Tsakas, Elias

Article

Robust scoring rules

Theoretical Economics

Provided in Cooperation with:
The Econometric Society

Suggested Citation: Tsakas, Elias (2020) : Robust scoring rules, Theoretical Economics, ISSN 1555-7561, The Econometric Society, New Haven, CT, Vol. 15, Iss. 3, pp. 955-987, https://doi.org/10.3982/TE3557

This Version is available at:
https://hdl.handle.net/10419/253469

Theoretical Economics 15 (2020), 955–987 1555-7561/20200955

Robust scoring rules

Elias Tsakas
Department of Economics (AE1), Maastricht University

Is it possible to guarantee that the mere exposure of a subject to a belief elicitation task will not affect the very same beliefs that we are trying to elicit? In this paper, we introduce mechanisms that make it simultaneously strictly dominant for the subject (a) not to acquire any information that could potentially lead to belief updating as a response to the incentives provided by the mechanism itself, and (b) to report his beliefs truthfully. Such mechanisms are called *robust scoring rules*. We prove that robust scoring rules always exist under mild assumptions on the subject's costs for acquiring information. Moreover, every scoring rule can become approximately robust, in the sense that if we scale down the incentives sufficiently, we will approximate with arbitrary precision the beliefs that the subject would have held if he had not been confronted with the belief-elicitation task.

Keywords. Noninvasive belief elicitation, prior beliefs, rational inattention, posterior-separability, Shannon entropy, population beliefs.

JEL classification. C91, D81, D82, D83, D87.

1. Introduction

Subjective beliefs constitute one of the most common latent variables of interest in economics (Manski 2004). Having recognized this, statisticians and economists have developed mechanisms, called *proper scoring rules*, that incentivize the economic agent to reveal his true latent belief by—roughly speaking—rewarding reports that are close to the realized state and punishing reports that are further away from it (Brier 1950, Good 1952). Due to their solid theoretical foundations (i.e., the fact that they are incentive-compatible) together with the overwhelming experimental evidence which suggests that incentives matter (Harrison and Rütstrom 2008, Harrison 2014), proper scoring rules have been extensively used both in applications and in (lab and field) experiments.

One of the main concerns with proper scoring rules is that, by rewarding accurate reports, they provide the agent not only with direct incentives to report truthfully, but

Elias Tsakas: e.tsakas@maastrichtuniversity.nl

This paper previously circulated under the title "Eliciting prior beliefs." I am greatly indebted to Glenn Harrison, Antonio Penta, John Rehbeck, Burkhard Schipper, Peter Wakker, and three anonymous referees for their valuable comments at different stages of this project. I would also like to thank Chris Chambers, Madhav Chandrasekher, Paul Heidhues, Peter Katuščák, Dorothea Kübler, Fabio Maccheroni, Massimo Marinacci, Marcus Pivato, Arno Riedl, Marciano Siniscalchi, Jakub Steiner, Mathias Staudigl, Stefan Terstiege, Nikolas Tsakas, Mark Voorneveld and the audiences in LOFT (Bocconi), BGSE Summer Forum (Pompeu Fabra), Bayesian Crowds (Tinbergen Institute), EEA-ESEM (Cologne), RWTH Aachen University, University of Athens, and University of Bielefeld for helpful comments and fruitful discussions. I am also grateful to Lars Wittrock for his research assistance. Finally, I would like to thank the Economics Department at UC Davis for its hospitality while working on this project.

 https://doi.org/10.3982/TE3557

also with indirect incentives to acquire information before stating their report. Such information acquisition will typically lead to belief updating. Thus, even if the direct effect is strong enough to induce truthful reporting in the end, the reported beliefs might not be the ones that the agent would have held in the absence of the elicitation task. In other words, as Schotter and Trevino (2014, p. 109) eloquently put it,

> *"the very act of belief elicitation may change the beliefs of subjects from their true latent beliefs or the beliefs they would hold (respond to) if those beliefs were not elicited (we might have a type of Heisenberg problem)."*

Hence, our aim is to find mechanisms that provide strong enough incentives to induce truth-telling, but at the same time not so strong to lead to information acquisition.

But why would one care about the beliefs the agent would have held before the elicitation task, rather than the (perhaps more accurate) ones that are typically formed after information has been acquired as a response to the incentives provided by the elicitation task? Take the example of an investigator whose aim is to elicit the distribution of beliefs in a population. For instance, consider a marketing campaign interested in eliciting the average subjective belief in a population of consumers (e.g., about a new product being superior to the existing ones), or a political campaign interested in the median belief in a population of voters (e.g., about a proposed project being successful or about the outcome of the election). Such statistics of the population beliefs can be used as explanatory variables for population behavior and, therefore, they are often crucial for pending strategic decisions by the respective campaign. The bottom line is that the investigator is not interested in learning the true state of nature per se, but rather in finding out what the population believes about the state of nature. Thus, she draws a representative sample from the respective population, she elicits individual beliefs from the sample, and then she uses the empirical frequency to estimate the distribution of the population beliefs. Crucially, the investigator wants the individuals in the sample to report the beliefs they would have held, had the survey not taken place. Otherwise, her estimate of the population beliefs will be biased.

As an alternative example, consider a lab experiment where both beliefs and choices are observed. One of the main assumptions that we—implicitly or explicitly—impose is stationarity of beliefs across the different tasks. For instance, if we want to use the beliefs as an explanatory variable for choice, we would ideally want to make sure that the elicited beliefs are the ones the subject would be holding at the time of his decision, or at least they do not deviate much from those benchmark beliefs.

Thus, the general question addressed in this paper is whether *we can construct noninvasive scoring rules that elicit the beliefs the agent would have held in the absence of our elicitation task*. In other words, we want to *guarantee that the agent will not find it beneficial to acquire any information that could potentially distort his beliefs before stating his report*. Let us stress that we are not aiming at discussing the practical implementation of such mechanisms, nor do we have something significant to say in relation to role of individual characteristics (e.g., risk-preferences) or biases (e.g., failure to do Bayesian updating) that empirically affect the elicited beliefs. Instead, our contributions are (i) to provide a theoretical benchmark for formally modeling and studying invasiveness of belief-elicitation tasks, and (ii) to establish conditions under which noninvasive scoring rules can be constructed.

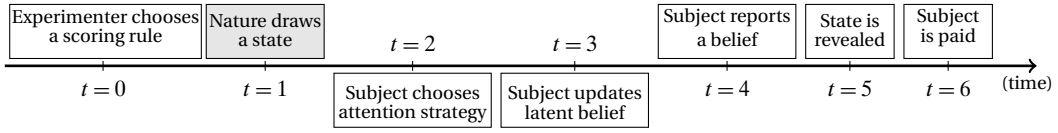


FIGURE 1. Boxes above the line are observed symmetrically by the subject and the experimenter. Boxes below the line are only observed by the subject. The shaded box is observed with a delay, i.e., it is realized at $t = 1$ and observed at $t = 5$.

Formally, we consider scoring rules in a model with hidden information costs, which typically emerge as an expression of rationally inattentive preferences (for an overview, see Caplin 2016). In our formal model, there is a (male) agent—henceforth called the subject—who has a latent (prior) probabilistic belief for some fixed event.[1] A (female) agent—henceforth called the experimenter— wants to elicit this belief, and to this end she asks the subject to report it. In order to incentivize him to report truthfully, she designs a scoring rule that rewards the subject on the basis of his report and the realization of the event. Before stating his report, the agent can acquire information through a *costly attention strategy* and then reports his belief after having perhaps updated his prior (see Figure 1 for the timeline).

In order to guarantee that the subject's prior belief is elicited, the scoring rule must make it simultaneously (a) strictly dominant to not acquire any information (i.e., to choose the degenerate zero-attention strategy), and (b) strictly dominant to report truthfully (i.e., the scoring rule must be proper). Such a mechanism is called *robust scoring rule*. Two natural questions arise then. *Does a robust scoring rule exist?* And if yes, *how does it look like?* Note that, in expectation, every attention strategy yields a benefit (due to the fact that reporting is postponed until after information has been acquired and the beliefs have been updated) and a cost (due to the fact that information acquisition is costly). Thus, the experimenter's problem boils down to finding a scoring rule that provides sufficiently strong incentives for the agent to report truthfully, but not so strong to offset the costs of acquiring information. In this sense, our work can be seen as part of a larger literature that studies the tradeoff between material payoffs and cognitive costs (Alaoui and Penta 2016, 2018, Alaoui et al. 2019). The novelty of our work is that we try to exploit—rather than to overcome—the presence of such costs.

Before moving on with our results, let us make some important remarks on our basic model. First, our entire analysis can be directly extended from a rational inattention framework (where costs are cognitive) to any costly information acquisition framework (where costs may even be material) (e.g., Cabrales et al. 2013, 2017). For instance, in a marketing survey like the one in our motivating example, an attention strategy may correspond to a costly (Bayesian) experiment that the subject can undertake before reporting a belief, e.g., he may elect to buy a sample product and try it. Second, it is not necessarily the case that the subject has at his disposal all possible attention strategies.

[1] Throughout the paper—being aligned with the rational inattention literature—the term "prior belief" refers to the belief held by the subject in the absence of elicitation.

In fact, depending on the specific environment/application, there may exist hard restrictions on the attention strategies that he can use. Nevertheless, since the aim of this paper is to provide conditions under which *no attention strategy is beneficial*, we focus on the unrestricted case.

