_a, \gamma \rangle; \tag{1}$$

while the value function is given by,

$$V(\gamma, A) := \max_{a \in A} \langle u_a, \gamma \rangle. \tag{2}$$

3.2 Information Structures

Even though the state is uncertain at the outset the DM can reduce the uncertainty by acquiring information. This is modeled by the choice of an information structure defining conditional distributions over signals for each realization of the state. Upon observing a signal the DM updates her prior information in a Bayesian manner and selects the best available action according to the corresponding posterior.

Information structures can be formally represented in different ways and the usefulness of a particular representation depends mostly on how explicit one wants to make the signal structure inducing the posteriors. For analyzing the decision making, where prior and posterior beliefs play the pivotal role, it is convenient to ignore signals and focus solely on the induced posteriors. For analyzing the induced choice probabilities it is more convenient to represent information structures as a family of probability distributions over a signal space which makes explicit that we have a conditional distribution over posteriors for each realization of the state. Since the primary focus is on analyzing the induced stochastic choice behavior in each state I define information structures using the signal space formulation and derive induced posteriors from this structure.

Definition 1 (Information structure)**.**
Let $\mathcal{S}$ be an arbitrary signal space. For a finite subset, $S \subset \mathcal{S}$, an information structure is a mapping (Markov kernel), $P \colon \Omega \mapsto \Delta(S)$, specifying conditional distributions over signals for each state. For ease of notation denote the conditional probabilities as $P_\omega \in \Delta(S)$ for each $\omega \in \Omega$. Let $\mathcal{P}(\Omega)$ be the space of all information structures defined over the same state space, Ω, with finitely many signals in their support.

Later I will show that under a general class of information cost functions it is never optimal for the DM to observe more signals than the number of available actions in the choice set. Given finite choice sets the finite support assumption on the information structures is without loss of generality.

[1] Assume that the best response correspondence is single valued, if needed consider any tie-breaking rule.

Given a prior distribution over states, $\mu \in \Delta(\Omega)$, we can integrate over states and denote the unconditional probability of signal s under μ as,

$$P_\mu(s) := \sum_\omega \mu(\omega) P_\omega(s). \tag{3}$$

The corresponding posterior distribution given signal $s \in \mathcal{S}$ is,

$$\gamma^s(\omega) := \frac{\mu(\omega) P_\omega(s)}{P_\mu(s)}. \tag{4}$$

Together with the prior the information structure defines a distribution over posteriors. Denote the support of the information structure as $\mathcal{S}(P) := \text{supp}(P_\mu)$.

3.3 The Cost of Information

The restriction on the signals the DM can potentially acquire is implicitly defined through a cost function over information structures. This is in the spirit of rational inattention (RI) theory following the seminal work of Sims [2003], who considered the possibility of acquiring all possible signals subject to a cost proportionate to the mutual information between states and signals. The flexibility to acquire any kind of signal is a plausible assumption when information acquisition is not about *producing* new information but rather about *digesting* existing one. When information is abundant the DM's primary problem is to allocate her attention to deliberately digesting certain pieces of information.

The use of mutual information ensures analytical tractability at the cost of restrictive behavioral implications, for this reason there are several other cost functions considered in the literature. I follow this more flexible approach in my main analysis.

An *information cost* function is a mapping from information structures and priors to the extended reals, $K \colon \mathcal{P}(\Omega) \times \Delta(\Omega) \mapsto \overline{\mathbb{R}}$. I follow the standard assumption that more information—in the Blackwell sense—has higher costs.

Definition 2 (Blackwell informativeness)**.**
Information structure $P^1 \in \mathcal{P}(\Omega)$ is Blackwell more informative than information structure $P^2 \in \mathcal{P}(\Omega)$, denoted as $P^1 \succeq_B P^2$ if for $\mu \in \text{int}\Delta(\Omega)$,

$$\sum_{s \in \mathcal{S}(P^1)} \phi(\gamma^s) P^1_\mu(s) \geq \sum_{s \in \mathcal{S}(P^2)} \phi(\gamma^s) P^2_\mu(s), \tag{5}$$

for every convex continuous function $\phi \colon \Gamma \mapsto \mathbb{R}$.

Note, that the value function, $V(\gamma, A)$, is convex in γ for any decision problem, A, hence it is straightforward that Blackwell informativeness implies higher achieved utility in expectation before observing a signal,

$$P^1 \succeq_B P^2 \implies \sum_{s \in \mathcal{S}(P^1)} V(\gamma^s, A) P^1_\mu(s) \geq \sum_{s \in \mathcal{S}(P^2)} V(\gamma^s, A) P^2_\mu(s).$$

Absent monotonicity in the Blackwell order the DM would always change the information structure in directions which are more informative but not costlier. From a behavioral perspective the resulting

choices could be rationalized by another cost function that is monotone in the Blackwell order. The above definition of Blackwell informativeness also provides a recommendation for defining posterior separable cost functions through convex functions over the space of posteriors.

Definition 3 (Posterior Separable[2] Cost function)**.**
An information cost function is posterior separable *if it can be written in the following form,*

$$K(P,\mu) = \kappa \sum_{s\in\mathcal{S}(P)} \phi(\gamma^s) P_\mu(s), \tag{6}$$

for a convex function $\phi\colon \Gamma \mapsto \mathbb{R}$ *and positive scalar* κ*.*

Posterior separable cost functions provide analytical tractability while also allow for flexible behavioral patterns that could not arise under the mutual information cost. Denti [2018] and Caplin et al. [2017] following a revealed preference approach give characterizations of posterior separable cost functions in terms of their behavioral implication derived from revealed posteriors in the data. Hébert and Woodford [2017] show equivalence of the posterior separable class with the outcome of a sequential information sampling problem and study neighborhood based cost functions that break symmetry across states. Neighborhood based cost functions have also been studied by Morris and Yang [2016] in the context of continuous play in coordination games. Other more general cost functions are also considered in the literature [de Oliveira et al., 2017, Chambers et al., 2018].

Posterior separability of the cost functions allows the use of standard convex analytic tools to solve for the optimal information structure and geometrically renders the attentional problem to a concavification exercise. As a result, it provides a balance between flexibility and tractability for analyzing comparative statics.

3.4 The Attention Allocation Problem

The attentional problem of the DM is to choose an information structure that maximizes ex ante expected utility net of the information costs. Given the assumption of posterior separable cost functions, the problem lends itself to an intuitive geometric interpretation. I consider the case when ϕ is strictly convex.

$$\max_{P\in\mathcal{P}(\Omega)} \left\{ \sum_{\mathcal{S}(P)} V(\gamma^s, A) P_\mu(s) - \kappa \sum_{\mathcal{S}(P)} \phi\left(\gamma^s\right) P_\mu(s) \right\} \tag{7}$$

The formulation of problem 7 makes it explicit that the entire objective function is separable in the induced posteriors.

$$V(\gamma^s, A) - \kappa\phi\left(\gamma^s\right) = \max_{a\in A} \langle u_a, \gamma^s\rangle - \kappa\phi\left(\gamma^s\right) = \max_{a\in A} \left\{\langle u_a, \gamma^s\rangle - \kappa\phi\left(\gamma^s\right)\right\}, \tag{8}$$

as for a given posterior the function ϕ is independent of the chosen action. The right hand side of expression 8 is the upper envelope of a collection of concave functions, it is typically not concave. As the distribution over posteriors induced by the information structure has to satisfy the Bayesian consistency of averaging back to the prior the maximum value of problem 7 is given by evaluating the concavified upper

[2] Caplin et al. [2017] call this class Uniformly Posterior Separable cost functions.

envelope at the prior. For a more detailed discussion on the relation to convex conjugates see appendix A.

Since the function ϕ is strictly convex there is a unique solution to the problem which we denote by P^*. The support of P^* is constrained by the cardinality of the action set, $|A|$. A straightforward implication of strict convexity of ϕ which we prove in appendix A.1.

Proposition 1 (Support of P^*)**.**
The support of the optimal information structure can not contain more than $|A|$ distinct signals, each inducing separate actions. That is, $\forall s, r \in \text{supp}(P^*_\mu)$ with $\gamma^s \neq \gamma^r$ we have,

$$b(\gamma^s, A) \neq b(\gamma^r, A). \tag{9}$$

This implies that without loss of generality we can take the signal space to be of the same cardinality as the action space and identify each signal as a recommendation for choosing a certain action. For the remainder of the paper I act accordingly and let $S = A$, unless otherwise noted.

3.5 Stochastic Choice Functions

We describe the DM's stochastic behavior in all potential choice sets via a stochastic choice function.

Definition 4 (Stochastic choice function)**.**
A stochastic choice function is a mapping, $\rho\colon \mathcal{Z}^\Omega \times \mathcal{A} \mapsto [0,1]$, such that for any $A \in \mathcal{A}$,

$$\rho(a, A) > 0 \quad \textit{implies that} \quad a \in A;$$

$$\sum_{a \in A} \rho(a, A) = 1.$$

The attention strategy P^* specifies a collection of conditional stochastic choice functions, one for each possible realization of the state. Given the prior distribution, integrating over the states gives the unconditional stochastic choice function. Which of these choice functions is applicable in a given context is mostly dependent on the information set of the analyst relative to the DM. In case the inattentive DM is learning about an observable feature of the environment that is available for the analyst, like the tax rate of a certain product, the use of conditional stochastic choice functions is granted. On the other hand, if the information acquisition regards features of the environment that are uncertain for both the DM and the analyst, like the quality of the match with a given alternative, then the unconditional stochastic choice function is more applicable. I elaborate on this in section 6 while discussing inference from observational data.

Given the primitives describing the environment, $(\Omega, \mathcal{A}, \mu, u, \kappa, \phi)$, we define the stochastic choice functions of the RI model through the induced distributions over posteriors. For the conditional stochastic choice functions, for all $\omega \in \Omega$, we have,

$$\rho^{\text{RI}}_\omega(a, A) = P^*_\omega \{\, s \in \mathcal{S}(P^*) : a = b(\gamma^s, A) \,\}. \tag{10}$$

The unconditional stochastic choice function is the prior weighted average of the conditional choice prob-

abilities.

$$\rho^{\text{RI}}(a, A) = P^*_\mu \{ s \in \mathcal{S}(P^*) : a = b(\gamma^s, A) \}. \tag{11}$$

Importantly the choice probabilities P^* are defined endogenously relative to a given choice set A, hence key features of the RI model will typically exhibit context dependence.

3.5.1 A Digression on Random Utility Models

The stochastic choice function formulation makes explicit the relationship between RI and random utility models (RUM). Given the uncertain context we consider random expected utility models as in Gul and Pesendorfer [2006].

Taking the same state space, action space and prior as before, $(\Omega, \mathcal{A}, \mu)$, the new primitives of the problem are a set of utility functions over outcomes, $\mathcal{V}$, and a distribution over the utility functions, $Q \in \Delta(\mathcal{V})$. The decision maker is assumed to be an expected utility maximizer, and given the primitives $(\Omega, \mathcal{A}, \mu, \mathcal{V}, Q)$ the stochastic choice function is defined as,

$$\rho^{\text{RUM}}(a, A) = Q \left\{ \nu \in \mathcal{V} : a = \arg\max_{a \in A} \langle \nu_a, \mu \rangle \right\}. \tag{12}$$

The stochastic choice function makes explicit the source of randomness in the different representations. Stochasticity in the RI model stems from the DM's noisy information processing under stable preferences, while in the case of RUMs it stems from exogenous shocks to preferences. Given choice set A the observational equivalence results in the literature [Matějka and McKay, 2015, Fosgerau et al., 2017] focus on finding utility and information cost functions for the RI model that generate the same multinomial distribution over A as does a distribution over utility functions in the case of RUM. However, even these matching pairs of representations diverge as we change the choice set A.

The endogeneity of the optimal information structure P^* generates patterns of behavior across choice sets that are not reproducible with random utility models. A prime example is monotonicity, a property shared by all random utility models—that is adding a new alternative to the choice set can not increase the choice probability of other alternatives already in the choice set. Monotonicity can be violated in RI models as shown theoretically by Matějka and McKay [2015] and empirically by Dean and Neligh [2017].

Comparative statics under the two model can be starkly different. Next, using the stochastic choice function the RI model induces, I show that the attentional problem can turn alternatives that are substitutes in utility to complements in behavior.

4 Comparative Statics

In this section I analyze comparative statics of the stochastic choice function resulting from the RI model. In particular, I'm interested in how predicted choice probabilities change with changes in the environment. Since the available alternatives are essentially lotteries I analyze changes in the stochastic choice function with the unilateral change of an alternative in a single state. With a slight abuse of notation I keep the

action labels fixed and denote the change in utility due to an improvement in the outcome of action a in state ω as $\partial u_a(\omega)$.

I define attentional complements by the increasing choice probability of an action due to another action becoming unilaterally more attractive. That is the improvement in action a does not cause the utility of action c to increase as in the case of complementarities stemming from utility.

Definition 5 (Attentional Complements)**.**
For a given choice set with $a, b \in A$ action b is an attentional complement of action a if there is a state, ω, such that,

$$\frac{\partial \rho^{RI}(b, A)}{\partial u_a(\omega)} > 0. \tag{13}$$

As introduced in Caplin, Csaba, Leahy, and Nov [2018a] I approach the attentional problem analogous to the problem of a competitive firm and apply standard tools from microeconomics to decompose the overall effect of the change in $u_a(\omega)$ on choice probabilities.

Keeping the primitives constant write the auxiliary problem for a fixed budget of attention as,

$$\max_{P \in \mathcal{P}(\Omega)} \sum_{s \in \mathcal{S}(P)} V(\gamma^s, A) P_\mu(s)$$

$$\text{such that} \qquad K(P, \mu) \leq C. \tag{14}$$

Denote the solution for action set, A, and the constant attention budget, C, as $P^H(A, C)$ and the corresponding Lagrange multiplier as $\lambda(A, C)$. Denote the solution to the original flexible attention budget problem adjusting the cost by multiplying it with the Lagrange multiplier $\lambda(A, C)K$ as $P^*(A, \lambda(A, C))$. The Lagrange multiplier ensures that,

$$P^*(A, \lambda(A, C)) = P^H(A, C). \tag{15}$$

Proposition 2 (Decomposition)**.**
The change in the conditional stochastic choice function due to an increase in $u_a(\omega)$ can be decomposed for any state $\tilde{\omega} \in \Omega$ and any action $b \in A$ as,[3]

$$\frac{\partial \rho^{RI}_{\tilde{\omega}}(b, A)}{\partial u_a(\omega)} = \frac{\partial P^H_{\tilde{\omega}}(A, C)(b)}{\partial u_a(\omega)} - \frac{\partial P^*_{\tilde{\omega}}\left(A, \lambda(U, C)\right)(b)}{\partial \lambda(A, C)} \frac{\partial \lambda(A, C)}{\partial u_a(\omega)}. \tag{16}$$

The above decomposition follows straight from taking the total derivative of equation 15, analogous to the standard Slutsky decomposition of demand functions. The first term on the right-hand-side of equation 16 captures the change in choice probability due to the pure substitution effect. Keeping the overall attention budget fixed the curvature of the information cost function defines the impact of changing a payoff—tilting the hyperplane of iso-expected utilities. The second term is the attentional budget effect, the change in the choice probability due to changing incentives to invest more or less attention before making a decision. The overall change in the choice probability is determined by the two effects.

For example, in case an increase in $u_a(\omega)$ reduces the utility gap between the best and worst action in state ω the incentive to invest attention decreases. Hence the Lagrange multiplier is decreasing in utility and

[3] See Rockafellar and Wets [2009, Theorem 1.17] for sufficient regularity conditions such that the solution changes continuously.

the change in the information structure with respect to the Lagrange multiplier defines how decreasing invested attention affects the choice probability.

Since the unconditional choice probability is the prior-weighted average of the conditional choice probabilities attentional complements necessitate the positive cross derivative of conditional choice probabilities in at least one state.

5 An Illustrative Example

To demonstrate the introduced concepts consider a simple example with three actions and three states. Three is the minimum number of actions such that there is a possibility of observing co-movements of unconditional choice probabilities—as choice probabilities have to sum up to one.