Our first main theorem shows that robust scoring rules exist under a mild condition on the functional form of the attention costs (Theorem 1). In particular, it suffices that the attention costs satisfy posterior-separability, a property that generalizes Sims's (2003) usual (Shannon) entropic specification of the cost function. Posterior-separability has recently attracted interest within the rational inattention literature, primarily due to its solid theoretical foundations (Caplin et al. 2019, Zhong 2017, Tommaso 2020) and the presence of supporting experimental evidence (Dean and Neligh 2019). Later in the paper, we show that posterior-separability is a rather tight condition, in the sense that minor relaxations lead to nonexistence of robust scoring rules (Section 6). The proof of our theorem is constructive. Notably, not only do we show existence, but we also explicitly identify an entire class of robust scoring rules for each posterior-separable cost function. In this respect, our theory has strong empirical content.

We subsequently weaken our notion of robustness, in order to (simultaneously) address two issues that frequently appear in practice. First, suppose that the experimenter restricts herself to a specific family of scoring rules (e.g., she wants to use a quadratic scoring rule), but unfortunately there is no robust scoring rule within this family. Second, suppose that the experimenter is uncertain about the subject's cost function (e.g., she knows that the subject's attention costs are entropic but does not know the multiplier parameter). This is often the case when the experimenter does not have enough data to calibrate each individual subject's actual cost structure, and instead she has formed a probabilistic estimate over the set of possible cost specifications based on past (individual or population) data. The latter is particularly common in surveys that aim at eliciting prior beliefs in a sample of individuals with heterogeneous cost functions. Then we ask the following question: *when one (or both) of the previous issues arises, how closely can the experimenter approximate the subject's prior beliefs?*

A scoring rule scoring rule is called (ε, δ)-robust if, it elicits a belief sufficiently close to the subject's prior (viz., not further than ε-away) with sufficiently high probability (viz., with probability at least $1-\delta$). In other words, the experimenter is sufficiently certain that the subject will not find it optimal to acquire a lot of information which could potentially lead to a posterior far away from his prior. For practical purposes, an approximation of the prior belief is good if the scoring rule is (ε, δ)-robust for small ε and δ.

Our second main result proves that, if we take an arbitrary proper scoring rule, any $\varepsilon > 0$ and any $\delta > 0$ bounded by the probability of the cost function being nonposterior-separable, we can weaken the incentives by proportionately reducing the rewards of the scoring rule until it becomes (ε, δ)-robust (Theorem 2).[2] The first implication of this result is that essentially every scoring rule can be used to arbitrarily approximate the

[2]In fact, in our formal treatment we prove a stronger result that allows the scoring rule to be even weakly proper.

subject's prior, thus suggesting that we do not need to resort to exotic mechanisms in order for our theory to have a bite, e.g., simply put, the quadratic scoring rule always does the job sufficiently well. This addresses our first issue. The second implication is that we can always construct scoring rules that approximate the prior even if we are uncertain about the subject's cost function. This addresses our second issue.

This paper should be placed at the intersection of two different streams of literature, viz., belief elicitation via scoring rules and rational inattention. Scoring rules were originally introduced by meteorologists (Brier 1950), before being further developed by statisticians (Good 1952, McCarthy 1956, Savage 1971), and eventually being adopted by several disciplines, such as economics, accounting, business, management, psychology, political science and computer science (Offerman et al. 2009). For two recent literature reviews, we refer to Schotter and Trevino (2014) and Schlag et al. (2015). On the other hand, rational inattention models first appeared in macroeconomics (Sims 2003, 2006), before attracting interest of microtheorists. The latter have mostly focused on providing axiomatic foundations (De Oliveira et al. 2017, Ellis 2018) and on designing revealed-preference tests for identifying the attention costs from choice data (Caplin and Dean 2015, Chambers et al. 2018, Caplin et al. 2019). For recent literature reviews, see Caplin (2016) and Maćkowiak et al. (2018).

Of particular interest is the relationship between our paper and the one of Chambers and Lambert (2017) in that they are among the handful of papers that study dynamic belief elicitation. To the best of our knowledge, the only other paper is the one by Karni (2020).[3] In their paper, Chambers and Lambert (2017) consider an agent who has a latent prior belief and receives new information over time based on an exogenously given dynamic process. Then they construct a mechanism which makes it incentive-compatible for the agent to simultaneously reveal his prior, his anticipated information flow and his realized posteriors. The conceptual difference to our paper is that the agent does not strategically choose the process of his information flow (viz., the attention strategy in our terminology). Moreover, the two papers differ in the formal approaches that they employ, viz., as opposed to our paper, their mechanism does not rely on the usual subgradient characterization, but rather on a randomization technique originally introduced by Allais (1953). On the other hand, a major similarity is that both our paper and the one of Chambers and Lambert (2017) truthfully elicit the agent's prior beliefs.

Another paper that is closely related to our work is the one by Clemen (2002), who also studies the possible effect of scoring rules in information acquisition. However, unlike our paper, his aim is not to preclude information acquisition, as he considers scoring rules that are primarily used as incentive schemes for experts. This approach can be further explored in future research by studying the converse problem to the one we study in our paper, viz., how to encourage the subject to acquire as much information as possible in order to provide more accurate predictions.

Overall, ours is one of the few paper on mechanism design with rational inattention, with the distinctive feature that inattention is desired by the designer. In a different framework, Yang (2020) studies a security design problem, where the seller wants to

[3]I am indebted to Chris Chambers for pointing out these connections.

design the security in a way such that the optimal strategy of the rationally inattentive buyer is not to acquire any or to acquire limited limited information, in order to avoid adverse selection effects. One major difference is that our experimenter wants the subject to remain uninformed because she is inherently interested in her prior beliefs per se, whereas in the type of problems that Yang (2020) studies, the seller's interest in the buyer remaining uninformed is simply a byproduct of the designer's preference to maximize expected monetary payoff.

In Section 2, we introduce our basic framework. In Section 3, we study exact robustness and we present our first main result. In Section 4, we introduce approximate robustness, and we present our second main result. In Section 5, we study standard special cases of scoring rules. In Section 6, we revisit our posterior-separability condition. Section 7 contains a discussion. All proofs are relegated to the Appendix.

2. Preliminaries

2.1 *Scoring rules*

Consider a binary state space $\Omega = \{\omega_0, \omega_1\}$. A risk-neutral (male) experimental subject has a latent subjective belief $\mu_0 \in [0, 1]$ of ω_0 occurring, which is not observed by the (female) experimenter. The subject is asked to state μ_0 and reports some $r \in [0, 1]$, which is not necessarily equal to μ_0. A *scoring rule* is a function

$$S : [0, 1] \times \Omega \to \mathbb{R},$$

chosen by the experimenter, which takes the subject's report (r) and the realized state (ω) as an input and returns a monetary payoff ($S_r(\omega)$) as an output. In economics, we sometimes consider binarized scoring rules where the subject is paid in probabilities of winning a fixed prize. Binarized scoring rules are used to elicit the subject's belief for arbitrary risk attitudes. Our entire analysis is directly extended to binarized scoring rules, implying that our assumption of the subject being risk-neutral is without loss of generality and can therefore be dispensed with (Section 5.2).

The subject is assumed to maximize his Subjective Expected Utility (SEU), i.e., given the scoring rule (S) and a belief (μ), he chooses the report (r) that maximizes

$$\mathbb{E}_\mu(S_r) := \mu S_r(\omega_0) + (1 - \mu) S_r(\omega_1).$$

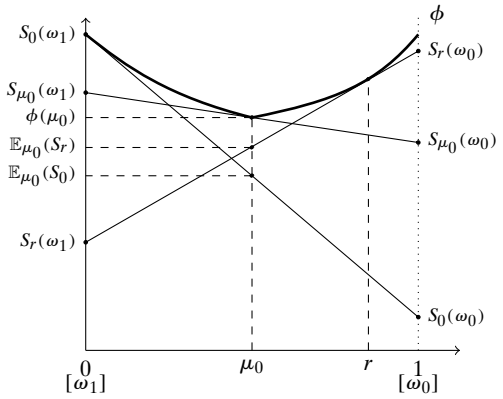
A scoring rule is called proper whenever, for every belief, it is strictly dominant to report truthfully (Brier 1950, Good 1952). Hence, whenever the subject says r, the experimenter directly infers that $\mu_0 = r$. Formally properness is defined as follows.

Definition 1 (Proper scoring rule). The scoring rule S is called proper, whenever

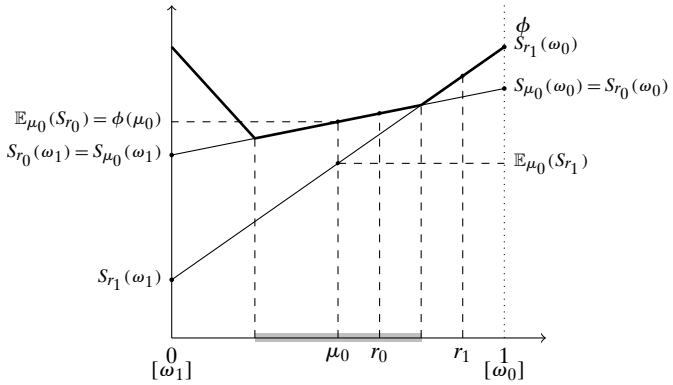
$$\mathbb{E}_\mu(S_\mu) > \mathbb{E}_\mu(S_r)$$

for every $r \neq \mu$ and every $\mu \in [0, 1]$.

It is well known in the literature that each proper scoring rule is characterized by a strictly convex function (McCarthy 1956, Savage 1971). Let us illustrate the idea behind



(a) PROPER SCORING RULE: The strictly convex function $\phi(\mu) := \mathbb{E}_\mu(S_\mu)$ generates a proper scoring rule by taking an arbitrary tangent for each r and evaluating it at 0 and 1, respectively. Properness follows from the fact that the tangent at μ_0 when evaluated at μ_0 lies higher than the tangent at any other r when evaluated at μ_0, i.e., $\mathbb{E}_{\mu_0}(S_{\mu_0}) > \mathbb{E}_{\mu_0}(S_r)$.



(b) WEAKLY PROPER SCORING RULE: The weakly convex function $\phi(\mu) := \mathbb{E}_\mu(S_\mu)$ generates a weakly proper scoring rule by taking an arbitrary tangent for each r and evaluating it at 0 and 1, respectively. Weak properness follows from the fact that the tangent at μ_0 when evaluated at μ_0 lies at least as high as the tangent at any other r when evaluated at μ_0. The subject is indifferent among all reports in the shaded subinterval $I_\phi(\mu_0)$, i.e., $\mathbb{E}_{\mu_0}(S_{\mu_0}) > \mathbb{E}_{\mu_0}(S_{r_1})$ while $\mathbb{E}_{\mu_0}(S_{\mu_0}) = \mathbb{E}_{\mu_0}(S_{r_0})$.

FIGURE 2. Subgradient characterization of properness and weak properness.

this characterization. For a graphical illustration, see Figure 2(a). First, define the subject's subjective expected utility from reporting truthfully as a function of his beliefs, viz., for each $\mu \in [0, 1]$, let

$$\phi(\mu) := \mathbb{E}_\mu(S_\mu).$$

It is not difficult to verify that if S is proper then ϕ is strictly convex. This is because properness implies $\phi(\mu) = \max\{\mathbb{E}_\mu(S_r)|r \in [0,1]\}$, i.e., ϕ can be written as the pointwise maximum of a family of linear functions. Interestingly, the converse is also true. Namely, every strictly convex and subdifferentiable function $\phi : [0,1] \to \mathbb{R}$ induces a unique class of (essentially equivalent) proper scoring rules. Let us elaborate. We first consider a subtangent $t_r(\mu) := a_r + b_r\mu$ of the function ϕ at $r \in [0,1]$. Then we define a scoring rule S induced by ϕ, by letting $S_r(\omega_0) := t_r(1)$ and $S_r(\omega_1) := t_r(0)$, i.e., if the subject reports r, we take the tangent t_r and evaluate it at 1 and 0 to obtain the two rewards that correspond to ω_0 and ω_1, respectively. Of course by strict convexity of ϕ, the tangent at μ_0 when evaluated at μ_0 will lie higher than any other tangent evaluated at the same point μ_0. Hence, we obtain $\mathbb{E}_{\mu_0}(S_{\mu_0}) > \mathbb{E}_{\mu_0}(S_r)$, implying that ϕ is proper.

Let us make two important remarks. First, by convexity of ϕ, a subtangent always exists at every interior $r \in (0,1)$. By requiring that ϕ is subdifferentiable, we guarantee that it also exists at the boundaries, i.e., ϕ does not becomes infinitely steep either at 0 or at 1. This is needed in order to guarantee that $S_0(\omega_0)$ and $S_1(\omega_1)$ remain finite. This last condition can be dispensed with, if we allow S to take values in $\overline{\mathbb{R}}$ instead of $\mathbb{R}$, as often done in statistics. Second, when the subtangent is not unique (e.g., when ϕ is not differentiable, like at μ_0 in Figure 2(a)) we arbitrarily select one of the infinitely many subtangents, implying that there is an entire class of scoring rules induced by ϕ. Nevertheless, irrespective of which of those scoring rules we pick, the subject's optimal expected utility will remain the same (equal to $\phi(\mu_0)$ in this case), which is why we have earlier said that all scoring rules derived from ϕ are essentially equivalent.

Example 1 (Quadratic scoring rule). The most commonly used proper scoring rule is the quadratic scoring rule (QSR), which is defined by $S_r(\omega_0) := \alpha - \beta(1-r)^2$ and $S_r(\omega_1) := \alpha - \beta r^2$, where $\alpha \in \mathbb{R}$ and $\beta > 0$. Accordingly, the subject is paid a fixed amount α minus a penalty which is proportional to the squared distance from the realized state. The strictly convex function that characterizes the quadratic scoring rule is

$$\phi_\beta(\mu) := \alpha - \beta\mu(1-\mu).$$

Notice that we explicitly index the characteristic function of the quadratic scoring rule with the parameter β that determines the strength of the incentives provided by the scoring rule. ◊

For a review of other standard proper scoring rules, we refer to Schlag et al. (2015, Section 2), while for an overview of the subgradient characterization that we presented above we recommend Gneiting and Raftery (2007).

A scoring rule is weakly proper if reporting truthfully is one of the optimal reports, but not necessarily the only one. Formally, weak properness is defined as follows.

Definition 2 (Weakly proper scoring rule). The scoring rule S is called weakly proper, whenever

$$\mathbb{E}_\mu(S_\mu) \geq \mathbb{E}_\mu(S_r)$$

for every $r \in [0,1]$ and every $\mu \in [0,1]$.

Geometrically, the function ϕ that characterizes S is (only) weakly convex (Figure 2(b)). The set of optimal reports (at some belief μ) is denoted by

$$I_\phi(\mu) := \arg\max_{r\in[0,1]} \mathbb{E}_\mu(S_r).$$

Geometrically, $I_\phi(\mu) := [r_\phi^-(\mu), r_\phi^+(\mu)]$ is the largest interval of μ where ϕ is linear, e.g., $I_\phi(\mu_0)$ is the shaded subinterval in Figure 2(b). It is straightforward to verify that $S_r = S_{r'}$ for every $r, r' \in \operatorname{int}(I_\phi(\mu))$, as all points in the interior of $I_\phi(\mu)$ share the same subtangent. The rewards of reports that lie on the boundary of $I_\phi(\mu)$ may or may not be the same too, as ϕ may have multiple subtangents at such a boundary point. This is the case when ϕ is not differentiable, like for instance at the kinks of the graph in Figure 2(b).

Definition 3 (ε-proper scoring rule). A weakly proper scoring rule ϕ is called ε-proper, for some $\varepsilon \geq 0$, whenever

$$\left|I_\phi(\mu)\right| \leq \varepsilon$$

for every $\mu \in [0,1]$, where $|I_\phi(\mu)| := r_\phi^+(\mu) - r_\phi^-(\mu)$ is the length of the interval $I_\phi(\mu)$.

Intuitively, ε puts a uniform bound on the errors that the (weakly proper) scoring rule can yield, i.e., ε-properness guarantees that the subject will never report further than ε-away from his true belief. Obviously, the scoring rule is proper if and only if $I_\phi(\mu) = \{\mu\}$ for every $\mu \in [0,1]$, i.e., whenever ϕ is 0-proper.

2.2 *Costly attention*

We now enrich the agent's preferences to allow for information acquisition by means of costly attention. An *attention strategy* is an experiment (viz., a Bayesian signal), designed by the subject in an attempt to acquire information about the state space, and it typically leads him to update his subjective beliefs. Given his prior μ_0, each attention strategy is identified by a (Bayes-plausible) distribution over posteriors, chosen from the set $\Pi(\mu_0) := \{\pi \in \Delta([0,1]) : \int_0^1 \mu \, d\pi = \mu_0\}$. We define the degenerate zero-attention strategy $\hat{\mu}_0 \in \Pi(\mu_0)$, as the one that does not carry any information and, therefore, yields the prior μ_0 with probability 1. On the other hand, by $\pi^*_{\mu_0} \in \Pi(\mu_0)$ we denote the most informative attention strategy given the prior μ_0, implying that $\pi^*_{\mu_0}$ attaches probability μ_0 to the posterior that puts probability 1 to ω_0 and probability $1-\mu_0$ to the posterior that puts probability 1 to ω_1. For notation simplicity, we henceforth denote by $\hat{\Pi}(\mu_0) := \Pi(\mu_0) \setminus \{\hat{\mu}_0\}$ the set of all nondegenerate attention strategies, observing that $\hat{\Pi}(\mu_0) = \varnothing$ whenever $\mu_0 \in \{0,1\}$.

As usual, attention is assumed to be costly. In particular, there is a continuous cost function,

$$C : \Delta\big([0,1]\big) \to \mathbb{R}_+$$

assigning a nonnegative cost to every attention strategy $\pi \in \Pi(\mu)$ for every prior $\mu \in [0,1]$. In Section 7.3, we clarify why our continuity assumption is without loss of generality. Attention costs can be identified from state-dependent stochastic-choice data

(Caplin and Dean 2015, Chambers et al. 2018) or from menu-choice data (De Oliveira et al. 2017, Ellis 2018). Throughout the literature, some structure on the cost function is either postulated or derived from primitive axioms of choice. In this paper, we consider cost functions that satisfy a property that has recently attracted interest in the rational inattention literature (Caplin et al. 2019).

DEFINITION 4 (Posterior-separability). A cost function C is said to be posterior-separable, if there is a strictly concave function $K : [0,1] \to \mathbb{R}$ such that

$$C(\pi) = K(\mu) - \langle K, \pi \rangle \tag{1}$$

for every $\pi \in \Pi(\mu)$ and every $\mu \in [0,1]$, where $\langle \cdot, \cdot \rangle$ denotes the inner product as usual.

Notice that for every posterior-separable cost function there is in fact an entire class of strictly concave functions that satisfy equation (1), viz., for any linear function $L : [0,1] \to \mathbb{R}$, the function $K + L$ satisfies (1) if and only if K satisfies it too. Throughout the paper, we uniquely identify the posterior-separable C by the function K from the aforementioned class such that $K(0) = K(1) = 0$, and with slight abuse of terminology we often refer to K as the subject's cost function. In this case, $K(\mu)$ is naturally interpreted as the cost of the most informative attention strategy for each prior μ, i.e., formally, $K(\mu) := C(\pi^*_\mu)$. Finally note that by continuity of C, the function K is also continuous.