The payoff matrix is given by,

	a	b	c
ω_1	9	18	15
ω_2	13	14	15
ω_3	20	10	15

Let the prior be uniform, $\mu = (1/3, 1/3, 1/3)$, so that under the prior the expected utilities corresponding to the three actions are,

$$\mathbb{E}_\mu \left[(u_a, u_b, u_c)\right] = (14, 14, 15).$$

All actions can be optimal given the realization of the state. Action a is risky and yields the highest payoff in state ω_3 and the lowest one in state ω_1; action b is a also risky and yields the highest payoff in state ω_1 and the lowest one in ω_3; action c is a riskless action, which is optimal if the realized state is ω_2.

There are two points of reference that do not require further analysis. In case of no costs of attention the DM is fully informed no matter what the realization of the state is and chooses the action that yields the highest possible utility in the given state. Correspondingly, conditional choice probabilities are degenerate and the unconditional choice probabilities reflect the prior probability over states, $P^*_\mu = (1/3, 1/3, 1/3)$. If the cost of acquiring information is prohibitive, the DM chooses the action that yields the highest expected payoff under the prior, in the given example action c. In this case, both the conditional and unconditional choice probabilities are degenerate with $P^*_\omega = (0, 0, 1)$ for all ω and corresponding $P^*_\mu = (0, 0, 1)$. These two extreme reference points could be captured by setting $\kappa = 0$ or $\kappa = \infty$, respectively. The more interesting cases in between these extreme scenarios are explored next.

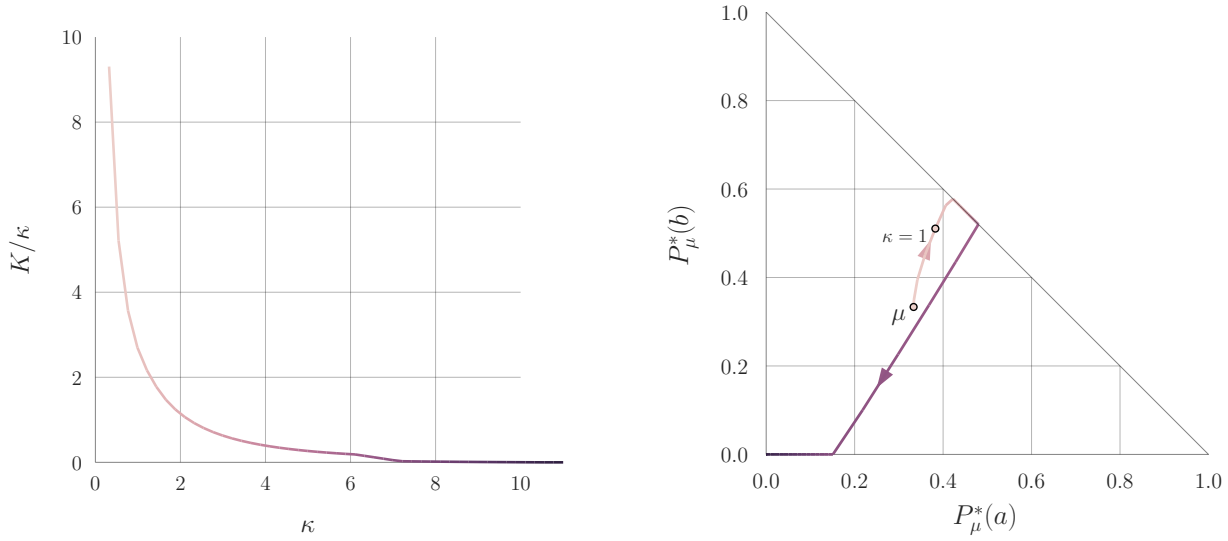
5.1 RI under the Shannon Cost

To showcase comparative statics consider the benchmark RI model with mutual information cost function. The (scaled) mutual information between signals and states is given by,

$$K(P, \mu) = \kappa \sum_{s \in \mathcal{S}(P)} P_\mu(s) \left(\sum_\omega \gamma^s(\omega) \log \frac{\gamma^s(\omega)}{\mu(\omega)} \right). \tag{17}$$

5.1.1 Uniform Changes in the Cost of Information

Consider the effect of changing the cost of information relative to the payoffs. This is achieved by varying κ, the multiplier on the cost of attention. The unconditional choice probabilities move from the uniform distribution—no attention costs—to the degenerate distribution centered at alternative c—prohibitive attention costs. The lines in the plots are color-coded and get darker as κ is increasing.



(a) Invested attention under varying κ

(b) Unconditional choice probabilities under varying κ

Figure 1: Impact of scaled attention costs

Figure 1a shows that as the costs of information increases the DM invests less attention—measured by the normalized mutual information.

Figure 1b maps the optimal unconditional choice probabilities on the probability simplex for the three actions in the choice set. When information is costless the DM chooses actions according to the prior probability μ, whereas when information is prohibitively costly the DM chooses action c with probability one—lower-left corner of the simplex. For intermediate costs the evolution of choice probabilities between these two extreme points demonstrates that probabilities can move non-monotonically in the total invested attention. For intermediate values of attention costs the DM is better off dropping the safe action c altogether, hence first lowering the probability of action c and then increasing it in all states as the costs of attention gets prohibitively high. This is because the DM directs attention to distinguishing states that yield the highest utility gain when making a correct decision.

Suppose that under some intermediate value of the attention costs the DM can choose to obtain signals that result in two posteriors, each characterized by being relatively confident about the realization of a state. Clearly, if there are no differences in informational costs across states the DM would direct these signals to help making a correct decision in the states where the relative differences in payoffs are highest across alternatives—states ω_1 and ω_3. Accordingly, state ω_2, is never revealed and the corresponding best action, c is never chosen.

Since the unconditional choice probabilities are the prior-weighted average of the conditional choice proba-

bilities, it is informative to break down the optimal information structures and analyze the choice patterns in each state. Figure 2 demonstrates their evolution with increasing costs.

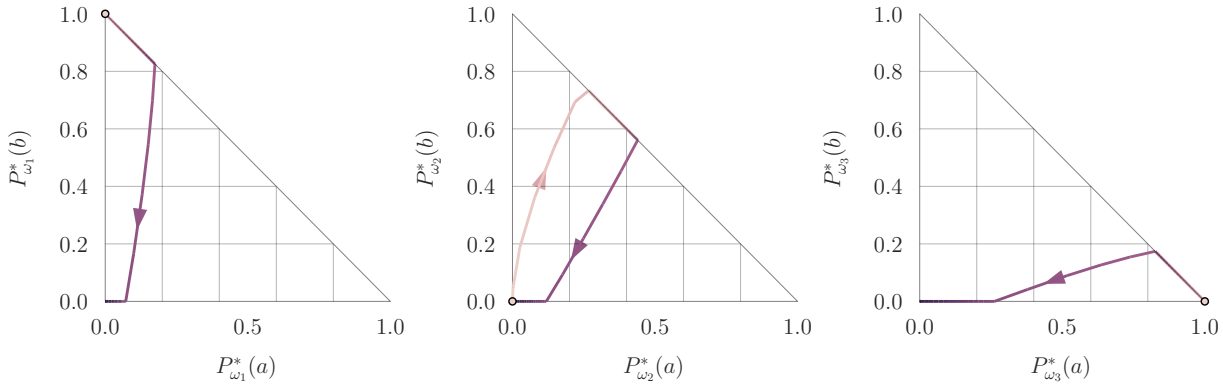


Figure 2: Conditional choice probabilities under varying κ

If attention is costless each alternative is chosen with probability one in the state in which it is optimal—each corresponding corner denoted by a circle. If information is prohibitively expensive, alternative c is chosen with probability one in each state. These explain the starting and ending points of the evolution of choice probabilities in each state. In states ω_1 and ω_3 we see a relatively symmetric switch from the optimal action in the corresponding state to action c. The case of state ω_2 is quite different, under intermediate costs of attention the safe option is dropped. Again, the explanation for such a behavior is that the DM attempts to distinguish states ω_1 and ω_3 where payoff differences are highest. However, once one of these states is considered highly probable it is not worth for the DM to further investigate and tell state ω_2 apart.

5.1.2 Change in the Payoff Structure

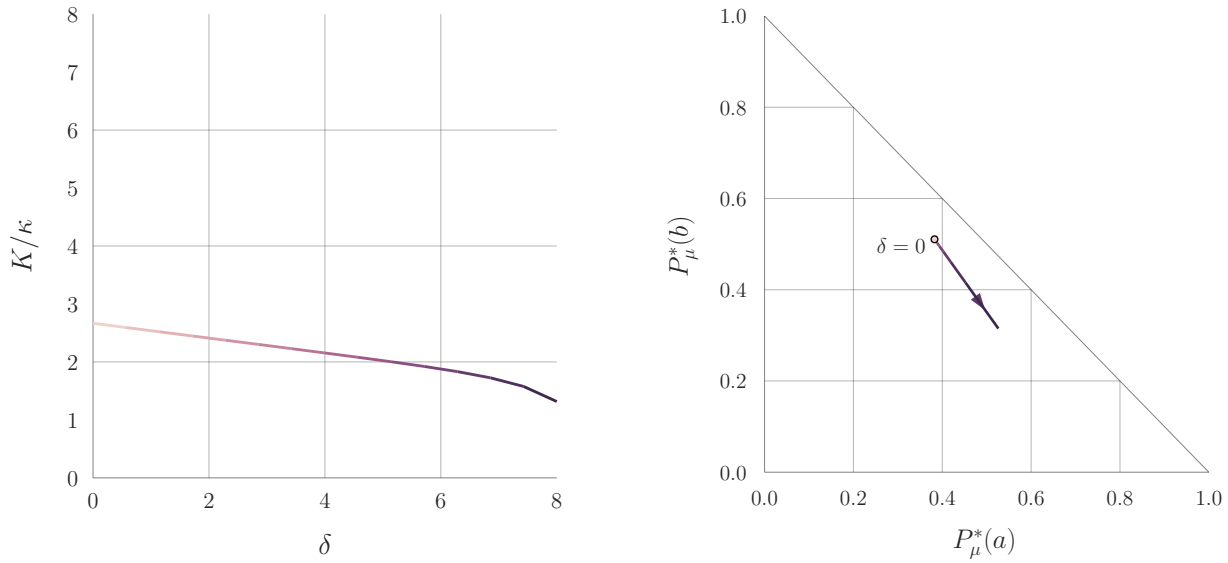
Next consider a change in the payoff of action a in state ω_1 in which action b would be optimal. More specifically, consider the modified payoff matrix,

	a	b	c
ω_1	$9+\delta$	18	15
ω_2	13	14	15
ω_3	20	10	15

,

where $\delta \in [0, 8]$. Figure 3 illustrates the evolution of unconditional choice probabilities and overall invested attention with changes in δ (the multiplier on the mutual information is held constant at $\kappa = 1$).

Notice that the overall invested attention is decreasing in δ as depicted in figure 3a. As action a becomes more attractive in state ω_1 the gain from making the optimal choice at ω_1—taking action b—diminishes and the DM has lower incentives for investing her attention in the exact realization of the state. Correspondingly, the overall level of invested attention decreases. Uniformly scaling up payoffs—lowering κ—would imply increasing information acquisition, however, changes in the relative stakes can change the amount of acquired information in both directions since it is changing the potential gains from information.



(a) Acquired information under varying δ

(b) Unconditional choice probabilities under varying δ

Figure 3: Impact of increased value of $u_a(\omega_1)$

Changes in the overall invested attention is indirectly shaping the patterns in stochastic choice.

By increasing the payoff of alternative a in state ω_1 we would expect that both the unconditional and the conditional choice probability in ω_1 are increasing for a, which in fact is the case. However, the unconditional choice probability of action c is also increasing. In fact, the magnitude of this change is roughly 5 percentage point, increasing the probability of c being chosen from 11% to 16%. Even though action a and c are substitutes in utility, they show up as complements in observed behavior due to the endogenously invested attention.

Figure 4 breaks down the unconditional choice probabilities across states. We see that the comovement of actions a and c is driven by the conditional choice probabilities in state ω_2

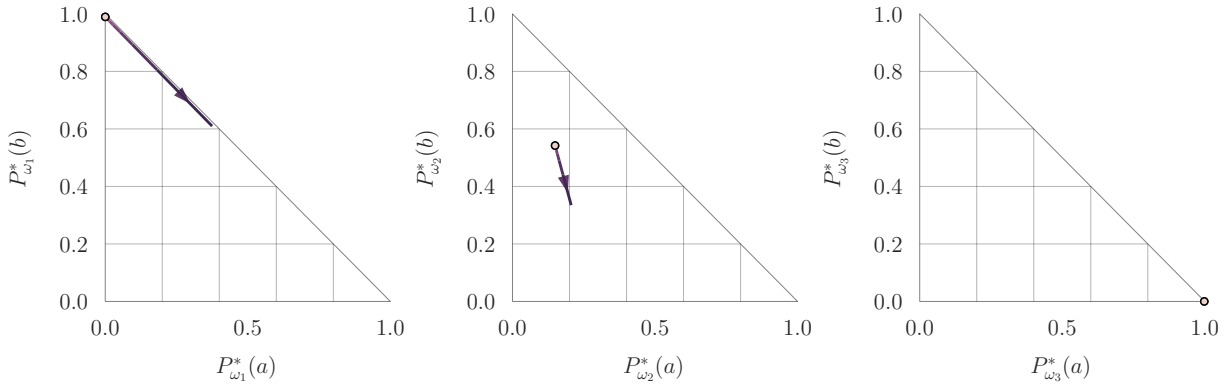


Figure 4: Conditional choice probabilities under varying δ

We can decompose the total change in choice probability into attention budget and substitution effects as demonstrated in proposition 2. The dynamics due to changes in the total invested attention is tantamount

to changing κ. As the payoff of $u_a(\omega_1)$ increases the relative gains from information diminish and the DM acquires less information overall—as shown in figure 3a. In the relevant range, investing less attention implies that the optimal probability of action c decreases—shown in figure 1b. Hence the attentional budget effect is negative. Since, the overall change in the probability of c is positive this implies that the change in P^H must be positive and larger in magnitude relative to the attentional budget effect.

6 Estimation and Inference for the RI model

This section discusses estimation and inference for the rational inattention model from observational data. The primary focus is on recovering the utility and prior parameters given a cost function and then conducting counterfactual predictions using the estimated parameters.

Making inference about attentional problems typically requires rich datasets. This is because the optimal allocation of attention is determined by all potential payoffs both across alternatives and across states, creating context dependence in two dimensions.

Following Caplin and Dean [2015] and Caplin and Martin [2015] a large strand of the literature on attention builds on the availability of state dependent stochastic choice data (SDSC), that is, for each potential realization of the state the analyst observes the choice probability over alternatives. This assumption is plausible in experimental settings where the information sets of the analyst and the DM differ—the analyst observes the state whereas the DM acquires costly information about it—and the analyst can manipulate the state.

Manipulation of the state is rarely plausible in observational data. Since optimal choice probabilities are determined by all the payoffs in states that could potentially realize, making inference based on choices being distributed according to one of the conditional choice probabilities for a single realization of the state will not allow for counterfactual predictions. Choice has to reflect information about all conditional choice probabilities.

All approaches that aim to make inference from observational data assume that choices are distributed according to the unconditional choice probability. Caplin et al. [2015] take a steady state approach whereby agents can infer the distribution of types from the equilibrium market shares of available alternatives. The types define the quality of the match with specific products. This allows one to disentangle utility and attentional parameters in the steady state. Joo [2018], on the other hand, takes an indirect approach that allows for the direct estimation of unconditional choice probabilities and then backs out prior probabilities over states that justify those as the outcome of a rational inattention model. He assumes that certain covariates affect beliefs but do not enter the valuation directly. For this estimation strategy to be valid all unconditional choice probabilities have to be interior and hence no consideration sets are allowed.

Similar to the two examples above, I model individuals as acquiring information about the quality of the match with a given alternative and not about the observable characteristics of those alternatives. In this sense, there are latent valuation types that the DM tries to learn about through acquiring costly information. From an ex ante point of view, the analyst and the DM have the same information set.

The approach I take allows for estimation of the RI model under any posterior separable cost function. It builds on the moment conditions defined by the optimal information structures irrespective of what information cost function generated them. This circumvents issues with the presence of corner solutions

that are specific to likelihood based approaches. Further, the gradient free estimation based on sampling from the quasi-posterior also works in the case when choice probabilities are not all interior. The possibility of the formation of individual level consideration sets and accommodating a large variety of information cost functions are key features that distinguish my approach from other works in the literature.