Posterior-separability is supported by recent experimental findings (Dean and Neligh 2019) and has solid theoretical foundations (Caplin et al. 2019, Zhong 2017, Tommaso 2020). Later in the paper we further elaborate regarding how restrictive our posterior-separability assumption is, both in a general axiomatic sense, as well as in the context of our results (Section 6).

EXAMPLE 2 (Entropic attention costs). The most common functional form of attention costs in the literature is the entropic specification (Sims 2003, 2006, Caplin et al. 2019), which among other nice properties, allows us to provide microeconomic foundations to the multinomial logit model (Matějka and McKay 2015, Steiner et al. 2017) and is moreover assumed in various applications (Matějka 2016, Martin 2017, Yang 2020). Accordingly, the cost of $\pi \in \Pi(\mu_0)$ is equal to

$$C_\kappa(\pi) := \kappa\big(H(\mu_0) - \langle H, \pi \rangle\big),$$

where $H(\mu) := -\mu \log \mu - (1-\mu)\log(1-\mu)$ is the Shannon entropy (Shannon 1948), and $\kappa > 0$ is a multiplier parameter. It is straightforward to verify that C_κ is posterior separable with $K = \kappa H$ being the corresponding function whose expected decrease gives the cost of attention. ◊

2.3 *Cost-benefit analysis*

Given a prior μ_0 and a weakly proper scoring rule ϕ, the (expected) benefit of an attention strategy $\pi \in \Pi(\mu_0)$ is equal to

$$B_\phi(\pi) := \langle \phi, \pi \rangle - \phi(\mu_0). \tag{2}$$

By convexity of ϕ, we obtain $B_\phi(\pi) \geq 0$, with equality holding if and only if $\text{supp}(\pi) \subseteq I_\phi(\mu_0)$. Clearly, whenever ϕ is proper, by strict convexity, every nondegenerate attention strategy yields a strictly positive expected benefit.

Taking into account simultaneously the expected benefits and the costs, the subject will choose an attention strategy in $\Pi(\mu_0)$ that maximizes the value

$$V_\phi(\pi) := B_\phi(\pi) - C(\pi).$$

The idea is that, after (optimally) choosing some $\pi \in \Pi(\mu_0)$, the subject will first update his beliefs to some (also latent) posterior $\mu \in \text{supp}(\pi)$, and then—as ϕ is weakly proper—he will report some posterior belief that yields expected utility equal to $\phi(\mu)$, i.e., he will report some posterior belief in $I_\phi(\mu)$. The fact that V_ϕ obtains a maximum in $\Pi(\mu_0)$ follows from C being continuous. Now, take the function

$$\psi := K + \phi,$$

and, by (1) and (2), rewrite the subject's value as follows:

$$V_\phi(\pi) = \langle \psi, \pi \rangle - \psi(\mu_0).$$

The latter implies, that the subject's optimal value is equal to $\overline{\psi}(\mu_0) - \psi(\mu_0)$, where $\overline{\psi}(\mu) := \max\{\langle \psi, \pi \rangle | \pi \in \Pi(\mu)\}$ is the concave closure of ψ.

The optimal attention strategies can be easily computed, using the (concavification) method which was first introduced in the repeated games literature by Aumann and Maschler (1995), and was later extensively used in the Bayesian persuasion literature starting with Kamenica and Gentzkow (2011). Accordingly, we first define the interval

$$J_\psi(\mu) := \big[a_\psi(\mu), b_\psi(\mu)\big], \tag{3}$$

which is the longest interval of μ where $\overline{\psi}$ is linear (see shaded subinterval in Figure 3(b)). Obviously, if $\overline{\psi}$ is not linear in any interval of μ, then $J_\psi(\mu) := \{\mu\}$. Then an attention strategy $\pi \in \Pi(\mu)$ is optimal if and only if $\pi(\{\nu \in J_\psi(\mu) : \overline{\psi}(\nu) = \psi(\nu)\}) = 1$. Clearly, the strategy π that satisfies $\pi(\{a_\psi(\mu), b_\psi(\mu)\}) = 1$ is the most informative among all optimal attention strategies.

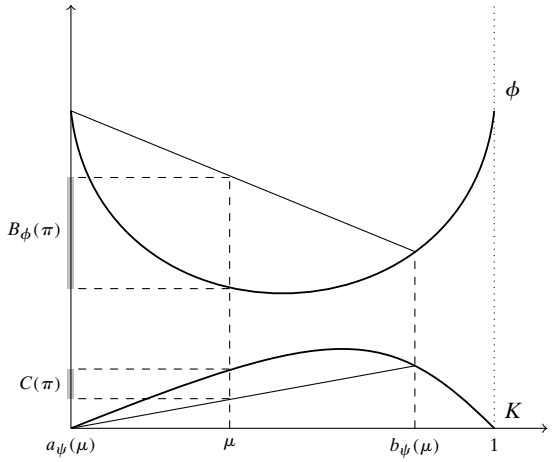
3. Exact robustness

Recall that the experimenter wants to elicit the subject's prior. In order to guarantee that the subject will not acquire any information (thus making sure that he will not update his prior belief), she must design a scoring rule that makes $\hat{\mu}$ a strictly dominant attention strategy for every $\mu \in [0, 1]$. If moreover the scoring rule is proper, then it will also be strictly dominant to report the prior. A scoring rule that satisfies these two conditions is called robust.

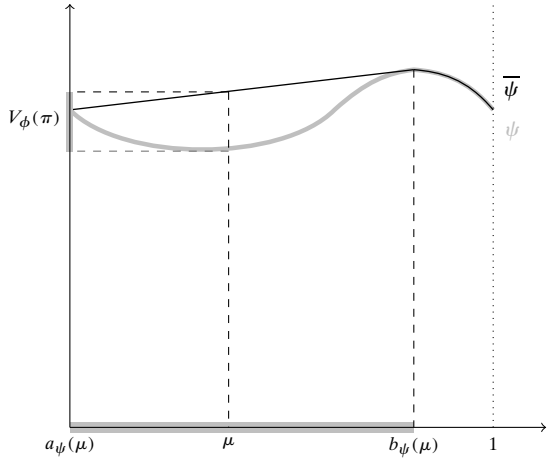
Definition 5 (Robust scoring rules). A scoring rule ϕ is robust, if it is proper and satisfies

$$V_\phi(\hat{\mu}) > V_\phi(\pi)$$

for every $\pi \in \hat{\Pi}(\mu)$ and every $\mu \in [0, 1]$.



(a) COST-BENEFIT ANALYSIS: The expected cost of the attention strategy $\pi \in \Pi(\mu)$ is given by the expected increase of ϕ, whereas the corresponding cost is given by the expected decrease of K. The value of π is the length difference of the respective shaded areas in the vertical axis.



(b) CONCAVIFICATION OF THE VALUE FUNCTION: Take the concave closure $\overline{\psi}$ of the function $\psi := K + \phi$, and find the longest subinterval $J_\psi(\mu)$ of μ where $\overline{\psi}$ is linear (shaded area). The optimal attention strategies are those that put positive probability only to posteriors in $J_\psi(\mu)$ such that $\overline{\psi}$ coincides with ψ.

FIGURE 3. Optimal attention strategy.

The overall idea is that the scoring rule provides strong enough incentives to induce truth-telling (viz., ϕ must be strictly convex), but not so strong to lead the subject to acquire information (viz., ϕ should not be "too convex"). The following intermediate result

uses the concavification method that we presented in the previous section to characterize those proper scoring rules that satisfy our robustness criterion, under the assumption that the cost function is posterior-separable.

LEMMA 1. *Fix a posterior-separable cost function K and a proper scoring rule ϕ. Then ϕ is robust, if and only if, $K+\phi$ is strictly concave.*

The intuition is quite obvious: $K+\phi$ is is strictly concave if and only if K is strictly "more concave" than $-\phi$, which is in turn equivalent to the costs of attention always offsetting the benefits, as required by robustness.

EXAMPLE 3 (Robust QSR when the costs are entropic). Let the attention costs be entropic (as in Example 2) and take a quadratic scoring rule (as in Example 1). Is there a specification (i.e., parameters α and β) such that the quadratic scoring rule is robust (given the cost parameter κ)? First, we take the function $\psi_\beta(\mu) := \kappa H(\mu)+\alpha-\beta\mu(1-\mu)$. By Lemma 1, the scoring rule ϕ_β is robust if and only if $\psi''_\beta \leq 0$ with equality holding at finitely many points in $[0,1]$. Solving the previous inequality yields $\beta \leq 2\kappa$, implying that we must bound the incentives provided by the scoring rule from above, in order to guarantee robustness. Finally, observe that only the parameter β is relevant for robustness. This is not surprising, given that the incentives of a scoring rule are measured in terms of its convexity, and the constant α does not affect the degree of convexity of ϕ_β, but rather it simply rescales the payments by adding a constant. ◊

Our following first main result shows that robust scoring rules exist, not only when costs are entropic, but rather for all posterior-separable specifications.

THEOREM 1. *If the cost function is posterior-separable, there is a robust scoring rule.*

The proof is constructive, thus allowing us not only to prove that a robust scoring rule exists, but also to identify its functional form. Let us sketch the main steps here, while the full proof is relegated to Appendix A.

SKETCH OF THE PROOF. We begin by considering the function

$$f := a - bK,$$

where $a \in \mathbb{R}$ and $b \in (0,1)$. If a function $\phi : [0,1] \to \mathbb{R}$ is such that $f-\phi$ is convex, then $K+\phi$ is strictly concave. Therefore, if ϕ is strictly convex and subdifferentiable, we can use it as our robust scoring rule (by Lemma 1). The next intermediate result shows that such a ϕ exists, thus completing our proof.