6.1 Assumptions and Data

I discuss identification and estimation of the RI model from individual level choice data with both alternative specific—actions are identified with the selection of an *alternative* in the choice set—and individual specific demographic covariates. For ease of notation, subscripts denoting individuals are abandoned when dealing with generic observations and the finite $|A| =: N$ number of alternatives from choice set A are enumerated as $a_1, \ldots, a_n, \ldots, a_N$, with a denoting a generic alternative.

The vector of alternative-specific covariates is denoted by $X_a \in \mathbb{R}^{d_X}$ for each $a \in A$—for generic covariates I use X without the alternative specific subscript. Denote the stacked covariates for all of the available alternatives in the choice set by $X_A \in \mathbb{R}^{d_X \times N}$. In addition to the observable covariates on alternatives, the individual specific demographic covariates are denoted by $Z \in \mathbb{R}^{d_Z}$. Lastly, the random choice vector $Y \in \mathbb{R}^N$ denotes choices with the n-th entry being $Y_n = 1$ if the chosen alternative is a_n and zero otherwise. Denote the potentially available data for a generic observation as $D := (Y, X_A, Z)$ with the random variables (Y, X_A, Z) having range $\mathcal{Y} \times \mathcal{X}^N \times \mathcal{Z}$ and realizations (y, x_A, z). Hence, for each decision problem we have information about which alternative is chosen, covariates describing alternatives within the choice set, and demographic covariates on the individual.

Assume that the utility level corresponding to an alternative is defined by the observable covariates on the alternative and an unobservable valuation vector that is random from both the decision maker's and the econometrician's perspective. The valuation vector defines the type of the consumer specifying the quality of her match with different alternatives. This latent valuation type is unknown for both the DM and the econometrician. The prior probability of the valuation vector can depend on the observable demographic covariates. Correspondingly, the relevant uncertainty that the DM can acquire information about concerns her valuation type.

For the sake of simple exposition, I assume linear models for both the utility levels and the prior probability of the valuation vectors. Specifically, assume that the utility of alternative a in state ω can be described by a simple linear function of alternative specific attributes, X_a, and the valuation vector, $\beta(\omega) \in \mathbb{R}^{d_X}$. The DM's prior belief μ is essentially defined over the vectors of valuations, $\{\beta(\omega) \in \mathbb{R}^{d_X} : \omega \in \Omega\}$. Hence, the finite state space, $J := |\Omega|$, describes the types of valuation vectors in the population about which the DM acquires information. When convenient, denote the stacked valuation vectors as $\beta \in \mathbb{R}^{d_X \times J}$ without a subscript and for a specific state ω_j, use the notation β_j for $\beta(\omega_j)$.

The prior probability of the DM having a certain valuation type, μ, depends on individual specific covariates Z such that the log-odds ratio follows a linear model. For the J distinct states this dependence is described by the stacked parameters $\alpha \in \mathbb{R}^{d_Z \times (J-1)}$, normalizing the coefficients corresponding to state ω_0 to zero. Coefficients corresponding to state ω_j are denoted by α_j. Note that the logistic form implies that the prior probabilities of valuation types for any value of the covariates are interior. As discussed before this does not prevent the RI model to produce corner solutions for the probabilities of choosing certain alternatives.

Assumption 1.
Given the observable covariates of the alternative, the utility of choosing alternative a in state ω is defined as,

$$u(a(\omega)) = \beta(\omega)' X_a.$$

Given the observable demographic covariates, the prior probability of having a valuation type $\beta(\omega_j)$ for $j \neq 0$ is defined as,

$$\log \frac{\mu(\omega_j \mid Z; \alpha)}{\mu(\omega_0 \mid Z; \alpha)} = \alpha_j' Z.$$

As noted, the primary focus is on estimating the parameters defining the potential utility levels of each alternative and the prior probability of valuation types for a given information cost function. As in section 3 I assume that the cost function belongs to the posterior separable class so that the solution to the attention problem can be characterized by standard Lagrangean methods and computed numerically using convex optimization techniques.

In principle, one might parametrize the convex function ϕ and estimate the utility and cost parameters jointly, however, this is not the focus of the current paper. Knowledge about the underlying information costs can come from experimental studies. For instance, in our recent paper [Caplin et al., 2018a] we discuss recoverability of the cost function from experimental data given prior beliefs and utility parameters and carry out such a cost recovery exercise in practice.

Assumption 2.
The information cost function takes the form,

$$K(P, \mu) = \kappa \sum_s \phi(\gamma^s) P_\mu(s),$$

for some $\phi \in C^1(\Delta(\Omega))$ being strictly convex and continuously differentiable, and κ a positive scalar.

The behavioral assumption of the RI model is that the stochasticity in choice is defined entirely by the noisy information acquisition process preceding the decision. As a result, given assumption 1 on functional forms and assumption 2 defining information costs, the outcome of the RI model is a choice probability over alternatives in the choice set for each valuation type and the corresponding unconditional choice probability integrating over these valuations with respect to the prior,

$$Y \mid \omega \sim P^*_\omega \qquad \forall \omega \in \Omega, \qquad \text{and} \qquad Y \sim P^*_\mu. \tag{18}$$

Given the representative agent approach and the assumptions on the functional forms, the utility parameters and the multiplier on the information cost only enter the model through $\beta(\omega)' X_a / \kappa$ and hence one can identify them only up to their ratio. Correspondingly, I normalize the multiplier to $\kappa = 1$. For counterfactual predictions under changing information costs the multiplier should be varied relative to this normalized value.

In applications with an outside alternative it is without loss of generality to normalize the utility of the outside alternative to zero across all states. Since the optimal information structure is a collection of J number of probability distributions defined over N alternatives, it has the dimensionality of $\mathbb{R}^{J \times (N-1)}$. It

is determined only by the relative differences in payoffs since $P_\omega(a_N) = \sum_{n=1}^{N-1} P_\omega(a_n)$,

$$\sum_a V(\gamma^a, A) P_\mu(a) = \sum_{a,\omega} \mu(\omega) P_\omega(a) u_a(\omega)$$

$$= \sum_\omega \left(\sum_{n=1}^{N-1} \mu(\omega) P_\omega(a_n) \left[u_{a_n}(\omega) - u_{a_N}(\omega) \right] + \mu(\omega) u_{a_N}(\omega) \right). \tag{19}$$

Correspondingly, in case there is an outside alternative it is without loss of generality to make the normalization $u_{a_N}(\omega) = 0$ for all ω.

The free parameters of the problem are $\alpha \in \mathbb{R}^{d_Z \times (J-1)}$ describing the individual specific prior probabilities of valuation types and the utility parameters $\beta \in \mathbb{R}^{d_X \times J}$. Given parameters $\theta := (\alpha, \beta)$ the model defines conditional and unconditional choice probabilities over the alternatives for each choice set described by X_A and each individual specific prior described by Z.

$$Y \mid \omega, X_A, Z; \alpha, \beta \sim P^*_\omega(Y \mid X_A, Z; \alpha, \beta); \tag{20}$$

$$Y \mid X_A, Z; \alpha, \beta \sim P^*_\mu(Y \mid X_A, Z; \alpha, \beta) = \sum_\omega \mu(\omega \mid Z, \alpha) P^*_\omega(Y \mid X_A, Z; \alpha, \beta). \tag{21}$$

I work with the mixture distributions conditional on the observable covariates (X_A, Z) given in equation 21 as the exact realization of valuation types—denoted by ω—are not observable by the econometrician.

6.2 Identification

In order to obtain identification of the true parameter, $\theta_0 = (\alpha_0, \beta_0)$, we exploit the structure of the mixture model implied by the RI model. Since there is a finite number of available alternatives the choice distributions implied by the RI model are multinomial. Parameters (α^1, β^1) and (α^2, β^2) are observationally equivalent if the corresponding conditional distributions over the choice vector Y given (X_A, Z) have identical distributions,

$$P^*_\mu(y \mid x_A, z; \alpha^1, \beta^1) = P^*_\mu(y \mid x_A, z; \alpha^2, \beta^2) \qquad \forall\ (y, x_A, z) \in \mathcal{Y} \times \mathcal{X}^N \times \mathcal{Z}.$$

That is, the parameters of interest are identified if for any two distinct parameter vectors $(\alpha^1, \beta^1) \neq (\alpha^2, \beta^2)$ we can find $(y, x_A, z) \in \mathcal{Y} \times \mathcal{X}^N \times \mathcal{Z}$ such that the corresponding probability mass functions are different.

With specified functional forms for the prior, the utility function, and the attention cost function, the RI model provides a fully parametric econometric model over observables. The issue that complicates identification of the parameters is the endogenous formation of consideration sets. These imply zero probabilities of selection for certain alternatives. At parameter values where the consideration sets change the probability mass function is non-differentiable. In order to prevent such issues we assume individual level stability of consideration sets.

For each individual $i \in I$, denote instances of choice—time or market—by $t \in T$. Without loss of generality suppose that the number of attributes describing the alternatives are constant. Let $X_{it} \in \mathbb{R}^{d_X \times N}$ describe

the attributes of alternatives in choice set A_{it} that individual i faces at time/market t. The observed choices are recorded in the vector Y_{it} such that $Y_{ita} = 1$ if the chosen alternative is a and zero otherwise. The individual specific covariates on individual i at time t are denoted as $Z_{it} \in \mathbb{R}^{d_z}$. For individual i at time t the available data is then $D_{it} = (y_{it}, x_{it}, z_{it})$.

Assumption 3.
There exist a neighborhood of θ_0, $\mathcal{N}(\theta_0) \subset \Theta$, such that individual level consideration sets are stable. That is,

$$supp\left(P^*_\mu(x_{it}, z_{it}; \theta)\right) \quad \text{is constant } \forall \theta \in \mathcal{N}(\theta_0),\ \forall i \in I, t \in T.$$

The stability condition in assumption 3, together with smoothness conditions on functional forms, ensures differentiability of the implied choice probabilities around the true parameter value θ_0. We can then establish local identification of the parameters using the moment conditions—defined in the next section—as in Chen, Chernozhukov, Lee, and Newey [2014].

6.3 Estimation Strategy

For a certain value of the parameters and realization of covariates the RI model produces conditional and unconditional multinomial choice probabilities over the alternatives.

Using the predicted multinomial choice probabilities to form the total likelihood of the data is problematic as many of the observation specific predictions have corner solutions effectively putting zero probability mass on certain alternatives. This causes numerical problems in standard optimization algorithms: even if at the true parameter value, θ_0, all choice probabilities are interior, for many values of θ in the parameter space there may exist corner solutions, and hence choice probabilities being zero, rendering the criterion function infinite at these values of θ. This causes complications when fed into conventional "black-box" numerical optimization algorithms. Standard nested-fixed point algorithms maximizing the total likelihood—as in Rust [1987]—are therefore not applicable in the context of RI models. For this reason, I follow a moment-based estimation strategy that is well-defined even if some of the predicted probabilities at the observed choices are zero and remains efficient if the model is correctly specified.

Another problem arising due to the corner solutions is that gradient based search algorithms might get stuck at the corner solutions. This is because the presence of corner solutions induce flat regions in the sample criterion function. To circumvent this issue I take a quasi-Bayesian approach as in Chernozhukov and Hong [2003] and Chen, Christensen, and Tamer [2018]. The GMM criterion is converted to a quasi-posterior. I then use the derivative-free sequential Monte Carlo algorithm to sample from the quasi-posterior and obtain estimates minimizing the GMM criterion. Under a flat prior the quasi posterior modes coincide with the standard frequentist estimates without relying on the gradient for parameter search.

6.3.1 Method of Moments Estimation

The model produces conditional probabilities for choosing a certain alternative given individual and choice covariates. The orthogonality conditions implied by the model can be concisely written as,

$$\mathbb{E}[Y \mid x_A, z] = P^*_\mu(x_A, z; \theta_0). \tag{22}$$

Since probabilities have to add up to one only $N-1$ orthogonality conditions are needed. Borrowing notation from programming syntaxes, let

$$f\left(Y, X_A, Z; \theta\right) := Y\texttt{[:-1]} - P_\mu^*\left(X_A, Z; \theta\right)\texttt{[:-1]} \tag{23}$$

be all but the last element of the corresponding vectors. The conditional moment conditions are,

$$\mathbb{E}\left[f\left(Y, X_A, Z; \theta_0\right) \mid x_A, z\right] = \mathbf{0}. \tag{24}$$

Assuming homoskedasticity[4] of the residuals, where

$$\mathbb{E}\left[f\left(Y, X_A, Z; \theta_0\right) f\left(Y, X_A, Z; \theta_0\right)' \mid x_A, z\right] = \Sigma, \tag{25}$$

the optimal instruments (cf. Newey [1990]) take the form,

$$W^*(x_A, z) = \mathbb{E}\left[\frac{\partial f\left(Y, X_A, Z; \theta_0\right)}{\partial \theta} \;\middle|\; x_A, z\right]' \Sigma^{-1}. \tag{26}$$

Since the instruments are functions of the covariates, we can take expectations and obtain the population orthogonality conditions,

$$\mathbb{E}\left[W(X_A, Z) f(Y, X_A, Z; \theta_0)\right] = \mathbf{0}. \tag{27}$$

The criterion for the asymptotically efficient GMM estimator is given by the sample analog of the cross product between instruments and residuals. Denoting $f_{it}(\theta) := f(y_{it}, x_{it}, z_{it}; \theta)$ and $W_{it} := W(x_{it}, z_{it})$,

$$\begin{aligned}\hat{\theta} &:= \arg\max_{\theta \in \Theta} -\frac{1}{2}\left[\frac{1}{n}\sum_{it} W_{it} f_{it}(\theta)\right]' \left[\frac{1}{n}\sum_{it} W_{it} W_{it}'\right]^{-1} \left[\frac{1}{n}\sum_{it} W_{it} f_{it}(\theta)\right] \\ &:= \arg\max_{\theta \in \Theta} Q_n(\theta).\end{aligned} \tag{28}$$

We obtain efficient estimates of the parameters in two steps. First, we obtain consistent estimates, $\hat{\theta}$ by using some instruments $W(x_A, z)$, then we use the estimates from the first stage to estimate the covariance matrix and then the optimal instruments through,

$$\widehat{\Sigma} = \frac{1}{n}\sum_{it} f\left(y_{it}, x_{it}, z_{it}; \hat{\theta}\right) f\left(y_{it}, x_{it}, z_{it}; \hat{\theta}\right)', \tag{29}$$

$$\widehat{W}(x_A, z) = -\frac{\partial P_\mu^*(x_A, z; \hat{\theta})}{\partial \theta'} \widehat{\Sigma}^{-1}. \tag{30}$$

The gradient of the vector of choice probabilities is well defined at parameters such that within a small enough neighborhood the corner solutions do not change. In case the cost function is proportionate to the mutual information we can obtain the gradient of the optimal choice probabilities using the implicit function theorem. Appendix F details the derivation of the gradients.

Under the assumption of correct specification, the above-described instrumental GMM approach yields consistent and efficient estimates as would an MLE approach yield without the zero-likelihood problem

[4] For the heteroskedastic case, see Chamberlain [1987].

arising due to corner solutions. However, as mentioned at the beginning of the section, the corner solutions still render gradient based optimization techniques impractical. A corner solution at a given parameter value implies that the gradient of the choice probabilities will be zero at the zero probabilities. That is, if an alternative had zero probability of being selected under some value of the parameters, a slight perturbation of the parameters would still leave it at zero. This hinders moving in directions where the criterion function would increase.

This problem is solved by using quasi-Bayesian estimation of the parameters as in Chernozhukov and Hong [2003] and Chen et al. [2018]. Given the criterion function, Q_n, and prior over the parameters, π, we form the quasi-posterior as,

$$\pi_n(\theta) \propto e^{nQ_n(\theta)}\pi(\theta). \tag{31}$$

This quasi-posterior puts relatively larger mass over parameters which make the criterion function larger. An adaptive Sequential Monte Carlo (SMC) algorithm is used to sample from the quasi-posterior following Chen et al. [2018, Appendix A] and Herbst and Schorfheide [2014, Section 3]. In contrast with standard MCMC algorithms using the Metropolis-Hastings procedure, this SMC algorithm is robust to multimodality of the quasi-posterior.

7 Experimental Design

I conduct an experiment to study the adjustments in attention strategies as incentives change. Specifically, I analyze the evolution of choice probabilities and the overall level of invested attention under two treatments. First, I vary the gain-to-cost ratio of information acquisition. Second, I vary the payoff structure of the problem while keeping the cost of attention fixed.

The experimental design closely mimics the example introduced in section 5. I consider a decision problem under uncertainty with three available actions, each of which is optimal given the realization of one of three possible states. The task which allows subjects to acquire information follows the design introduced in Caplin, Csaba, Leahy, and Nov [2018a].

In each round, a subject is shown a number of geometric objects, each of which is one of five regular polygons with four to eight sides as depicted in figure 5. The state is defined by which of three polygons, five-sided, six-sided or seven-sided, is most common on the screen. Four- and eight-sided polygons are decoys and the highest-paying action in any given round is independent of their numbers. Their presence lets us vary the total number of the most common shape across rounds even when the total number of all five shapes remains constant. This ensures that by counting just one of the three shapes subjects can not conclude which of them is most common with certainty. Counting only one or two of the shapes, or sampling freely, give subjects partial information about the realization of the state. The simple geometric counting task minimizes ability differences across subjects.

The location and rotation of the polygons and the total number of non-decoy shapes are generated randomly in each round as described in appendix C. This ensures that information from one round does not carry over to another. The prior probability of any of the non-decoy shapes being the most common one is uniform in each round, and subjects are conveyed this information before starting the experiment.

The experiment is conducted on Amazon Mechanical Turk which has a number of advantages for experiments on costly attention. First, subjects have a clear opportunity cost in terms of how much they could

earn by completing other tasks on MTurk. Second, they are not constrained to dedicate a fixed amount of time to the experiment, which is usually the case in lab experiments. The subject pool consists of U.S. MTurkers with at least 100 prior completed tasks with an overall approval rating of 95%.

Subjects are rewarded based on a the action they choose in round that is randomly selected at the end of the experiment. In order to ensure credibility of the randomization device I use the millisecond counter of subjects' built-in computer clock. This is linked to the round numbers in a loop, and subjects are asked to stop the clock to determine which round gets selected for final payment. Subjects do not have control over stopping the counter simply because of limits on human reaction time.

I follow a between subject design in estimating choice probabilities and conduct the two treatments on two separate samples. The specifics of each treatment are discussed in the following subsections. Experimental instructions are in appendix B.

7.1 Varying the Cost of Information

I vary the gain-to-cost ratio of acquiring information by changing the total number of shapes appearing in a round. The total number can take values of 8, 24 or 64. These effectively scale the difficulty level—and hence the cost—of acquiring information. An example of the task with 24 shapes is presented in figure 5.

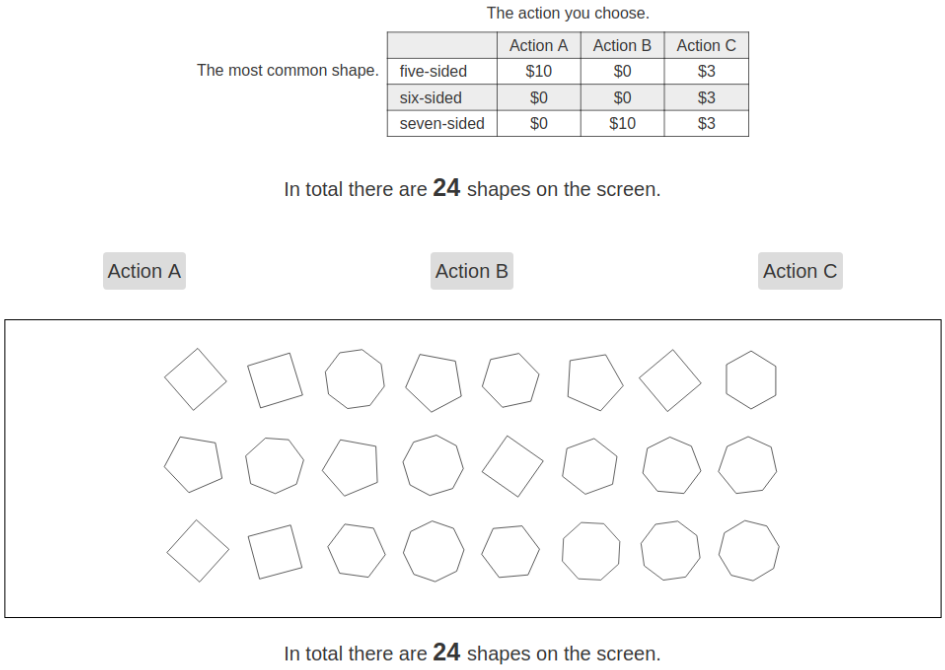


Figure 5: Example of an experimental task with 24 shapes

The payoff matrix depicted in figure 5 is kept constant across all rounds. Subjects have to choose one of three available actions. Two of these actions, Action A and Action B, are risky: they pay off \$10 in one of the three states and nothing in the other two. Action C, on the other hand, is riskless: it pays off \$3 irrespective of the realization of the state. Subjects can acquire information by inspecting the geometric

shapes appearing on the screen. Counting all three non-decoy shapes reveals the state, while counting only one, two or sampling freely give partial information.

Subjects face 30 rounds with the number of total shapes varying on the screen. The number of rounds with a given number of total shapes is summarized in table 1. The order of the rounds is randomized.

Total number of shapes	Number of rounds
8	5
24	20
64	5
Total	30

Table 1: Total number of shapes and number of rounds within a session

In order to prevent subjects from clicking through the experiment always selecting Action C and making a certain payoff of \$3, eligibility for payment is conditional on matching the state with the corresponding highest-paying action above chance. That is, since in each round selecting the highest-paying action has 1/3 chance, in expectation, under completely inattentive strategy, the highest-paying action is selected 10 of 30 rounds. In order to be eligible for payment, subjects are required to select the highest-paying action in at least 15 of 30 rounds. Under full inattention this binomial tail probability is around 4.5%. If a subject gets the 5 rounds with only 8 shapes correct and employs the fully inattentive strategy during the rest of the rounds, she is eligible for payment with roughly 30% probability.

7.1.1 Experimental Results

There are 143 subjects participating at the experiment varying the cost of information, each of them responding in 30 rounds resulting in 4290 points of observation. Response times show considerable heterogeneity in the sample.

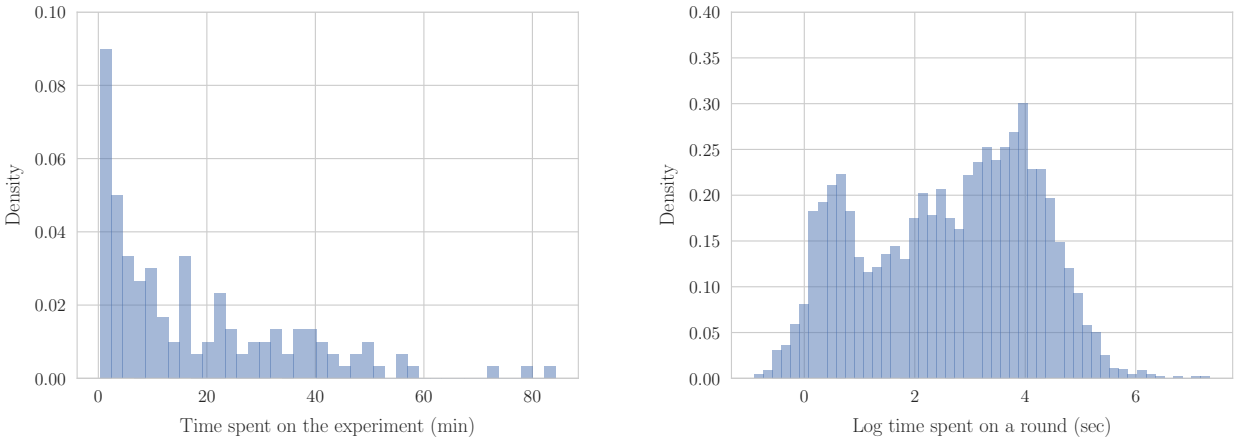


Figure 6: Distribution of time spent on the experiment and time spent on rounds

Figure 6 shows that there is a considerable fraction of subjects who spend less than 2 minutes on all 30 rounds. There is also heterogeneity in response times across rounds, which is partly driven by the varying

difficulty levels.

Altogether, subjects who spend more time on the instructions tend to spend more time on the experiment, as well. Those who spend more time on the experiment, in turn, tend to select the highest-paying action given the most common shape more frequently as shown in the right hand panel of figure 7.

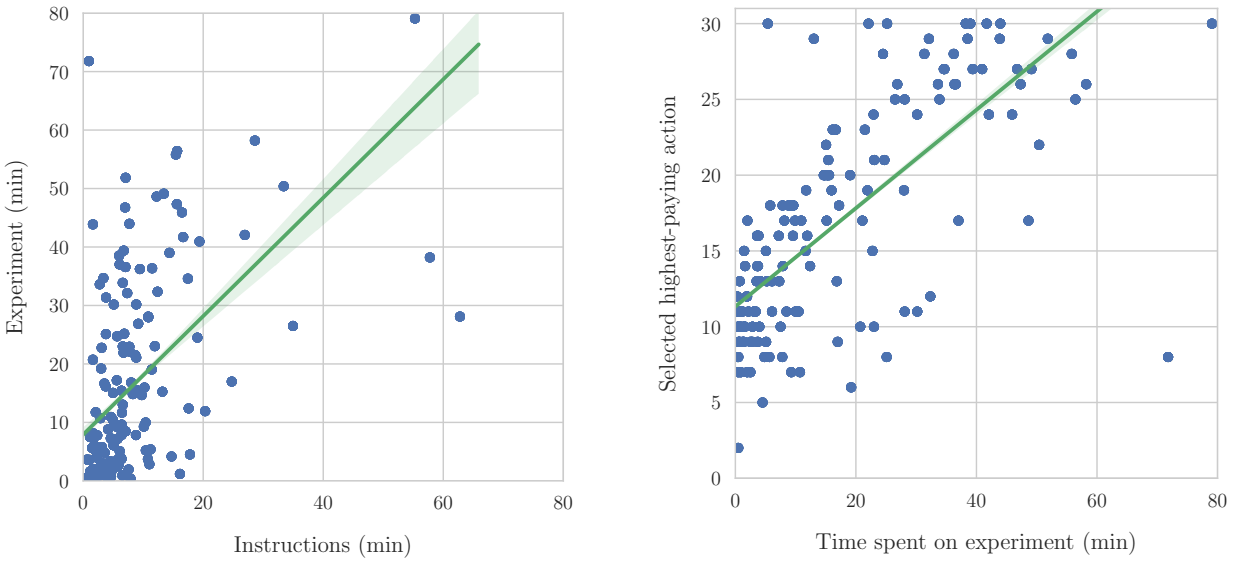


Figure 7: Distribution of time spent on the experiment and time spent on rounds

There are 118 subjects who spend at least 2 minutes on the experiment and 25 who spend less than that. Labeling them attentive and inattentive, respectively, figure 8 shows their median response times and the probability of selecting the highest-paying action across rounds. On the left hand panel, we see that response times decrease over rounds for the attentive subsample, while it remains constant and close to 1-2 seconds for the inattentive sample. The right hand panel shows that there is no clear performance change for either of the two samples, with the performance in the inattentive sample being close to that of pure chance. The average probability of selecting the highest-paying action do not change significantly across rounds for either of the two groups—for the regression table see appendix D.1.

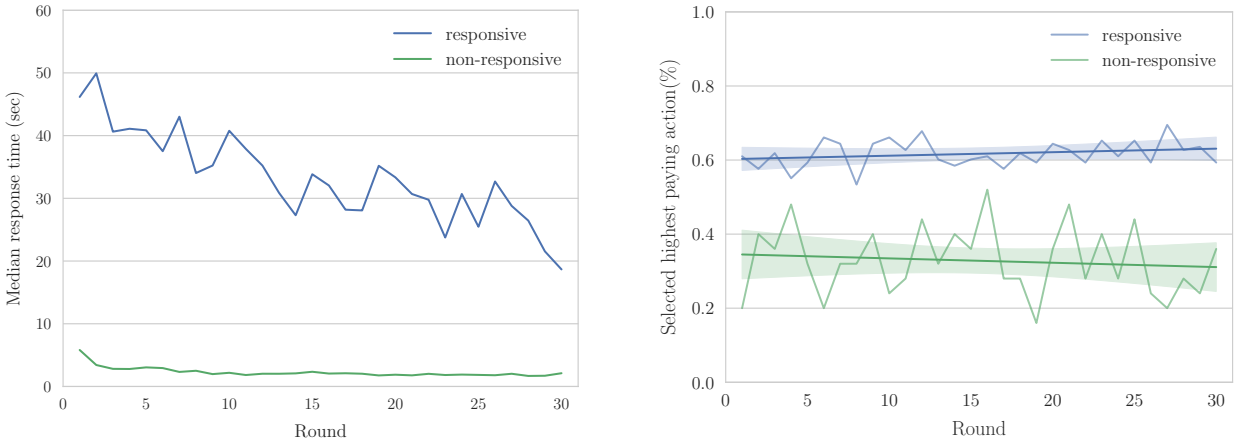


Figure 8: Response times over rounds and probability of matching the state

Subjects who employ the fully inattentive strategy are not inconsistent with the theory. It might be that their cost of completing the task is so high, that they are better off by participating in a lottery that yields no rewards with over 95% chance. However, in order to pool subjects for estimating choice probabilities one needs sufficient homogeneity and hence subjects who don't show any variation in their attention strategies are not included in the estimation.

The 2-minute threshold is set such that there is no significant performance change in the inattentive sample as the total number of shapes vary. The strategies employed by the attentive and inattentive subjects look quite different based on their response times across difficulty levels as illustrated in figure 9. While attentive subjects spend considerably longer time on the medium difficulty rounds with 24 total shapes, inattentive subjects show no significant variation in response time.

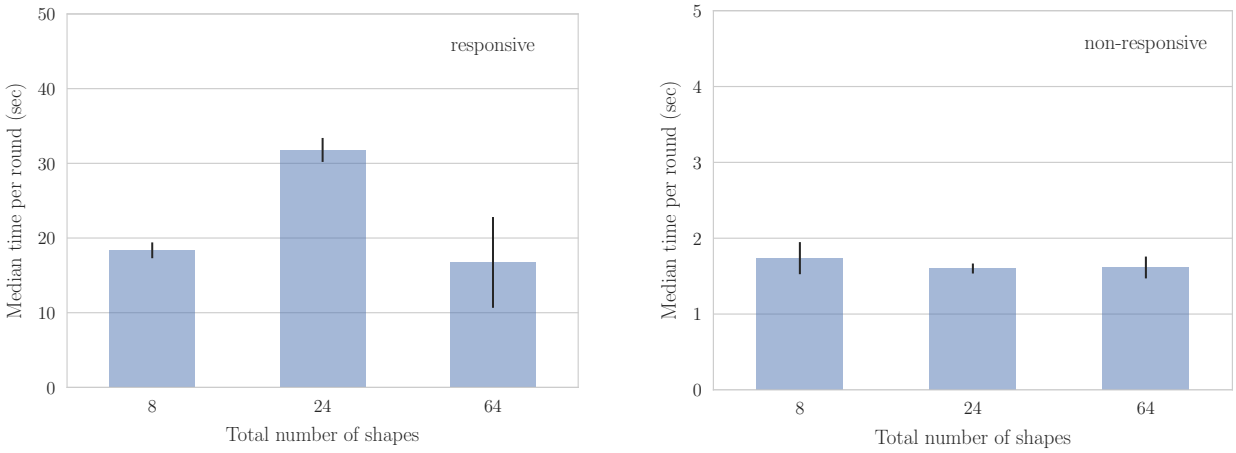


Figure 9: Median response times per round across difficulty levels

How these translate to actual performance in terms of selecting the highest-paying action in a given round is depicted in figure 10. We see that the attentive sample stays significantly above chance—1/3 probability—for all difficulty levels, while performance in the inattentive sample is not significantly different from that of pure chance under any of the difficulty levels.