LEMMA 2. *Consider a strictly convex function $f : [0,1] \to \mathbb{R}$. Then there exists a strictly convex and subdifferentiable function $\phi : [0,1] \to \mathbb{R}$ such that $f-\phi$ is convex.*

The proof of Lemma 2 is constructive, implying that not only do we prove existence, but we also identify an entire family of robust scoring rules. Let us illustrate the proof of the previous lemma for a function f which is continuously differentiable in $(0,1)$ and becomes infinitely steep at 0 and 1. Define $F:[0,1]\to\overline{\mathbb{R}}$ as the slope of f, i.e., $F(x):=f'(x)$ for every $x\in(0,1)$, while $F(0)=-\infty$ and $F(1)=\infty$. Then, define the composition $\Phi(x):=\tan^{-1}(F(x))$, and let ϕ be a primitive of Φ. Since both $\tan^{-1}$ and F are strictly increasing, so is Φ, implying that ϕ is strictly convex. Moreover, since $\tan^{-1}$ is Lipschitz with constant less than 1, $F-\Phi$ is increasing and, therefore, $f-\phi$ is convex. The proof is easily extended to the nondifferentiable case. □

Note that the proof of our previous theorem describes the construction of only one class of robust scoring rules. In fact, there are many more scoring rules that serve the same purpose of eliciting the subject's prior beliefs, e.g., when the costs are entropic, the robust scoring rule that we obtain following the steps of our proof is one possibility, while the robust QSR from Example 3 is another. Hence, we naturally ask whether everything goes, viz., does every scoring rule become robust if we shrink the incentives via multiplication with a sufficiently small constant? The following example illustrates that this is not the case, implying that (exact) robustness may rule out widely-used families of scoring rules, such as for instance the quadratic scoring rule. This observation is further discussed in the next section.

Example 4. Consider the posterior-separable cost function $K(\mu)=\mu-\mu^3$. Note that the second derivative of K is not bounded away from 0, i.e., the cost function becomes arbitrarily flat (close to 0). Now consider a quadratic scoring rule ϕ_β, and observe that its second derivative is bounded away from 0, irrespective of how small β is. Thus, we obtain $\psi''_\beta(\mu)=2\beta-6\mu$, implying that $\psi_\beta=K+\phi_\beta$ is convex in $[0,\beta/3]$. Hence, by Lemma 1, it follows that ϕ_β is not robust for any $\beta>0$. ◊

4. Approximate robustness

Two natural questions are posed at this point. First, if we fix some scoring rule, how close can we get to the subject's prior by weakening the incentives? Second, if we relax the assumption that the experimenter knows the subject's cost function, can we still approximate the subject's prior beliefs?[4] In this section, we will answer both questions simultaneously through a single result.

Let us begin by observing that often times the experimenter accepts small errors in the elicitation of the subject's beliefs, either because she understands that there are restrictions in the experimental technology or because she does not care about minor mistakes. In our context, for a given small $\varepsilon\geq 0$, assume that the experimenter is satisfied if she can elicit a belief not further than ε away from the subject's prior. A scoring rule that achieves this goal will be called ε-robust.

[4] I am indebted to Burkhard Schipper for suggesting this approach.

DEFINITION 6 (ε-robust scoring rule). A weakly proper scoring rule ϕ is ε-robust, if for all $\mu \in [0,1]$, for all $\pi \in \arg\max_{\rho \in \Pi(\mu)} V_\phi(\rho)$ and for all $\nu \in \text{supp}(\pi)$, it is the case that

$$I_\phi(\nu) \subseteq B_\varepsilon(\mu),$$

where $B_\varepsilon(\mu) := \{\nu \in [0,1] : |\mu - \nu| \leq \varepsilon\}$ is the closed ε-neighborhood of μ.

Intuitively, there are two forces that may bring the optimal report away from the prior. On the one hand, ϕ can be locally linear, due to the fact that the scoring rule is just weakly convex. On the other hand, $K + \phi$ can be locally nonconcave, due to the incentives to acquire information. So the idea behind ε-robustness is that even when combined, these two forces will not lead to a posterior further than ε away from the prior belief. Clearly, a proper scoring rule is ε-robust, if every optimal attention strategy yields posteriors at most ε away from the prior, i.e., in this case, only the second force (viz., the incentives to acquire information) may lead the subject to misreport.

Now turning to our second question, sometimes the experimenter cannot pin down the subject's cost function with certainty. This is typically due to the experimenter not having enough data to calibrate the subject's (actual) cost function. In such cases, she instead resorts to an estimated probability distribution over cost functions. Can the experimenter then be sufficiently certain that she will approximate the subject's prior beliefs with sufficient precision? Formally speaking, can the experimenter find a scoring rule which is ε-robust with sufficiently high probability?

Let $\mathcal{C}$ denote the space of continuous (weakly) concave functions $K : [0,1] \to \mathbb{R}$, together with the topology induced by the sup norm $\|\cdot\|_\infty$. Let $\mathcal{K} \subset \mathcal{C}$ be the space of strictly concave functions such that $K(0) = K(1) = 0$. As we have already mentioned, each posterior-separable cost function is identified by a unique $K \in \mathcal{K}$. Nonposterior-separable costs that correspond to functions $K \in \mathcal{C} \setminus \mathcal{K}$ are studied in Section 6. Uncertainty about the subject's costs is described by a distribution $P \in \Delta(\mathcal{C})$. Whenever $P(\mathcal{K}) = 1$, the experimenter is certain that the subject's cost function is posterior-separable.

For each scoring rule ϕ, define the function $\varepsilon_\phi : \mathcal{C} \to \mathbb{R}_+$ by

$$\varepsilon_\phi^K := \inf\{\varepsilon \geq 0 : \phi \text{ is } \varepsilon\text{-robust given } K\},$$

which provides a uniform bound on the approximation of the prior that the experimenter can achieve with ϕ for some $K \in \mathcal{C}$. Note that ε_ϕ is upper semicontinuous (Lemma B.1). Hence, $\{K \in \mathcal{C} : \varepsilon_\phi^K \leq \varepsilon\}$ is Borel measurable. That is, the event that "the scoring rule ϕ is ε-robust" is expressible in the experimenter's language and, therefore, the experimenter assigns some probability to it.

DEFINITION 7 ((ε,δ)-robust scoring rule). A weakly proper scoring rule ϕ is (ε,δ)-robust, if

$$P\big(\{K \in \mathcal{C} : \varepsilon_\phi^K \leq \varepsilon\}\big) \geq 1 - \delta,$$

given some fixed $\varepsilon \geq 0$ and $\delta \geq 0$.

In other words, approximate robustness guarantees that the probability of eliciting a belief further than ε away from the prior is smaller than δ. Although both ε and δ place bounds in the precision of the scoring rule, the two bounds operate in different probability spaces, viz., ε is a bound in the space of (the subject's) probabilities over Ω, whereas δ is a bound in the space of (the experimenter's) probabilities over $\mathcal{C}$. Obviously, the concept has a bite when both ε and δ become sufficiently small. When they both collapse to 0, we are essentially back to exact robustness.

THEOREM 2. *Let $P(\mathcal{K}) = p$, and fix an arbitrary $\tilde{\varepsilon}$-proper scoring rule ϕ. Then, for every $\varepsilon > \tilde{\varepsilon}$ and every $\delta > 1 - p$, there exists some $\lambda > 0$ such that $\lambda\phi$ is (ε, δ)-robust.*

The intuition of the previous result is that we can begin with any weakly proper scoring rule, and by shrinking its incentives enough we can arbitrarily approximate the subject's prior beliefs. Below we present the main ideas underlying our proof, and the full proof is relegated to Appendix B.

SKETCH OF THE PROOF. Fix some $\tilde{\varepsilon}$-proper scoring rule ϕ and some posterior-separable cost function K. For starters, we trivially observe that $\lambda\phi$ is also $\tilde{\varepsilon}$-proper for every $\lambda > 0$. Indeed, for every belief in $[0, 1]$, the optimal reports under ϕ and coincide with the optimal reports under $\lambda\phi$. Then, for each prior belief $\mu \in [0, 1]$, the most informative among all optimal attention strategies (given the scoring rule $\lambda\phi$ and the cost function K) distributes all probability mass between the posteriors $a_\lambda(\mu) \leq \mu$ and $b_\lambda(\mu) \geq \mu$, i.e., all posteriors that the subject can rationally form belong to the interval $[a_\lambda(\mu), b_\lambda(\mu)]$. Hence, by $\tilde{\varepsilon}$-properness of the scoring rule, the subject will eventually report a belief in $[a_\lambda(\mu) - \tilde{\varepsilon}, b_\lambda(\mu) + \tilde{\varepsilon}]$. The crucial step is then to prove that $a_\lambda(\mu) \uparrow \mu$ and $b_\lambda(\mu) \downarrow \mu$ as $\lambda \downarrow 0$ (Lemma B.2). Therefore, for every $\varepsilon > \tilde{\varepsilon}$ there exists a sufficiently small $\lambda > 0$ such that

$$\left[a_\lambda(\mu) - \tilde{\varepsilon}, b_\lambda(\mu) + \tilde{\varepsilon}\right] \subseteq [\mu - \varepsilon, \mu + \varepsilon].$$

Finally, notice that while λ approaches 0, the set of cost functions that satisfy the previous inclusion becomes larger approaching the entire set $\mathcal{K}$. Hence, the probability of the actual cost function satisfying this inclusion approaches the probability of $\mathcal{K}$ from above, thus completing the proof. □

Let us now present two examples which illustrate that the previous result simultaneously answers the two questions we posed at the beginning of this section.