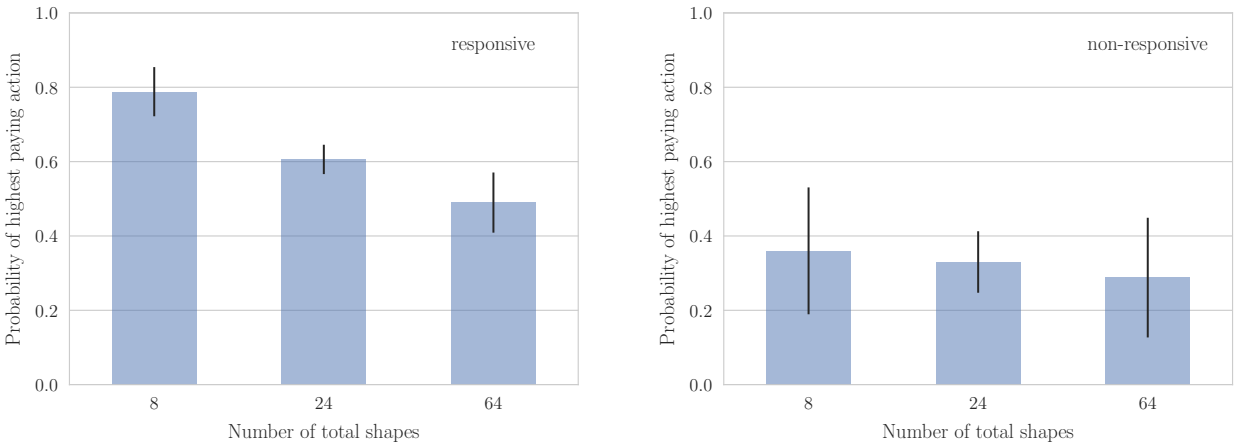


Figure 10: Probability of selecting the highest paying action

The average final payment in the attentive sample is \$4.03, while in the inattentive sample it is \$0.17. Average time spent on the experiment is 21.79 minutes (31.52 with instructions) in the attentive and 0.98 minutes (4.32 with instructions) in the inattentive sample. These result in an hourly wage between \$7.64 and \$11.09 in the attentive sample.

The unconditional choice probabilities of selecting the highest-paying action hide valuable information. For example, one could select the highest paying action 1/3 of the time by always selecting Action C, or by completely randomizing across actions—a vastly different strategy. Revealed posteriors—first introduced by Caplin and Martin [2015]—are much more informative in this respect.

Definition 6 (Revealed posteriors.)**.**
Suppose the experimental data provides the joint empirical distribution over states and actions given by, $P_e \in \Delta(\Omega \times A)$—this is State Dependent Stochastic Choice data [Caplin and Martin, 2015]. The revealed posterior under action a is given by

$$\gamma^a(\omega) := \frac{P_e(\omega, a)}{P_e(a)}, \qquad \forall \omega, a \in \Omega \times A, \tag{32}$$

where $P_e(a) = \sum_\omega P_e(\omega, a)$ is the marginal distribution of selecting action a.

Revealed posteriors and the unconditional probabilities over actions jointly determine the attention strategies employed by subjects. For instance, they show how much information subjects have while selecting an alternative. Figure 11 presents these posteriors in the probability simplex over the actions for both the attentive and inattentive subsamples. The corners denote degenerate beliefs about states corresponding to one of the actions being optimal with certainty. For instance, the lower right corner denotes the posterior that puts probability 1 on the five-sided shapes being the most common, which corresponds to Action A being the highest-paying alternative. The revealed posteriors are denoted based on which action they correspond to under what difficulty level. For instance, γ^A_{24}, denotes the revealed posterior corresponding to Action A and the treatment with the total number of shapes being 24. Other actions and treatments are denoted analogously.

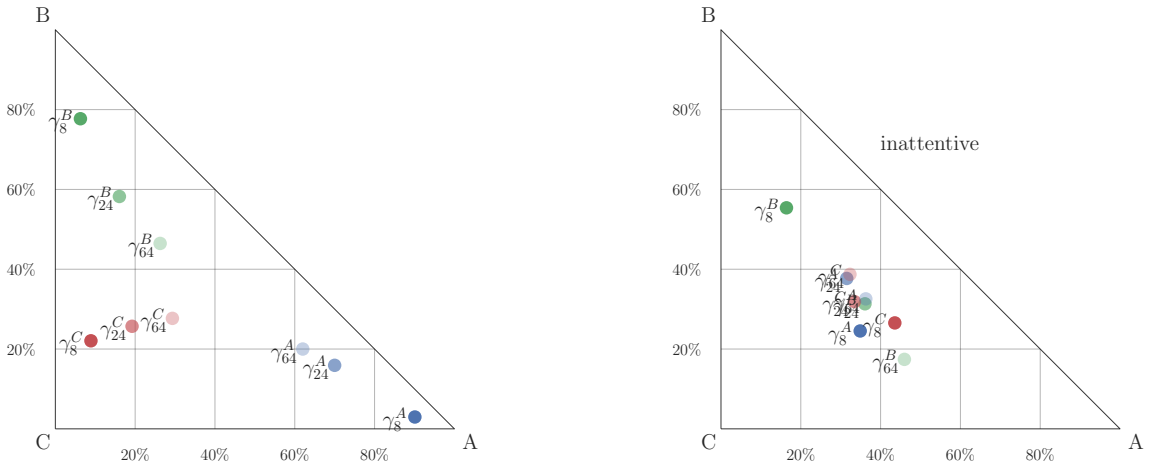


Figure 11: Revealed posteriors – colors fade as total number of shapes increases

The left panel shows the revealed posteriors under different actions and cost treatments for the attentive sample. We see that for lower difficulty levels subjects are more informed, the posteriors essentially being

a dilation of previous posteriors under lower difficulty levels—that is, the employed information structures are Blackwell more informative when the cost of information is lower.

Using the revealed posteriors we can test for the rationalizability of the data by an information cost function. The two, necessary and sufficient, conditions are No Improving Action Switches (NIAS) [Caplin and Martin, 2015] and No Improving Attention Cycles (NIAC) [Caplin and Dean, 2015]. The former essentially requires that each action is rational given the revealed posteriors in the data, while the latter requires that expected utility can not be improved on average by reallocating attention strategies. Following Dean and Neligh [2017], I conduct statistical tests of the NIAS and NIAC conditions.

The NIAS conditions state that,

$$\sum_{\omega} P_e(\omega, a)u(a(\omega)) \geq \sum_{\omega} P_e(\omega, a)u(b(\omega)) \qquad \forall a, b \in A. \tag{33}$$

For the present experimental task, denoting states by ω_k when the k-sided shapes are most common, for Action A we have the following two conditions,

$$P_e(\omega_5, A) - P_e(\omega_7, A) \geq 0; \tag{34}$$

$$P_e(\omega_5, A)u(\$10) - \left(P_e(\omega_5, A) + P_e(\omega_6, A) + P_e(\omega_7, A)\right)u(\$3) \geq 0. \tag{35}$$

The NIAS conditions for the other Action B and C are analogous. In case the utility levels cancel, we can use the empirical probabilities without any further assumption. In general, however, the conditions depend on the curvature of the utility function. For the sake of simple exposition I use the risk neutral case taking the dollar amounts at face value—a different route would be to provide bounds on risk aversion such that the NIAS conditions hold. Out of the 18 NIAS tests 16 are passed at the 1% significance level for the attentive sample while only 2 of the tests are passed at the 5% significance level for the inattentive sample. The p-values for each individual test is provided in appendix D.2.

The NIAC conditions test for rationalizability of the data under a stable cost function. Therefore, for these to be applicable in the present context one has to assume how the cost varies in the total number of shapes appearing on the screen. Assuming only monotonicity in the total number of shapes the NIAC conditions boil down to the following two equations—the superscript on the SDSC data denotes the total number of shapes.

$$\left(P_e^8(\omega_5, A) - P_e^{24}(\omega_5, A)\right) + \left(P_e^8(\omega_7, B) - P_e^{24}(\omega_7, B)\right) + \left(P_e^8(C) - P_e^{24}(C)\right)\frac{u(\$3)}{u(\$10)} \geq 0 \tag{36}$$

$$\left(P_e^{24}(\omega_5, A) - P_e^{64}(\omega_5, A)\right) + \left(P_e^{24}(\omega_7, B) - P_e^{64}(\omega_7, B)\right) + \left(P_e^{24}(C) - P_e^{64}(C)\right)\frac{u(\$3)}{u(\$10)} \geq 0 \tag{37}$$

Both of these are passed at the 1% significance level for the attentive sample, while the difference is insignificant for the inattentive sample.

Altogether, the revealed attention strategies in the attentive sample are largely consistent with the existence of a rationalizing cost function. Under the specification considered in the experimental task, subjects do not exhibit non-monotonicity in choice probabilities and do not show patterns of consideration sets. This does not contradict the model given the unknown preferences of the subject pool.

7.2 Varying the Payoff Structure

The second treatment varies the payoff structures while keeping the cost of information fixed. The experiment is conducted on a separate sample of 114 subjects. Similar to the previous treatment subjects face 30 rounds in one session, and for each of them there are 24 geometric shapes identical to the intermediate cost treatment from the previous section. These reward schemes are presented in figure 12.

	Action A	Action B	Action C
five-sided	**$10**	**$0**	**$3**
six-sided	**$0**	**$0**	**$3**
seven-sided	**$0**	**$10**	**$3**

Reward scheme I

	Action A	Action B	Action C
five-sided	**$10**	**$0**	**$3**
six-sided	**$0**	**$0**	**$3**
seven-sided	**$5**	**$10**	**$3**

Reward scheme II

Figure 12: The two reward schemes

The only difference between reward scheme I and reward scheme II is that the payoff of Action A when the seven-sided shapes are most common is increased to $5 from $0 under reward scheme II. This is the state in which Action B is the highest-paying, hence, the increase in the payoff of Action A narrows the gap between the highest- and lowest-paying action when the seven-sided shapes are most common. In principle, this lowers the incentives to acquire information.

Similar to the previous treatment, subjects are only eligible for final payment if they perform above chance and select the highest-paying action at least 15 of 30 rounds. The reward schemes are presented in blocks of 15 rounds.

The aggregate response time patterns are similar to the previous treatment, many of the subjects clicking through the experiment. In fact, out of the 114 subjects 25 spend less than 2 minutes on all 30 rounds combined. The distribution of the logarithm of response times per round shows substantial heterogeneity, with many of the subjects spending only around one second on a large fraction of the rounds.

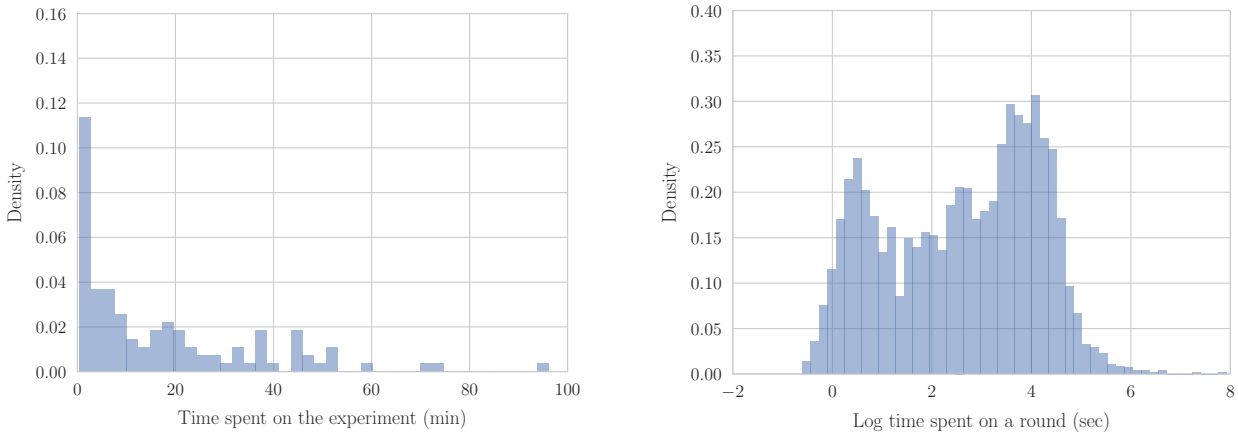


Figure 13: Distribution of time spent on the experiment and time spent on rounds

The correlation between time spent on the instructions and time spent on the experiment is positive again, just as the correlation between the number of highest-paying actions selected and the time spent on the

experiment. Setting the same 2-minute threshold to separate the attentive and inattentive samples, figure 14 depicts the evolution of median response times and the probability of selecting the highest-paying action across rounds. The change in performance is not significantly different from zero for either of the two groups.

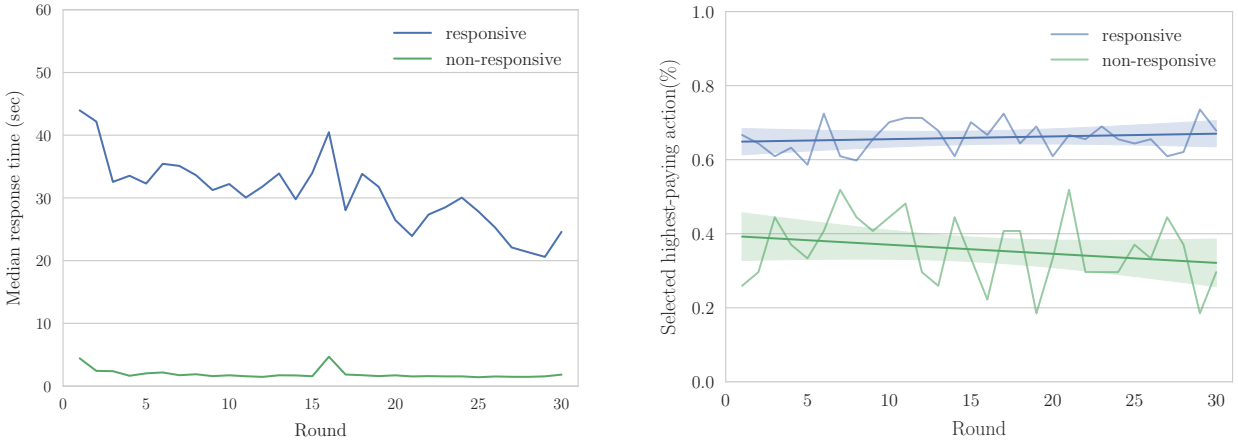


Figure 14: Response times over rounds and probability of matching the state

The slight bump in response times right after round 15 indicates that subjects register the change in the reward scheme.

The average final payment in the attentive sample is $4.48, while in the inattentive sample it is $0.59. Average time spent on the experiment is 22.40 minutes (31.17 with instructions) in the attentive and 1.07 minutes (4.61 with instructions) in the inattentive sample. These result in an hourly wage between $8.72 and $12.13 in the attentive sample.

Increasing the payoff of Action A has a spillover effect on the choice probability of Action C. Conditional on the six-sided shapes being the most common—when Action C is the highest-paying one—the probability of selecting C increases by 2 percentage point. At the current sample size, this change is insignificant.

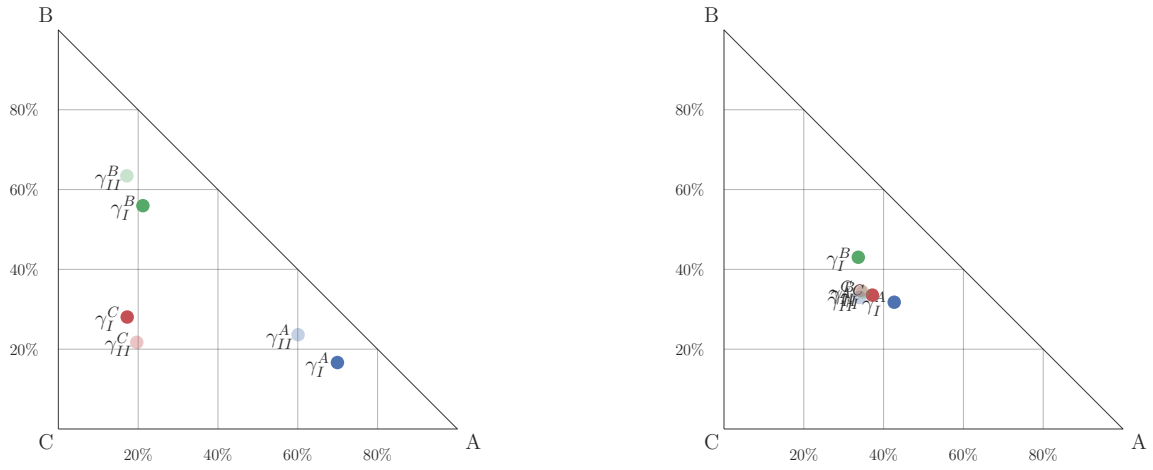


Figure 15: Revealed posteriors – colors fade when switching to reward scheme II

Figure 15 plots the revealed posteriors in the probability simplex under the two reward schemes for both

attentive and inattentive samples. The attention strategies under reward scheme I and II tell the following story. While Action A becomes more attractive subjects are less informed when choosing it—the revealed posterior at Action A corresponding to reward scheme II is closer to the center of the simplex relative to reward scheme I. This is not the case for Action B and Action C, where the revealed posteriors indicate that subjects become more confident under reward scheme II when selecting these actions. Therefore, it is not immediate whether or not subjects acquire more or less information under reward scheme II. We can use the revealed posteriors and the unconditional choice probabilities to give an estimate in the change of acquired information by computing the corresponding values of mutual information. These indicate that the acquired information reduces by 2% due to the increase in the payoff of Action A.

For reward scheme I the NIAS conditions are identical to the previous treatment. For reward scheme II, for Action A they are given by,

$$P_e(\omega_5, A) - P_e(\omega_7, A)\left(1 - \frac{u(\$5)}{u(\$10)}\right) \geq 0; \tag{38}$$

$$P_e(\omega_5, A)u(\$10) + P_e(\omega_7, A)u(\$5) - \left(P_e(\omega_5, A) + P_e(\omega_6, A) + P_e(\omega_7, A)\right)u(\$3) \geq 0. \tag{39}$$

Altogether, there are 12 NIAS tests and all of them are passed at the 1% significance level for the attentive sample while 5 of these are passed at the 1% significance level for the inattentive sample. Details can be found in appendix D.

The NIAC condition simplifies to a single inequality under the treatment varying the reward schemes,

$$P_e^{II}(\omega_7, A) - P_e^{I}(\omega_7, A) \geq 0. \tag{40}$$

The NIAC test is passed on the 1% for the attentive sample and at the 10% level for the inattentive sample. Altogether, the behavior revealed in the data is consistent with the theory of RI.

8 Empirical Application

I demonstrate the estimation strategy proposed in section 6 using a workhorse data set of applied microeconometrics, a revealed preference data set of transportation mode between Sydney and Melbourne, Australia [Greene and Hensher, 1997].[5]

For each individual there are four modes of transportation available, air, bus, car, and train. Table 2 below shows the averages of attributes describing the choice sets conditioning on the chosen alternative. The sample size is $n = 210$.

There are $N = 4$ alternatives that are available for each individual with varying covariates. There are two individual specific covariates, household income (in \$1000) and party size (in number of persons). There are potentially 3 covariates describing the payoff structure, travel time (in minutes), cost of vehicle (in dollars), and wait time at the terminal (in minutes). I base the specification using the BIC from a latent class logit model in order to have a reasonable point of reference for comparing the RI model. The specification entails two latent types and uses two covariates—`vcost, wait`—to describe alternatives, $p_x = 2$, and two covariates—`income, size`—plus an intercept to describe the likelihood of valuation

[5] Attentional frictions are implausible to play a large role in this context, so the application is purely for the purpose of demonstrating the approach.

		Conditional on choice				
		air	bus	car	train	overall
air	travel	124.83	145.70	132.15	137.63	133.71
	vcost	97.57	89.33	76.25	80.40	85.25
	wait	46.53	67.83	65.86	66.54	61.01
bus	travel	618.88	618.83	648.85	626.11	629.46
	vcost	34.28	33.73	33.78	32.27	33.46
	wait	43.07	25.20	46.29	43.86	41.66
car	travel	559.93	671.83	527.37	581.38	573.20
	vcost	23.36	26.80	15.64	21.06	21.00
	wait	0.00	0.00	0.00	0.00	0.00
train	travel	606.90	710.37	638.49	532.67	608.29
	vcost	58.47	62.27	53.59	37.46	51.34
	wait	38.48	36.33	40.27	28.52	35.69
	income	41.72	29.70	42.22	23.06	34.55
	size	1.57	1.33	2.20	1.67	1.74
in sample proportion		0.2762	0.1429	0.2809	0.3000	

Table 2: table: Conditional averages for the Travel Mode Choice dataset

types. A state is then identified with a valuation type of alternative specific covariates, the number of states is $J = 2$ corresponding to the two latent classes. This implies that for each individual the RI model puts positive mass on at most two alternatives exhibiting "consideration sets". The total number of parameters are thus, $(J-1)p_z + Jp_x = 7$.

8.1 Specification – The Shannon Model

In the empirical example I take the benchmark case with information costs being defined proportionate to the mutual information. This implies that the conditional choice probabilities take the form (I replace realization of y with corresponding choice to ease notation),

$$P^*_\mu(a \mid x_A, z; \alpha, \beta) = \sum_k \frac{\exp(\alpha_k' z)}{1 + \sum_{j=1}^{J-1} \exp(\alpha_j' z)} \frac{\exp\left(\beta_k' x_a + \ln P^*_\mu(a \mid x_A, z; \alpha, \beta)\right)}{\sum_{\tilde{a}} \exp\left(\beta_k' x_{\tilde{a}} + \ln P^*_\mu(\tilde{a} \mid x_A, z; \alpha, \beta)\right)}. \tag{41}$$

The optimum[6] in the Shannon model is characterized by,

- If $P^*_\mu(a \mid x_A, z; \alpha, \beta) > 0$:

$$\sum_k \frac{\exp(\alpha_k' z)}{1 + \sum_{j=1}^{J-1} \exp(\alpha_j' z)} \frac{\exp\left(\beta_k' x_a\right)}{\sum_{\tilde{a}} \exp\left(\beta_k' x_{\tilde{a}} + \ln P^*_\mu(\tilde{a} \mid x_A, z; \alpha, \beta)\right)} = 1;$$

[6]Efficient computation of the optimal information structure can be obtained via the Blahut-Arimoto algorithm, which is detail in appendix E.

- If $P_\mu^*(a \mid x_A, z, \alpha, \beta) = 0$:

$$\sum_k \frac{\exp(\alpha_k' z)}{1 + \sum_{j=1}^{J-1} \exp(\alpha_j' z)} \frac{\exp\left(\beta_k' x_a\right)}{\sum_{\tilde{a}} \exp\left(\beta_k' x_{\tilde{a}} + \ln P_\mu^*(\tilde{a} \mid x_A, z; \alpha, \beta)\right)} \leq 1.$$

The gradients with respect to the parameters for the estimation, and with respect to covariates for the computation of marginal effects can be obtained via an application of the implicit function theorem. Computations are detailed in appendix F.

8.2 Empirical Results

We use the estimates of a latent class logit model with the same specification to compare elasticity estimates. For the SMC algorithm I use 10,000 particle draws, 100 tempering steps and 4 draws from a Metropolis-Hastings algorithm with flat prior for each particle at each tempering step to mutate the particles to avoid particle impoverishment. The particles are drawn from initial estimates and the correction and mutation steps are corrected for this.

Figure 16 shows the quasi-posterior of the parameters based on sampling through the adaptive SMC algorithm. To obtain the parameter estimates optimizing the GMM criterion function we take the mode of the densities.

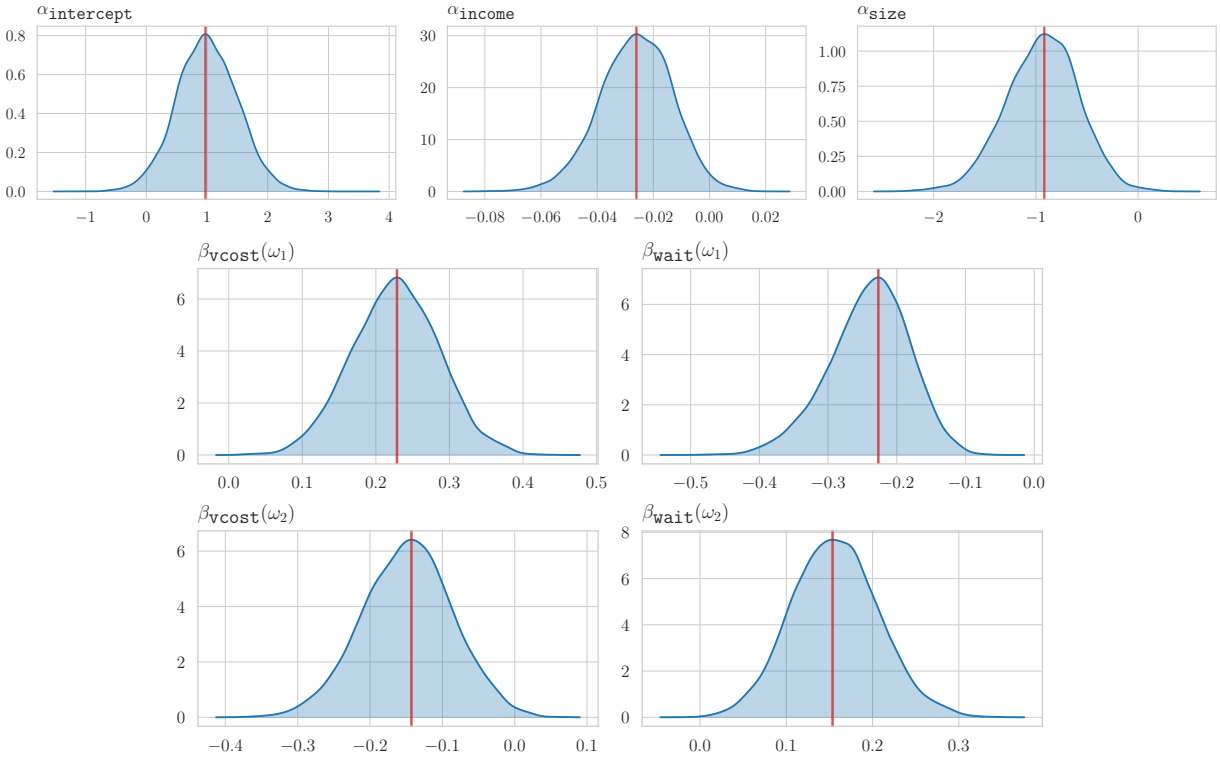


Figure 16: Weighted particle means and the quasi-Bayesian posterior density

I report estimates from the latent class logit model and the RI model with Shannon cost in appendix G.1. These use the same covariates, however, the parameter estimates themselves can not be meaningfully compared between the two models. Instead, table 3 provides the predicted marginal effects on choice

probabilities with changes in the covariates. These are computed using the individual level RI model for each observation and then averaging over the resulting marginal effects. We take different magnitudes of changes so that effects are meaningfully comparable, that is, an increase of $10,000 in yearly household income, 1 person in party size, $10 in cost and 30 minutes in waiting time.

	Latent Class Logit				Rational Inattention			
	$P(\texttt{air})$	$P(\texttt{bus})$	$P(\texttt{car})$	$P(\texttt{train})$	$P^*(\texttt{air})$	$P^*(\texttt{bus})$	$P^*(\texttt{car})$	$P^*(\texttt{train})$
income	0.008	-0.023	0.014	0.001	0.004	-0.027	0.017	0.006
size	0.019	-0.055	0.034	0.001	0.015	-0.095	0.059	0.021
$\texttt{vcost}^{\texttt{air}}$	0.036	0.004	-0.026	-0.013	-0.003	0.001	0.002	0.000
$\texttt{wait}^{\texttt{air}}$	-0.166	-0.002	0.110	0.058	0.010	-0.002	-0.007	-0.001
$\texttt{vcost}^{\texttt{bus}}$	0.004	-0.017	0.004	0.009	0.001	-0.002	0.001	0.001
$\texttt{wait}^{\texttt{bus}}$	-0.002	-0.001	0.012	-0.009	-0.002	0.007	-0.002	-0.003
$\texttt{vcost}^{\texttt{car}}$	-0.026	0.004	0.059	-0.038	0.002	0.001	-0.003	0.000
$\texttt{wait}^{\texttt{car}}$	0.110	0.012	-0.284	0.162	-0.007	-0.002	0.009	-0.000
$\texttt{vcost}^{\texttt{train}}$	-0.013	0.009	-0.038	0.042	0.000	0.001	0.000	-0.001
$\texttt{wait}^{\texttt{train}}$	0.058	-0.009	0.162	-0.211	-0.001	-0.003	-0.000	0.004

Table 3: Marginal effects for the latent class logit and RI models

The rational inattention model predicts lower marginal effects for the alternative specific covariates—close to zero—implying that the choice probabilities are largely defined by the demographic covariates. Changes in the alternative specific characteristics do not lead to substantial changes in choice probabilities. The own-price elasticity is negative for each alternative under the RI model and is only negative for the bus under the latent class logit model. The sign of marginal effects coincide under the two models for the demographic covariates with the RI model predicting larger effects.

The main implication of attentional frictions then is that choices are more rigid in alternative specific changes under attentional constraints and largely driven by individual specific covariates defining valuation types. This is due to the costly information acquisition about types—demographic covariates have a larger impact by swaying the initial probabilities over valuation types, information about which is costly to refine.

9 Conclusion

In this paper I analyze patterns in demand arising under constraints on attention. I identify a novel pattern of complementarity whereby substitute goods appear as complements in observed behavior due to limits on attention. This is orthogonal to the traditional notion of complementarity stemming from preferences. I study empirical inference for the rational inattention model from observational data. For this end, I pose conditions for identification from standard stochastic choice data and employ an estimation strategy that fits peculiar features of the model, such as the formation of consideration sets. I conduct an experiment to study how attention strategies are adjusted with changes in incentives. Finally, I demonstrate the estimation strategy on a dataset of travel mode choice.

References

Jason Abaluck and Abi Adams. What Do Consumers Consider Before They Choose? Identification from Asymmetric Demand Responses. 2018.

Roy Allen and John Rehbeck. Complementarity in Perturbed Utility Models. 2017.

Sandro Ambuehl, Axel Ockenfels, Colin Stewart, et al. Attention and selection effects. 2018.

Suguru Arimoto. An algorithm for computing the capacity of arbitrary discrete memoryless channels. *IEEE Transactions on Information Theory*, 18(1):14–20, 1972.

Steve Berry, Ahmed Khwaja, Vineet Kumar, Andres Musalem, Kenneth C Wilbur, Greg Allenby, Bharat Anand, Pradeep Chintagunta, W Michael Hanemann, Przemek Jeziorski, et al. Structural models of complementary choices. *Marketing Letters*, 25(3):245–256, 2014.

Richard Blahut. Computation of channel capacity and rate-distortion functions. *IEEE transactions on Information Theory*, 18(4):460–473, 1972.

Pedro Bordalo, Nicola Gennaioli, and Andrei Shleifer. Salience theory of choice under risk. *The Quarterly journal of economics*, 127(3):1243–1285, 2012.

Pedro Bordalo, Nicola Gennaioli, and Andrei Shleifer. Salience and consumer choice. *Journal of Political Economy*, 121(5):803–843, 2013.

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, February 2015.

Andrew Caplin, John Leahy, and Filip Matějka. Social learning and selective attention. 2015.

Andrew Caplin, Mark Dean, and John Leahy. Rationally inattentive behavior: Characterizing and generalizing shannon entropy. 2017.

Andrew Caplin, Dániel Csaba, John Leahy, and Oded Nov. Rational Inattention, Competitive Supply, and Psychometrics. *Working Paper, National Bureau of Economic Research*, 2018a.

Andrew Caplin, Mark Dean, and John Leahy. Rational Inattention, Optimal Consideration Sets and Stochastic Choice. *forthcoming in the Review of Economic Studies*, 2018b.

Guillaume Carlier and Alfred Galichon. Exponential convergence for a convexifying equation. *ESAIM: Control, Optimisation and Calculus of Variations*, 18(3):611–620, 2012.

Gary Chamberlain. Asymptotic efficiency in estimation with conditional moment restrictions. *Journal of Econometrics*, 34(3):305–334, 1987.

Christopher Chambers, Ce Liu, and John Rehbeck. Costly Information Acquisition. 2018.

Xiaohong Chen, Victor Chernozhukov, Sokbae Lee, and Whitney K Newey. Local identification of nonparametric and semiparametric models. *Econometrica*, 82(2):785–809, 2014.

Xiaohong Chen, Timothy Christensen, and Elie T. Tamer. Monte carlo confidence sets for identified sets. *forthcoming in Econometrica*, 2018.

Anton Cheremukhin, Anna Popova, and Antonella Tutino. Experimental Evidence on Rational Inattention. Working Paper 1112, Federal Reserve Bank of Dallas, December 2011.