EXAMPLE 4 CONTINUED. Recall from the previous section that there is no (exactly) robust QSR $\phi_\beta(\mu) = \alpha - \beta\mu(1 - \mu)$ when the posterior-separable cost function is $K(\mu) = \mu - \mu^3$. So, how far can we get using only a QSR? We begin by noticing that the concave closure of the function $\psi_\beta = K + \phi_\beta$ is strictly concave in the interval $[\beta/2, 1]$ and linear in the interval $[0, \beta/2]$. The latter implies that the subject will choose the zero-attention strategy when his prior is larger than $\beta/2$. On the other hand if his prior is smaller than $\beta/2$, he will optimally choose an attention strategy that will yield posterior beliefs either equal to 0 or equal to $\beta/2$. In both cases, since the QSR is proper, he will truthfully report

his posterior. Hence, the prior will be misreported only in the interval $(0, \beta/2)$, in which case the subject will say either 0 or $\beta/2$. Therefore, for any fixed $\varepsilon > 0$, by making β sufficiently small (and in particular, by taking $\beta < 2\varepsilon$), we guarantee that the scoring rule is ε-robust. We generalize our analysis of ε-robust QSR in Section 5.1.

EXAMPLE 5 (Approximate robustness under unknown entropic costs). Let P be uniformly distributed over the set of entropic cost functions in $\mathcal{E} := \{C_\kappa | \kappa \in [0,1]\} \subset \mathcal{K}$. Obviously, if we knew the exact value of κ, we would be able to find an exactly robust QSR (Example 2). Is it still possible to approximately elicit the subject's beliefs with a QSR even without knowing κ? Observe that the smaller β becomes the more likely it is for QSR to be robust. In particular, for every $\beta \in (0,1)$, the QSR ϕ_β is robust for all cost functions with $\kappa \geq \beta/2$. Hence, if we fix some $\delta > 0$, then for every $\beta \leq 2\delta$ the probability of the scoring rule being robust is at least $1 - \delta$. ♢

So far our approach has been to fix a pair (ε, δ) and then shrink ϕ until it becomes (ε, δ)-robust. Let us now briefly discuss a dual approach. Namely, if we begin with a fixed ϕ, what is the best we can achieve in terms of approximating the subject's prior?

Let us first observe that since (ε, δ)-robustness specifies two upper bounds, typically we cannot obtain a unique "best approximation" of the prior beliefs. This is because there is often a tradeoff between our two bounds, i.e., the more permissive we are in terms of how far from the prior we are willing to allow the elicited beliefs, the higher the probability of the subject's cost function being such that the elicited beliefs are within the margin of error that we allow. Formally, for each $\varepsilon \geq 0$, the lowest $\delta \geq 0$ such that ϕ is (ε, δ)-robust is defined by

$$\delta_\phi(\varepsilon) := \inf\{\delta \geq 0 : P(\{K \in \mathcal{C} : \varepsilon_\phi^K \leq \varepsilon\}) \geq 1 - \delta\}.$$

It is not difficult to verify that the function δ_ϕ is (weakly) decreasing, indicating that there is indeed a tradeoff between ε and δ. Finally, observe that weakening the incentives of ϕ shifts the entire graph of δ_ϕ downwards. Indeed, for every $\varepsilon \geq 0$, it is the case that $\delta_{\lambda\phi}(\varepsilon)$ is decreasing in λ. If we then fix some $\tilde{\varepsilon} > 0$ and $\delta > 1 - p$, there is some small enough $\lambda \in (0,1)$ such that $\delta_{\lambda\phi}(\varepsilon) \leq \delta$ for all $\varepsilon > \tilde{\varepsilon}$, which is another way of stating Theorem 2.

5. USUAL SCORING RULES

5.1 *Quadratic scoring rules*

As we have already illustrated, an approximately robust QSR exists, even in cases where an exactly robust QSR does not (Example 4). In fact, it follows as a direct consequence of our Theorem 2 that this is always the case, i.e., an approximately robust QSR always exists.

COROLLARY 1. *Let $P(\mathcal{K}) = p$. Then, for every $\varepsilon > 0$ and every $\delta > 1 - p$, there exists some $\beta > 0$ such that the quadratic scoring rule ϕ_β is (ε, δ)-robust.*

Clearly, if the (posterior-separable) cost function is known to the experimenter, the previous result can be further refined along the lines of Example 4, viz., for a given $K \in \mathcal{K}$ and for an arbitrary $\varepsilon > 0$, there exists some $\beta > 0$ such that the ϕ_β is ε-robust.

5.2 *Binarized scoring rules*

A binarized scoring rule is one that pays in probability units of winning a fixed prize (Hossain and Okui 2013, Schlag and van der Weele 2013). Formally, a binarized scoring rule S takes values in $[0,1]$, with the interpretation that $S_r(\omega) \in [0,1]$ is the objective probability of the subject winning the prize when he reports r and the realized state is ω. In this case, the subject's expected utility is linear in the probability of winning the prize irrespective of his risk preferences, and our analysis follows verbatim except for one small detail, viz., in order for a function ϕ to characterize a proper (resp., weakly proper) binarized scoring rule, not only should it be subdifferentiable, but it should also have at every point a subtangent that takes values in $[0,1]$ both when evaluated at 0 and at 1. It is not difficult to see that for every (weakly) convex and subdifferentiable function f there exists some $\lambda > 0$ and some $c \in \mathbb{R}$ such that $\phi := c + \lambda f$ satisfies $\partial\phi(\mu) \cap [0,1] \neq \varnothing$ for every $\mu \in [0,1]$, i.e., if a function is subdifferentiable, we can multiply it with a sufficiently small λ so that the subdifferential become sufficiently small to take values in $[0,1]$. Then the following result shows that our two main theorems hold for binarized scoring rules, too.

COROLLARY 2. *The following two conditions hold:*

(i) If the cost function is posterior-separable there exists a robust binarized scoring rule.

(ii) Let $P(\mathcal{K}) = p$, and fix an $\tilde{\varepsilon}$-proper binarized scoring rule ϕ. Then, for every $\varepsilon > \tilde{\varepsilon}$ and every $\delta > 1 - p$, there exists some $\lambda > 0$ such that $\lambda\phi$ is an ε-robust binarized scoring rule.

It is important to mention that binarized scoring rules have been criticized based on experimental evidence (Selten et al. 1999), although such criticism is not unanimous (Harrison et al. 2013, 2014, 2015).

5.3 *Discrete scoring rules*

A scoring rule is called discrete whenever the subject can only give a report from a finite set $R \subseteq [0,1]$. In this case, the scoring rule becomes a function $S : R \times \Omega \to \mathbb{R}$. We henceforth focus on the most common discrete scoring rule, viz., one where $R = \{\frac{0}{n}, \frac{1}{n}, \ldots, \frac{n}{n}\}$ for an arbitrary $n \in \mathbb{N}$. We call such a scoring rule n-discrete. For instance, a 100-discrete scoring rule is one where the subject is asked to report a percentage.

The practical advantage is that discrete scoring rules are easier to implement, as they can be presented to the subject in the form of a list. On the negative side, the fact that there are only finitely many reports implies that there is no proper, and a fortiori there is no robust scoring rule. So we need to settle for the second best, i.e., to approximate the subject's prior beliefs. The best approximation that we can theoretically achieve with an n-discrete scoring would be to elicit a belief within $1/2n$ from his true prior. That is, whenever the subject reports k/n, we can guarantee that his prior belongs to the interval $[\frac{k}{n} - \frac{1}{2n}, \frac{k}{n} + \frac{1}{2n}]$.

COROLLARY 3. *Fix an arbitrary $n \in \mathbb{N} \setminus \{0\}$. Then the following hold:*

(i) If the cost function is posterior-separable, there exists a $\frac{1}{2n}$-robust n-discrete scoring rule.

(ii) Let $P(\mathcal{K}) = p$, and fix a weakly proper n-discrete scoring rule ϕ. Then, for every $\varepsilon > \frac{1}{2n}$ and every $\delta > 1 - p$, there exists some $\lambda > 0$ such that $\lambda\phi$ is a (ε, δ)-robust n-discrete scoring rule.

The trick to obtain the previous result is to view an n-discrete scoring rule as an $\frac{1}{2n}$-proper scoring rule, characterized by a convex piecewise linear function ϕ with kinks at every point in $\{\frac{1}{2n}, \frac{3}{2n}, \ldots, \frac{2n-1}{2n}\}$. In particular, whenever the subject reports k/n, he will be paid the act that ϕ would yield if a report in the interval $[\frac{k}{n} - \frac{1}{2n}, \frac{k}{n} + \frac{1}{2n}]$ was given.

6. POSTERIOR-SEPARABILITY REVISITED

6.1 *Axiomatic characterization*

Throughout the paper, we have focused on cost functions that satisfy posterior-separability. As we have already mentioned, this specification is supported by recent experimental evidence and is general enough to accommodate usual functions, such as costs that are proportional to the expected decrease in Shannon entropy. As the following result illustrates, posterior-separability also has solid theoretical foundations. In particular, it can be characterized by means of three mild axioms. A similar result has been proven by Zhong (2017) in a somewhat different context, relying on standard properties of mutual information (e.g., Cover and Thomas 2006).

PROPOSITION 1. *The cost function C is posterior-separable if and only if it satisfies:*

(C_1) NORMALIZATION: $C(\hat{\mu}) = 0$ *for all* $\mu \in [0, 1]$.