Victor Chernozhukov and Han Hong. An MCMC approach to classical estimation. *Journal of Econometrics*, 115(2):293–346, 2003.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience, 2006.

Imre Csiszár and Gábor Tusnády. Information geometry and alternating minimization procedures. *Statistics and decisions*, 1:205–237, 1984.

Henrique de Oliveira, Tommaso Denti, Maximilian Mihm, and Kemal Ozbek. Rationally inattentive preferences and hidden information costs. *Theoretical Economics*, 12(2):621–654, 2017.

Mark Dean and Nathaniel Neligh. Experimental tests of rational inattention. 2017.

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

Tommaso Denti. Posterior-separable cost of information. 2018.

Ambuj Dewan and Nathaniel Neligh. Estimating information cost functions in models of rational inattention. 2017.

Mogens Fosgerau, Emerson Melo, Matthew Shum, and André de Palma. Discrete Choice and Rational Inattention: A General Equivalence Result. 2017.

Xavier Gabaix. A sparsity-based model of bounded rationality. *The Quarterly Journal of Economics*, 129 (4):1661–1710, 2014.

Matthew Gentzkow. Valuing new goods in a model with complementarity: Online newspapers. *American Economic Review*, 97(3):713–744, 2007.

William H. Greene and David Hensher. Multinomial logit and discrete choice models. *W.H. Greene, LIMDEP*, 7, 1997.

Faruk Gul and Wolfgang Pesendorfer. Random expected utility. *Econometrica*, 74(1):121–146, 2006.

Benjamin Handel and Joshua Schwartzstein. Frictions or mental gaps: What's behind the information we (don't) use and when do we care? *Journal of Economic Perspectives*, 32(1):155–78, February 2018. doi: 10.1257/jep.32.1.155.

Benjamin Hébert and Michael Woodford. Rational Inattention and Sequential Information Sampling. 2017.

Edward Herbst and Frank Schorfheide. Sequential monte carlo sampling for dsge models. *Journal of Applied Econometrics*, 29(7):1073–1098, 2014.

Joonhwi Joo. Quantity-surcharged larger package sales as rationally inattentive consumers' choice. 2018.

Botond Kőszegi and Adam Szeidl. A model of focusing in economic choice. *The Quarterly journal of economics*, 128(1):53–104, 2012.

Bartosz Adam Maćkowiak, Filip Matějka, and Mirko Wiederholt. Survey: Rational inattention, a disciplined behavioral model. Technical report, CEPR Discussion Papers, 2018.

Paola Manzini, Marco Mariotti, and Levent Ülkü. Stochastic complementarity. *The Economic Journal*, March 2018. doi: 10.1111/ecoj.12601.

Filip Matějka and Alisdair McKay. Rational Inattention to Discrete Choices: A New Foundation for the Multinomial Logit Model. *American Economic Review*, 105(1):272–98, 2015.

Stephen Morris and Ming Yang. Coordination and continuous choice. 2016.

Whitney K. Newey. Efficient instrumental variables estimation of nonlinear models. *Econometrica*, pages 809–837, 1990.

Ralph Tyrrell Rockafellar and Roger J-B. Wets. *Variational analysis*, volume 317. Springer Science & Business Media, 2009.

John Rust. Optimal replacement of gmc bus engines: An empirical model of harold zurcher. *Econometrica: Journal of the Econometric Society*, pages 999–1033, 1987.

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

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

Michael Woodford. Inattentive valuation and reference-dependent choice. *Unpublished Manuscript, Columbia University*, 2012.

A Attention Allocation Problem and Duality

Equivalently, one can write problem 7 as minimizing information costs net of utility gains. This makes the application of convex analytic tools more convenient.

$$\min_{P\in\mathcal{P}(\Omega)}\left\{\sum_{\mathcal{S}(P)}\left(\kappa\phi\left(\gamma^s\right)-\max_{a\in A}\left\langle u_a,\gamma^s\right\rangle\right)P_\mu(s)\right\}=$$

$$\min_{P\in\mathcal{P}(\Omega)}\left\{\sum_{\mathcal{S}(P)}\left(\min_{a\in A}\underbrace{\left\{\kappa\phi\left(\gamma^s\right)-\left\langle u_a,\gamma^s\right\rangle\right\}}_{\text{denote by }\phi_a(\cdot)}\right)P_\mu(s)\right\}=\min_{P\in\mathcal{P}(\Omega)}\left\{\sum_{\mathcal{S}(P)}\left(\min_{a\in A}\underbrace{\left\{\phi_a\left(\gamma^s\right)\right\}}_{\text{convex}}\right)P_\mu(s)\right\}$$

Note that ϕ_a[7] is strictly convex, as it is the difference of a strictly convex and a linear function. The lower envelope function appearing in the sum plays a pivotal role, it is worth naming it,

$$\phi_A(\gamma)\colon\Gamma\mapsto\mathbb{R}:=\min_{a\in A}\left\{\,\phi_a\left(\gamma\right)\right\}. \tag{42}$$

ϕ_A is the lower envelope of the class of functions $\{\phi_a : a\in A\}$. It is not a convex function. From this formulation it is straightforward that the problem defined in equation 7 is equivalent to evaluating the convex hull[8] of ϕ_A at μ,

$$\min_{P\in\mathcal{P}(\Omega)}\left\{\sum_{\mathcal{S}(P)}\phi_A\left(\gamma^s\right)P_\mu(s)\right\}. \tag{43}$$

Since the function ϕ is strictly convex there is a unique solution to the problem which we denote by P^*. Since the distribution of posteriors satisfy the Bayesian consistency of averaging to the prior, the optimization problem in the RI model is essentially one with respect to a martingale. It results in tracing the convex hull of ϕ_A. The convex hull of ϕ_A can be expressed through double application of the Legendre-Fenchel transform—that is, the convex conjugate.

The value of equation 43 is equal to,

$$\phi_A^{**}(\mu). \tag{44}$$

(Where * denotes the convex conjugate operator.) This in fact equals to,

$$\left(\max_{a\in A}\{\phi_a^*\}\right)^*(\mu). \tag{45}$$

To see the equivalence take $h\in\mathbb{R}^{|\Omega|}$ and consider

$$\phi_A^*(h)=\sup_{\gamma\in\Gamma}\{\langle h,\gamma\rangle-\underbrace{\phi_A(\gamma)}_{\min_a\phi_a(\gamma)}\}=\sup_{\gamma\in\Gamma}\max_{a\in A}\{\langle h,\gamma\rangle-\phi_a(\gamma)\}=\max_{a\in A}\{\phi_a^*(h)\}$$

[7] We drop κ from the notation of $\phi_A(\cdot)$ as we can always renormalize the payoffs so that $\kappa=1$.
[8] The relationship to viscosity equations is discussed in Carlier and Galichon [2012].

The convex conjugates, ϕ_a^*, take a convenient form which simplifies computation—we only need to compute ϕ^* once.

$$\phi_a^*(h) = \sup_{\gamma\in\Gamma}\{\langle h,\gamma\rangle - \phi_a(\gamma)\} = \sup_{\gamma\in\Gamma}\{\langle h,\gamma\rangle + \langle u_a,\gamma\rangle - \phi(\gamma)\} = \phi^*(h+u_a)$$

With the above derivation we can write the original problem as

$$\left(\max_{a\in A}\{(\phi_a)^*\}\right)^*(\mu) = \left(\max_{a\in A}\{\phi^*(h+u_a)\}\right)^*(\mu) =$$

$$\sup_h\left\{\langle h,\mu\rangle - \max_{a\in A}\{\phi^*(h+u_a)\}\right\} = \sup_h\left\{\min_{a\in A}\{\langle h,\mu\rangle - \phi^*(h+u_a)\}\right\}. \tag{46}$$

Remark 1.
An application of the min-max inequality would state that the DM can not do worse than choosing the expected utility maximizing action with respect to the prior. That is,

$$\begin{aligned}\sup_h\left\{\min_{a\in A}\{\langle h,\mu\rangle - \phi^*(h+u_a)\}\right\} &\le \min_{a\in A}\left\{\sup_h\{\langle h,\mu\rangle - \phi^*(h+u_a)\}\right\}\\ &= \min_{a\in A}\{\phi_a(\mu)\}.\end{aligned} \tag{47}$$

A.1 Proof of Proposition 1

Suppose otherwise. Leaving all other signals intact define a new signal n based on γ^s and γ^r such that the induced posterior is,

$$\begin{aligned}\gamma^n &= \frac{1}{P_\mu^*(s)+P_\mu^*(r)}\left(P_\mu^*(s)\gamma^s + P_\mu^*(r)\gamma^r\right),\\ \tilde{P}_\mu(n) &= P_\mu^*(s) + P_\mu^*(r).\end{aligned} \tag{48}$$

Since expected utility is linear the value of the information structure $\tilde{P}$ is equivalent to that of P^*. However, by strict convexity of ϕ its cost is lower,

$$\tilde{P}_\mu(n)\phi(\gamma^n) < P_\mu^*(s)\phi(\gamma^s) + P_\mu^*(r)\phi(\gamma^r). \tag{49}$$

A contradiction. □

B Experimental Instructions

The experimental instructions and tasks are available at:

- http://distinguish-shapes-2.herokuapp.com/welcome for the treatment varying the gain-to-cost ratio of information acquisition;
- http://distinguish-shapes-3.herokuapp.com/welcome for the treatment varying the payoff structure.

C Generating the Polygons

The geometric shapes are generated by the following algorithm.

- N: the total number of polygons generated in each round
- $n = 5$: the number of different polygons to be present on the page
  - specify their types – e.g. [5-gon, 6-gon, 7-gon, 8-gon, 9-gon]
- $m = 3$: the number of non-decoy polygons which the subject actually has to consider
  - specify their types – e.g. [6-gon, 7-gon, 8-gon]
- $\Delta^{\max} \leq \frac{N}{n}(n-m)$: the difference between the highest number and lowest number within the non-decoy class. This difference is assigned randomly to one of the polygons in the non-decoy class.
- $\Delta^{\text{mid}} < \Delta^{\max}$: the difference between the middle number and the lowest number within the non-decoy class. This difference is assigned randomly to one of the remaining two polygons (the ones to which $\Delta^{\max}$ wasn't assigned) in the non-decoy class.
- $\Delta^{\text{mud}} \leq \frac{1}{m}\left(\frac{N}{n}(n-m) - \Delta^{\max} - \Delta^{\text{mid}}\right)$: shifts all the counts uniformly in the decoy class with a number in the range $[-\Delta^{\text{mud}}, \Delta^{\text{mud}}]$ picked randomly. Call the realization of this random number δ^{mud}.
- With these specified parameters (and the random mudding) make the $(n-m)$ decoy polygon(s) so that everything sums to N and the decoys are "as even as possible". That is look for the closest integer assignment to $\frac{N}{n} - \frac{\Delta^{\max}+\Delta^{\text{mid}}+m\delta^{\text{mud}}}{n-m}$.
- Degrees for rotating the polygons randomly.

Example:

- $N = 32$
- $n = 5$ [5-gon, 6-gon, 7-gon, 8-gon, 9-gon]
- $m = 3$ [6-gon, 7-gon, 8-gon] (non-decoys)

$$[\overbrace{7}^{\text{5-gon}},\ \underbrace{6,\ 6,\ 6}_{\text{6-gon, 7-gon, 8-gon}},\ \overbrace{7}^{\text{9-gon}}]$$

- $\Delta^{\max} = 2 \leq \frac{N}{n}(n-m) = 12.8$. Then randomly assign $\Delta^{\max} = 2$ to one of the non-decoy polygons.

$$[\overbrace{7}^{\text{5-gon}},\ \underbrace{6,\ 6,\ 8}_{\text{6-gon, 7-gon, 8-gon}},\ \overbrace{7}^{\text{9-gon}}]$$

- $\Delta^{\text{mid}} = 1 < \Delta^{\max} = 2$. Then randomly assign $\Delta^{\text{mid}} = 1$ to one of the remaining non-decoy polygons.

$$[\overbrace{7}^{\text{5-gon}},\ \underbrace{7,\ 6,\ 8}_{\text{6-gon, 7-gon, 8-gon}},\ \overbrace{7}^{\text{9-gon}}]$$

- $\Delta^{\text{mud}} = 3 \leq \frac{1}{m}\left(\frac{N}{n}(n-m) - \Delta^{\max}\right) = 5$. Pick a random integer between [-3, 3]. Suppose the realization is $\delta^{\text{mud}} = 1$. Shift all the non-decoys.

$$[\overbrace{7}^{\text{5-gon}},\ \underbrace{8,\ 7,\ 9}_{\text{6-gon, 7-gon, 8-gon}},\ \overbrace{7}^{\text{9-gon}}]$$

  Now change the decoys, so that they are closest to $\frac{N}{n} - \frac{\Delta^{\max}+\Delta^{\text{mid}}+m\delta^{\text{mud}}}{n-m} = 4$. So [4, 4] will do in this case. Now the final numbers are

$$[\overbrace{4}^{\text{5-gon}},\ \underbrace{8,\ 7,\ 9}_{\text{6-gon, 7-gon, 8-gon}},\ \overbrace{4}^{\text{9-gon}}]$$

  They sum to 32 as expected.

In case the total number of shapes appearing on the screen is 8, the numbers of the three non-decoy shapes are 1, 2, 3 and these values are assigned randomly. There is always only one of each decoy shapes in this case.

D Experimental Data Analysis

D.1 Performance Change

The following tables report results from regressing the round rank on the binary variable of selecting the highest-paying action—effectively mapping the probability of choosing the highest-paying action across rounds.

Table 4: Performance – Treatment I

	Attentive	Inattentive
Intercept	0.602***	0.346***
	(0.024)	(0.036)
round	0.001	-0.001
	(0.001)	(0.002)
No. observations	3540	750
Adj. R-squared	0.00	-0.00

Clustered standard errors in parentheses.
* $p < .1$, ** $p < .05$, *** $p < .01$

Table 5: Performance – Treatment II

	Attentive	Inattentive
Intercept	0.648***	0.395***
	(0.032)	(0.035)
round	0.001	-0.002
	(0.001)	(0.002)
No. observations	2610	810
Adj. R-squared	-0.00	0.00

Clustered standard errors in parentheses.
* $p < .1$, ** $p < .05$, *** $p < .01$

D.2 NIAS Tests

Treatment I

p-values for the NIAS tests for both attentive and inattentive samples under the treatment varying the gain-to-cost ratio of information costs.

Table 6: NIAS – Attentive sample

		A	B	C
8	A		0.000	0.000
	B	0.000		0.000
	C	0.000	0.000	
24	A		0.000	0.000
	B	0.000		0.000
	C	0.000	0.000	
64	A		0.000	0.000
	B	0.000		0.000
	C	0.404	0.198	

Table 7: NIAS – Inattentive sample

		A	B	C
8	A		0.039	0.125
	B	0.092		0.057
	C	0.999	0.014	
24	A		0.834	0.357
	B	0.723		0.358
	C	0.888	0.782	
64	A		0.580	0.402
	B	0.939		0.892
	C	0.364	0.968	

Out of the 18 NIAS conditions 16 are passed at the 1% significance level for the attentive sample, while non are passed at the 1% significance level for the inattentive and 2 of the violations are significant at the 5% level.

Treatment II

p-values for the NIAS tests for both attentive and inattentive samples under the treatment varying the reward scheme.

Table 8: Attentive sample

		A	B	C
I	A		0.000	0.000
	B	0.000		0.000
	C	0.000	0.000	
II	A		0.000	0.000
	B	0.000		0.000
	C	0.000	0.000	

Table 9: Inattentive sample

		A	B	C
I	A		0.008	0.009
	B	0.546		0.266
	C	0.066	0.040	
II	A		0.000	0.000
	B	0.000		0.311
	C	0.999	0.727	

All of the 12 NIAS conditions are passed at the 1% significance level for the attentive sample, while 5 of them are passed at the 1% significance level for the inattentive sample.

E Computation of RI Models

Choice probabilities in the Shannon model can be computed using the Blahut-Arimoto algorithm [Blahut, 1972, Arimoto, 1972]. The algorithm is a special case of the alternating-minimization algorithm due to Csiszár and Tusnády [1984]. We state the Blahut-Arimoto algorithm in the RI context making explicit its connections to alternating minimization—an excellent source for this is Cover and Thomas [2006].

E.1 The Blahut-Arimoto Algorithm

Problem 7 stated,

$$\min_{P\in\mathcal{P}(A|\Omega)} \left\{ \kappa K(P,\mu) - \sum_{s\in\mathcal{S}(P)} V(\gamma^s, A) P_\mu(s) \right\}.$$

The minimizer would not change if we reposed the problem as,

$$\min_{P\in\mathcal{P}(A|\Omega)} \left\{ K(P,\mu) - \frac{1}{\kappa} \sum_{s\in\mathcal{S}(P)} V(\gamma^s, A) P_\mu(s) \right\},$$

which could be interpreted as minimizing the cost of the information structure subject to keeping the expected utility above a certain threshold $\bar{U}$ corresponding to the Lagrange multiplier $1/\kappa$.

The Blahut-Arimoto algorithm starts with reposing the original minimization problem as a double minimization which is possible in the case of the mutual information being the cost function. Keeping the prior fixed an information structure induces a joint distribution over states and actions—we will use the fact that posteriors uniquely correspond to actions—through $P_\omega(\gamma^a) = P_\omega(a)$.

First, we use the fact that,

$$D_{\mathrm{KL}}\Big(P_\omega(a)\mu(\omega)\,\|\,P_\mu(a)\mu(\omega)\Big) = \min_{R\in\Delta(A)} D_{\mathrm{KL}}\Big(P_\omega(a)\mu(\omega)\,\|\,R(a)\mu(\omega)\Big). \tag{50}$$

The the optimizer in the original can be found by solving

$$\begin{aligned}\min_{R\in\Delta(A)}\min_{P\in\mathcal{P}(A|\Omega)} & D_{\mathrm{KL}}\Big(P_\omega(a)\mu(\omega)\,\|\,R(a)\mu(\omega)\Big)\\ \text{subject to} \quad & \sum_{a,\omega} u(a(\omega))P_\omega(a)\mu(\omega) \geq \bar{U}.\end{aligned} \tag{51}$$

Now denote the sets defining the constraints as follows,

$$B := \left\{ P\in\Delta(A\times\Omega) : \sum_a P(a,\omega) = \mu(\omega) \quad \forall\omega, \text{ and } \sum_{a,\omega} P(a,\omega)u(a(\omega)) \geq \bar{U} \right\}; \tag{52}$$

$$C := \Big\{ P\in\Delta(A\times\Omega) : P(a,\omega) = R(a)\mu(\omega) \quad \forall\omega, a \text{ for some } R\in\Delta(A) \Big\}. \tag{53}$$

Both B and C are convex sets in the same finite dimensional Euclidean space. Then problem 51 reduces to,

$$\min_{S\in C}\min_{Q\in B} D_{\mathrm{KL}}\left(Q\,\|\,S\right).$$

The alternating minimization algorithm generates a pair of sequences (Q^n, S^n) in the following manner:

- Start with arbitrary $S^0\in C$. For $n\geq 1$,
- $Q^n = \arg\min_{Q\in B} D_{\mathrm{KL}}\left(Q\,\|\,S^{n-1}\right)$
- $S^n = \arg\min_{S\in C} D_{\mathrm{KL}}\left(Q^n\,\|\,S\right)$

As Csiszár and Tusnády [1984] showed the Kullback-Leibler divergence satisfies certain geometric properties such that,

$$\lim_{n\to\infty} D_{\mathrm{KL}}\left(Q^n\,\|\,S^n\right) = \min_{Q\in B}\min_{S\in C} D_{\mathrm{KL}}\left(Q\,\|\,S\right).$$

The Blahut-Arimoto algorithm is efficient because,

- $S^n(a,\omega) = R^n(a)\mu(\omega)$ is computed by a simple marginal distribution: $R^n(a) = \sum_\omega Q^n(a,\omega)$.
- $Q^n(\,\cdot\,,\omega) = Q^n(\,\cdot\mid\omega)\mu(\omega)$ is computed using only $u(\,\cdot\,,\omega)$ and $R^n(a)$.

F Gradient of the Optimal Probabilities

$P^*_\mu(x_{it}, z_{it}; \alpha, \beta)$ has to satisfy a fixed point constraint for each individual specific choice situation it. We repeat these fixed point constraints for generic choice set and individual characteristics (x_A, z).

For each label $a \in A$ the optimal unconditional choice probabilities satisfy,

$$\left(\sum_{k \in J} \frac{\exp(\alpha_k' z)}{1 + \sum_{j=2}^{J} \exp(\alpha_j' z))} \frac{\exp\left(\beta_k' x_a\right)}{\sum_{\tilde{a}} P^*_\mu(\tilde{a} \mid x_A, z; \alpha, \beta) \exp\left(\beta_k' x_{\tilde{a}}\right)} - 1 \right) P^*_\mu(a \mid x_A, z; \alpha, \beta) = 0. \tag{54}$$

F.1 Gradient with respect to parameters

Keeping the covariates fixed, for any parameter value, θ, the optimal choice probabilities are described by the following implicit function.

$$\begin{aligned} &F \colon \mathbb{R}^{d_Z (J-1) + d_X J} \times \mathbb{R}^N \mapsto \mathbb{R}^N \\ &F(\theta, P_\mu) = \mathbf{0}. \end{aligned} \tag{55}$$

For each θ (and observables (x_A, z)—we suppress this dependence now) there is a unique vector of probabilities $P_\mu(\theta \mid x_A, z)$ that satisfies the fixed point constraint.

$$F(\theta, P^*_\mu(\theta)) = \mathbf{0}$$

We use the implicit function theorem to characterize the gradient of $\nabla P^*_\mu(\theta)$.

$$\begin{aligned} (D_\theta F) + (D_{P_\mu} F)(D_\theta P_\mu) &= \mathbf{0} \\ [D_\theta P^*_\mu]_{N \times d} &= \left[(D_{P_\mu} F)^{-1} \right]_{N \times N} [-(D_\theta F)]_{N \times d} \end{aligned} \tag{56}$$

Derivatives:

$$\begin{aligned} D_{P_\mu(a)} F(a) = &\left(\sum_{k \in J} \frac{\exp(\alpha_k' z)}{1 + \sum_{j=1}^{J-1} \exp(\alpha_j' z)} \frac{\exp\left(\beta_k' x_a\right)}{\sum_{\tilde{a}} \exp\left(\beta_k' x_{\tilde{a}}\right) P^*_\mu(\tilde{a})} - 1 \right) - \\ &\sum_{k \in J} \frac{\exp(\alpha_k' z)}{1 + \sum_{j=1}^{J-1} \exp(\alpha_j' z)} \left(\frac{\exp\left(\beta_k' x_a\right)}{\sum_{\tilde{a}} \exp\left(\beta_k' x_{\tilde{a}}\right) P^*_\mu(\tilde{a})} \right)^2 P^*_\mu(a) \end{aligned} \tag{57}$$

$$D_{P_\mu(b)} F(a) = - \sum_{k \in J} \frac{\exp(\alpha_k' z)}{1 + \sum_{j=1}^{J-1} \exp(\alpha_j' z)} \frac{\exp\left(\beta_k' x_a\right) \exp\left(\beta_k' x_b\right)}{\left(\sum_{\tilde{a}} \exp\left(\beta_k' x_{\tilde{a}}\right) P^*_\mu(\tilde{a}) \right)^2} P^*_\mu(a) \tag{58}$$

$$
\begin{aligned}
D_{\alpha_k} F(a) =& \frac{\exp(\alpha_k' z) z'}{\sum_j \exp(\alpha_j' z)} \frac{\exp\left(\beta_k' x_a\right) P_\mu^*(a)}{\sum_{\tilde{a}} \exp\left(\beta_k' x_{\tilde{a}}\right) P_\mu^*(\tilde{a})} - \\
& \sum_l \frac{\exp(\alpha_l' z) \exp(\alpha_k' z) z'}{\left(\sum_j \exp(\alpha_j' z)\right)^2} \frac{\exp\left(\beta_l' x_a\right) P_\mu^*(a)}{\sum_{\tilde{a}} \exp\left(\beta_l' x_{\tilde{a}}\right) P_\mu^*(\tilde{a})} \\
=& \mu(\omega_k) z' P^*(a \mid \omega_k) - \sum_l \mu(\omega_l) \mu(\omega_k) z' P^*(a \mid \omega_l) \\
=& \mu(\omega_k) \left[P^*(a \mid \omega_k) - P_\mu^*(a)\right] z'
\end{aligned} \tag{59}
$$

$$
\begin{aligned}
D_{\beta_k} F(a) =& \frac{\exp(\alpha_k' z)}{\sum_j \exp(\alpha_j' z)} \left(\frac{\exp\left(\beta_k' x_a\right) P_\mu^*(a) x_a'}{\sum_{\tilde{a}} \exp\left(\beta_k' x_{\tilde{a}}\right) P_\mu^*(\tilde{a})} - \sum_{\tilde{a}} \frac{\exp\left(\beta_k' x_a\right) P^*(a) \exp\left(\beta_k' x_{\tilde{a}}\right) P_\mu^*(\tilde{a}) x_{\tilde{a}}'}{\left(\sum_{\tilde{a}} \exp\left(\beta_k' x_{\tilde{a}}\right) P_\mu^*(\tilde{a})\right)^2} \right) \\
=& \mu(\omega_k) \left(P^*(a \mid \omega_k) x_a' - \sum_{\tilde{a}} P^*(a \mid \omega_k) P^*(\tilde{a} \mid \omega_k) x_{\tilde{a}}' \right) \\
=& \mu(\omega_k) P^*(a \mid \omega_k) \left(x_a' - \sum_{\tilde{a}} P^*(\tilde{a} \mid \omega_k) x_{\tilde{a}}' \right)
\end{aligned} \tag{60}
$$

F.2 Gradient with respect to covariates

Keeping the parameters fixed, for any covariate value, (z, x_A), the optimal choice probabilities are described by the following implicit function.

$$
\begin{aligned}
&G \colon \mathbb{R}^{d_Z + d_X N} \times \mathbb{R}^N \mapsto \mathbb{R}^N \\
&G(z, x_A, P_\mu) = \mathbf{0}.
\end{aligned} \tag{61}
$$

For each (z, x_A) there is a unique vector of probabilities $P_\mu(z, x_A \mid \theta)$ that satisfies the fixed point constraint.

$$
G(z, x_A, P_\mu^*(z, x_A)) = \mathbf{0}
$$

We use the implicit function theorem to characterize the gradient of $\nabla P_\mu^*(z, x_A)$. Denote the vector of all covariates $zx := (z, x_A)$.

$$
\begin{aligned}
(D_{zx} G) + (D_{P_\mu} G)(D_{zx} P_\mu) &= \mathbf{0} \\
[D_{zx} P_\mu^*]_{N \times d} &= \left[(D_{P_\mu} G)^{-1}\right]_{N \times N} \left[-(D_{zx} G)\right]_{N \times d}
\end{aligned} \tag{62}
$$

Derivatives:

$$D_{P_\mu(a)}G(a) = \left(\sum_{k\in J} \frac{\exp(\alpha_k' z)}{1+\sum_{j=1}^{J-1}\exp(\alpha_j' z)} \frac{\exp\left(\beta_k' x_a\right)}{\sum_{\tilde{a}} \exp\left(\beta_k' x_{\tilde{a}}\right) P_\mu^*(\tilde{a})} - 1 \right) -$$

$$\sum_{k\in J} \frac{\exp(\alpha_k' z)}{1+\sum_{j=1}^{J-1}\exp(\alpha_j' z)} \left(\frac{\exp\left(\beta_k' x_a\right)}{\sum_{\tilde{a}} \exp\left(\beta_k' x_{\tilde{a}}\right) P_\mu^*(\tilde{a})} \right)^2 P_\mu^*(a) \tag{63}$$

$$D_{P_\mu(b)}G(a) = -\sum_{k\in J} \frac{\exp(\alpha_k' z)}{1+\sum_{j=1}^{J-1}\exp(\alpha_j' z)} \frac{\exp\left(\beta_k' x_a\right)\exp\left(\beta_k' x_b\right)}{\left(\sum_{\tilde{a}} \exp\left(\beta_k' x_{\tilde{a}}\right) P_\mu^*(\tilde{a})\right)^2} P_\mu^*(a) \tag{64}$$

$$\begin{aligned} D_z G(a) =& \sum_{k\in J} \left(\frac{\exp(\alpha_k' z)}{\sum_j \exp(\alpha_j' z)} \alpha_k - \frac{\exp(\alpha_k' z)}{\left(\sum_j \exp(\alpha_j' z)\right)^2} \sum_{l\in J} \exp(\alpha_l' z)\alpha_l \right) \frac{\exp\left(\beta_k' x_a\right) P_\mu^*(a)}{\sum_{\tilde{a}} \exp\left(\beta_k' x_{\tilde{a}}\right) P_\mu^*(\tilde{a})} \\ =& \sum_{k\in J} \left(\mu(\omega_k)\alpha_k - \mu(\omega_k) \sum_l \mu(\omega_l)\alpha_l \right) P^*(a \mid \omega_k) \\ =& \sum_{k\in J} \mu(\omega_k) P^*(a \mid \omega_k) \left(\alpha_k - \bar{\alpha}\right) \end{aligned} \tag{65}$$

$$\begin{aligned} D_{x_a} F(a) =& \sum_{k\in J} \frac{\exp(\alpha_k' z)}{\sum_j \exp(\alpha_j' z)} \left(\frac{\exp\left(\beta_k' x_a\right) P_\mu^*(a)}{\sum_{\tilde{a}} \exp\left(\beta_k' x_{\tilde{a}}\right) P_\mu^*(\tilde{a})} \beta_k - \frac{\exp\left(\beta_k' x_a\right) P_\mu^*(a)}{\left(\sum_{\tilde{a}} \exp\left(\beta_k' x_{\tilde{a}}\right) P_\mu^*(\tilde{a})\right)^2} \exp\left(\beta_k' x_a\right) P_\mu^*(a)\beta_k \right) \\ =& \sum_{k\in J} \mu(\omega_k) \left(P^*(a \mid \omega_k) \left(1 - P^*(a \mid \omega_k)\right) \right) \beta_k \end{aligned} \tag{66}$$

$$\begin{aligned} D_{x_b} F(a) =& \sum_{k\in J} \frac{\exp(\alpha_k' z)}{\sum_j \exp(\alpha_j' z)} \left(-\frac{\exp\left(\beta_k' x_a\right) P_\mu^*(a)}{\left(\sum_{\tilde{a}} \exp\left(\beta_k' x_{\tilde{a}}\right) P_\mu^*(\tilde{a})\right)^2} \exp\left(\beta_k' x_b\right) P_\mu^*(b)\beta_k \right) \\ =& -\sum_{k\in J} \mu(\omega_k) P^*(a \mid \omega_k) P^*(b \mid \omega_k) \beta_k \end{aligned} \tag{67}$$

G Empirical Application

G.1 Estimated parameters

Table 10 compares estimates of the parameters from a latent class logit model and therational inattention model with corresponding specifications using the same covariates. Estimates for the RI model are obtained by sampling from the quasi-posterior corresponding to the GMM criterion. The tables presents the particle-weighted mode and quantiles.

Table 10: Parameter estimates

	Latent Class Logit	Rational Inattention
$\alpha_{\texttt{intercept}}$	1.2763*	0.9737
	(0.6921)	[0.0282, 1.9930]
$\alpha_{\texttt{income}}$	-0.0305**	-0.0260
	(0.0130)	[-0.0525, -0.0023]
$\alpha_{\texttt{size}}$	-0.7418**	-0.9176
	(0.3702)	[-1.6480, -0.2737]
$\beta_{\texttt{vcost}}(\omega_1)$	0.0649***	0.2287
	(0.01392)	[0.1118, 0.3435]
$\beta_{\texttt{wait}}(\omega_1)$	-0.0855***	-0.2268
	(0.0169)	[-0.3658, -0.1371]
$\beta_{\texttt{vcost}}(\omega_2)$	-0.0506***	-0.1428
	(0.0163)	[-0.2727, -0.0279]
$\beta_{\texttt{wait}}(\omega_2)$	0.0322**	0.1536
	(0.0145)	[0.0604, 0.2596]
Log Likelihood	-237.5	–
No. observations	210	210

Note: * $p < .1$, ** $p < .05$, *** $p < .01$
Standard errors in parentheses, 95% (particle weighted) confidence intervals in brackets.