(C_2) ATTENTION IS COSTLY: $C(\pi) > 0$ *for all* $\pi \in \hat{\Pi}(\mu)$ *and all* $\mu \in [0, 1]$.

(C_3) DYNAMIC CONSISTENCY FOR FULL INFORMATION: $C(\pi^*_\mu) = C(\pi) + \mathbb{E}_\pi(C \circ \pi^*)$ *for all* $\pi \in \Pi(\mu)$ *and all* $\mu \in [0, 1]$.

The first two axioms are quite standard in the literature, postulating that an attention strategy is costly if and only if it carries some information. The third axiom on the other hand is relatively new, postulating that the cost of learning the true state does not depend on the order of collecting information, i.e., only the "acquired information" matters, and not the "process of acquiring it."[5] Intuitively, suppose that the subject (with prior μ) chooses a sequential attention strategy, according to which he first picks some

[5]In a previous version of the paper, we used a stronger dynamic consistency axiom, viz., we postulated that the cost of any attention strategy (not only the one that reveals the state) depends only on the distribution of posteriors, and not on the underlying process that yields this distribution (Tsakas 2019). However, it turns out that the two systems of axioms are equivalent, which is why we present the weaker form of dynamic consistency here.

arbitrary $\pi \in \Pi(\mu)$ (first-period attention strategy) and then conditional on observing each posterior $\nu \in \text{supp}(\pi)$ he picks the most informative attention strategy π^*_ν (second-period attention strategy), implying that the subject learns the state in two steps. Then the total cost that he incurs is equal to the cost of his first-period strategy (viz., $C(\pi)$) plus the expected cost of his second-period strategies (viz., $\mathbb{E}_\pi(C \circ \pi^*) = \int_0^1 C(\pi^*_\nu)\, d\pi$). Dynamic consistency postulates that the cost $C(\pi^*_\mu)$ of directly choosing the most informative attention strategy π^*_μ is equal to the total cost of the aforementioned sequential attention strategy.[6]

It is not difficult to see that posterior-separability is a special case of the class of cost functions that satisfy the basic regularity conditions, viz., Blackwell monotonicity and convexity (Caplin and Dean 2015, De Oliveira et al. 2017). Formally, for two attention strategies $\pi, \rho \in \Pi(\mu)$, we say that π is Blackwell more informative than ρ, and we write $\pi \trianglerighteq \rho$, whenever $\langle f, \pi\rangle \geq \langle f, \rho\rangle$ for every convex function $f : [0,1] \to \mathbb{R}$ (Blackwell 1953). Then C is called *regular*, whenever the following two properties are satisfied:

(C_4) Blackwell Monotonicity: $C(\pi) \geq C(\rho)$ for all $\pi, \rho \in \Pi(\mu)$ with $\pi \trianglerighteq \rho$.

(C_5) Convexity: $C(\lambda\pi + (1-\lambda)\rho) \leq \lambda C(\pi) + (1-\lambda)C(\rho)$ for all $\pi, \rho \in \Pi(\mu)$ and all $\lambda \in (0,1)$.

Proposition 2. *Every posterior-separable cost function is regular.*

Of course the converse is not necessarily true, i.e., regularity does not imply posterior-separability. Intuitively, this is because (C_3) imposes some basic coherency on the costs across different priors, similarly to recent work on dynamic information acquisition (Hébert and Woodford 2017, Zhong 2017), as opposed to (C_4)–(C_5) that put structure on the attention costs given a fixed prior but remain silent on the relationship of the costs across the different priors (Caplin and Dean 2015, De Oliveira et al. 2017). Consider for instance the weaker version of posterior-separability which allows the strictly convex function K to be prior-dependent, i.e., formally, for each $\mu \in [0,1]$ there exists a strictly convex $K_\mu : [0,1] \to \mathbb{R}$ such that $C(\pi) = K_\mu(\mu) - \langle K_\mu, \pi\rangle$ for all $\pi \in \Pi(\mu)$. Clearly, such a cost function satisfies (C_4) and (C_5) without being posterior-separable. We further elaborate on this point below.

6.2 *Weakenings*

We are going to consider two variations of posterior-separability, corresponding to different weakenings of our axioms (C_1)–(C_3). Throughout this section, for simplicity purposes, we will maintain the assumption that the experimenter knows the subject's cost function. Nevertheless, our analysis can be fully generalized to the case where the experimenter holds probabilistic beliefs over the subject's cost functions.

[6]Our notion of dynamic consistency is similar in spirit to the one in the standard characterization of dynamic variational preferences (Maccheroni et al. 2006). Of course in their paper the interpretation is different in that costs are incurred by nature, rather than by the decision maker. Nevertheless their condition—similar to ours—guarantees that that costs are time-consistent. I am indebted to Fabio Maccheroni and Massimo Marinacci for pointing out this connection to me.

6.2.1 *Weak posterior-separability* Let us begin by removing (C_2) from our axiomatic system, i.e., we allow some attention strategies to be cost-free. In this case, costs are characterized by a property that weakens posterior-separability. In particular, we say that a cost function C is weakly posterior-separable if there exists a (weakly) convex function $K : [0,1] \to \mathbb{R}$ satisfying (1) for all $\pi \in \Pi(\mu)$ and all $\mu \in [0,1]$. In other words, every weakly posterior-separable cost function is identified by some $K \in \mathcal{C}$.

PROPOSITION 3. *A cost function is weakly posterior-separable if and only if it satisfies* (C_1) *and* (C_3).

The proof is a straightforward adjustment of the one of Proposition 1. Moreover, it is not difficult to verify that weak posterior-separability also implies regularity, thus extending Proposition 2, i.e., (C_1) and (C_3) suffice for (C_4) and (C_5). Nevertheless, again the converse does not hold, i.e., regularity does not suffice for weak posterior-separability either, for the same reasons why regularity does not imply the strong version of posterior-separability.

Intuitively, defining $[a_K(\mu), b_K(\mu)]$ as the largest interval of μ where K is linear, implies that every attention strategy $\pi \in \Pi(\mu)$ with $\text{supp}(\pi) \subseteq [a_K(\mu), b_K(\mu)]$ is necessarily costless. In fact, the attention strategy that distributes all probability to $a_K(\mu)$ and $b_K(\mu)$ is the most informative attention strategy that the subject can use for free. Obviously, if $a_K(\mu) = b_K(\mu)$ for all $\mu \in [0,1]$, the function K becomes strictly concave, and consequently the cost function becomes posterior-separable in the strong sense. Then we define $\hat{\varepsilon}_K := \sup\{b_K(\mu) - a_K(\mu) | \mu \in [0,1]\}$ as the largest deviation from the prior without the subject incurring any cost.

Now, the fact that K is weakly convex does not alter the logic of our analysis. In particular, Lemma 1 still holds verbatim, i.e., a proper scoring rule ϕ is robust if and only if $K + \phi$ is strictly concave. Obviously, if K is not strictly concave, $K + \phi$ will not be either, implying that we will not be able to find a robust scoring rule. Intuitively, this is because the subject will always prefer to choose an attention strategy that induces updating to $a_K(\mu)$ or $b_K(\mu)$ over the zero attention strategy $\hat{\mu}$. So we resort to our second best solution, i.e., to approximate the subject's prior beliefs.

PROPOSITION 4. *Consider a weakly posterior-separable K, and take an arbitrary proper scoring rule ϕ. Then, for every $\varepsilon > \hat{\varepsilon}_K$, there exists some $\lambda > 0$ such that $\lambda\phi$ is ε-robust.*

The intuition is straightforward. Given that, for any proper ϕ, the subject will always want to use an attention strategy that induces updating from μ to either $a_K(\mu)$ or $b_K(\mu)$, we will choose some $\lambda\phi > 0$ that provides weak enough incentives to at least guarantee that the updated beliefs will not be outside $[a_K(\mu) - \varepsilon, b_K(\mu) + \varepsilon]$.

6.2.2 *Prior-dependent posterior-separability* Let us now attempt to weaken (C_3). In fact, we will depart from (C_3) minimally, by considering for each $\mu \in (0,1)$ some (possibly different) strictly concave $K_\mu : [0,1] \to \mathbb{R}$ such that $C(\pi) = K_\mu(\mu) - \langle K_\mu, \pi \rangle$ for every $\pi \in \Pi(\mu)$. Let us call a cost function that satisfies the previous condition, prior-dependent posterior-separable. Such cost functions have also appeared in the literature

(Caplin et al. 2019), and obviously satisfy regularity (i.e., Blackwell monotonicity and convexity). Nevertheless, the relationship across costs for different priors can be quite arbitrary, as we do not impose any restrictions on the relationship between K_μ and K_ν for two different $\mu, \nu \in [0,1]$. This is exactly why we cannot always elicit the subject's prior beliefs. In fact, often we cannot even approximate the prior beliefs, as illustrated below.

EXAMPLE 6. Fix some strictly concave K, and let $K_\mu(\nu) := \mu K(\nu)$, implying that K_μ becomes arbitrarily flat for priors close to 0. Hence, for any proper scoring rule ϕ, there exists some $\tilde{\mu} \in (0,1)$ such that the optimal attention strategy at every $\mu < \tilde{\mu}$ is the most informative one, implying that there is not even an ε-robust scoring rule, for any $\varepsilon < 1$. $\Diamond$

The previous example also illustrates that regularity alone does not put enough structure to guarantee that we can elicit the subject's prior, even when it is augmented with additional assumptions (e.g., prior-dependent posterior-separability and continuity). Hence, in some sense the most essential part of posterior-separability (viz., our dynamic consistency axiom, (C_3)) seems to be rather tight.

7. DISCUSSION

7.1 *Alternative methodologies*

As we have already discussed, our theory is useful for eliciting population beliefs. Of course, this is not the only methodology that can be used for this purpose. So, let us provide a comparison between robust scoring rules and such alternative methods.

For starters, note that traditional surveys are typically not incentivized, mainly due to practical reasons, e.g., providing monetary incentives can sometimes be very costly (for an exception to this rule, see Grisley and Kellogg 1983). Nevertheless, as we have already mentioned, there is strong evidence that supports the use of monetary incentives (Harrison and Rütstrom 2008, Harrison 2014). Within the class of incentivized mechanisms, one can find widely-used methods such as prediction markets (Hanson 2003). While a prediction market is an incentive compatible mechanism, it suffers from two specific shortcomings that lead to biased estimates of population beliefs. First, the incentives are formed endogenously as the prices of the traded asset fluctuate over time, implying that the experimenter cannot intervene to weaken them in order to guarantee that information will not be acquired. Second, traders in prediction markets do not typically form a representative sample of the population. Of course, we should recognize that the whole aim of prediction markets is to obtain good forecasts, rather than unbiased estimates of population beliefs.

Let us also present two alternative approaches, inspired by the rational inattention literature, which can be used to estimate population beliefs.[7] First, following either the revealed-preference approach (Caplin and Dean 2015, Chambers et al. 2018) or the axiomatic approach (De Oliveira et al. 2017, Ellis 2018), instead of trying to supress information acquisition, the experimenter can just let it occur and identify the prior beliefs

[7]I would like to thank one of the referees for suggesting these two alternatives.

from choice data. A major drawback of this approach is that it requires a large number of choice observations for each individual, while our method works with a single observation. Of course, this is due to the fact that our method requires some ex ante knowledge of the cost function, as opposed to this alternative which identifies the costs simultaneously with the prior beliefs. A second drawback of this approach is that it relies on a common stationarity assumption, namely that the subject does not carry his updated beliefs from one decision problem to the next, thus maintaining the same prior across observations.

A second alternative method for estimating population beliefs is based on eliciting individual beliefs with a proper (but not necessarily robust) scoring rule from a large sample of homogeneous individuals (i.e., individuals with the same prior, the same vNM preferences and the same cost functions) and then average across individuals.[8] While this approach theoretically works smoothly, it crucially relies on two assumptions that make it relatively less appealing than robust scoring rules: first, it requires a large sample of homogeneous subjects which is difficult to recruit, and second it implicitly assumes that the realizations of the respective attention strategies are independent across subjects.

7.2 *Eliciting multinomial prior beliefs*

Throughout the paper, we have focused on binary state spaces, thus eliciting the probability of a single event. Now suppose that the subject has a (multinomial) prior belief $\mu_0 \in \Delta(\Omega)$ that the experimenter would like to elicit, where Ω is an arbitrary finite state space. The technical difficulty with directly extending Theorem 1 to this richer environment lies on the extension of Lemma 2 to higher-dimension euclidean spaces not being straightforward. In particular, it is not clear whether for every strictly convex function $f : \Delta(\Omega) \to \mathbb{R}$ there exists a convex and subdifferentiable $\phi : \Delta(\Omega) \to \mathbb{R}$ such that $f - \phi$ is convex. Intuitively, this is because all the directional derivatives of ϕ must decrease slower than those of f, and it is not clear how this could be achieved as the directional derivatives are not independent. In the unidimensional case on the other hand, there is a single derivative, which we compose with a Lipschitz function with constant less than 1 thus obtaining a ϕ with slower rate of change (see proof of Lemma 2), which constitutes the main step for constructing our robust scoring rule.

Nevertheless, for practical purposes, this problem is of minor concern. Indeed, on the one hand, whenever the strictly concave function $K : \Delta(\Omega) \to \mathbb{R}$ is subdifferentiable at the boundary of $\Delta(\Omega)$, this lemma is not needed and our Theorem 1 holds verbatim for any finite Ω. Furthermore, Theorem 2 also holds verbatim for an arbitrary finite state space Ω. Hence, our assumption on Ω being binary is essentially without loss of generality.

7.3 *Continuity of the cost function*

Throughout the paper, we have considered exclusively continuous cost functions C, having pointed out that this is without loss of generality. Let us elaborate on why this

[8]This approach was independently suggested by Jakub Steiner who I would also like to thank.

is the case. First, since we focus on (weakly) posterior-separable cost functions, K is (weakly) concave. Hence, it is continuous in the interior $(0,1)$, and possible discontinuities could only be encountered at the boundaries. Take some concave function $\tilde{K}:(0,1)\to\mathbb{R}$, and consider its continuous extension $\hat{K}:[0,1]\to\mathbb{R}$ and any other concave extension $K:[0,1]\to\mathbb{R}$. Then it is straightforward to verify that for every $\pi\in\Delta([0,1])$ we obtain $K(\mu)-\langle K,\pi\rangle\geq\hat{K}(\mu)-\langle\hat{K},\pi\rangle$. Hence, any scoring rule which is robust (resp., ε-robust) given the continuous cost $\hat{K}$ will also be robust (resp., ε-robust) given K.

Appendix A: Proofs of Section 3

Proof of Lemma 1. For arbitrary $\pi\in\hat{\Pi}(\mu)$ and $\mu\in(0,1)$, it is the case that

$$\begin{aligned}V_\phi(\pi)&=B_\phi(\pi)-C(\pi)\\&=\big(\langle\phi,\pi\rangle-\phi(\mu)\big)-\big(K(\mu)-\langle K,\pi\rangle\big)\\&=\langle K+\phi,\pi\rangle-(K+\phi)(\mu).\end{aligned}$$

Then, obviously $V_\phi(\pi)<0$ for all $\pi\in\hat{\Pi}(\mu)$ and all $\mu\in(0,1)$, if and only if, $K+\phi$ is strictly concave, which completes the proof. □

Proof of Lemma 2. If f is subdifferentiable in $[0,1]$ then the result follows trivially by setting $\phi:=f$. Therefore, we assume that there exists $x\in\{0,1\}$ such that the subderivative

$$\partial f(x):=\big\{t\in\mathbb{R}:f(y)\geq f(x)+t(y-x)\text{ for all }y\in[0,1]\big\}$$

is empty.

Step 1: For each $x\in[0,1]$, define the left $a_x:=f'_-(x)$ and right $b_x:=f'_+(x)$ derivative, respectively. We adopt the notational convention that $a_0=-\infty$ and $b_1:=\infty$. It follows from (strict) convexity of f that $\partial f(x)=[a_x,b_x]$, with $a_x=b_x$ whenever f is differentiable at x. Moreover, ∂f is strictly increasing, i.e., $x<y$ if and only if $a_x\leq b_x<a_y\leq b_y$. Obviously, f is subdifferentiable if and only if $-\infty<b_0<a_1<\infty$, in which case we simply set $\phi:=f$, as we have already mentioned above. Hence, we henceforth focus on the case where $\partial f(x)=\varnothing$ for some $x\in\{0,1\}$, i.e., $b_0=-\infty$ or $a_1=\infty$. Let $x_0\in[0,1]$ be the unique minimizer of f, and define the strictly increasing function $F:[0,1]\to\overline{\mathbb{R}}$ as follows: $F(x):=a_x>0$ for all $x\in(x_0,1]$, $F(x):=b_x<0$ for all $x\in[0,x_0)$, and $F(x_0)=0$.

Step 2: Since f is continuous in a closed interval, it is also absolutely continuous and, therefore, by the fundamental theorem of calculus, F is Lebesgue integrable and

$$f(x)=f(0)+\int_0^x F(t)\,dt. \tag{A.1}$$

Take a strictly increasing Lipschitz function $h:\overline{\mathbb{R}}\to[-1,1]$ (with Lipschitz constant $c\leq1$), and let $\Phi:=h\circ F$. Since F is Lebesgue integrable, so is Φ. Thus, we can define $\phi:[0,1]\to\mathbb{R}$ by

$$\phi(x):=f(0)+\int_0^x\Phi(t)\,dt. \tag{A.2}$$

Since Φ is strictly increasing, ϕ is strictly convex and, therefore, subdifferentiable in $(0,1)$. Moreover, since Φ takes values in $[-1,1]$, it is the case that $\int_0^x \Phi(t)\,dt \geq -2x$, implying that $\phi(x) \geq \phi(0) - 2x$ for every $x \in [0,1]$, i.e., ϕ is subdifferentiable at 0. We prove identically that ϕ subdifferentiable also at 1, implying that it is subdifferentiable in $[0,1]$.

Step 3: Let us finally prove that $f - \phi$ is convex. Consider arbitrary $0 \leq x_1 < x_2 \leq 1$. Since h is Lipschitz with constant $c \leq 1$, it is the case that $F(x_2) - F(x_1) \geq \Phi(x_2) - \Phi(x_1)$, implying that $F - \Phi$ is increasing. Moreover, by (A.1) and (A.2), we obtain $(f-\phi)(x) = \int_0^x (F(t) - \Phi(t))\,dt$, implying that $f - \phi$ is convex, which completes the proof. □

Proof of Theorem 1. For arbitrary $a \in \mathbb{R}$ and $b \in (0,1)$, define the strictly convex function $f := a - bK$. Then, by Lemma 2, there exists some strictly convex and subdifferentiable function ϕ, such that $f - \phi$ is convex. By strict convexity and subdifferentiability of ϕ, it follows that ϕ is a proper scoring rule. Finally, notice that $K + \phi$ is strictly concave, as it is the sum of a strictly concave function (viz., $K + f$) and a concave function (viz., $\phi - f$). Therefore, by Lemma 1, the scoring rule ϕ is robust. □

Appendix B: Proofs of Section 4

Lemma B.1. *For every weakly proper scoring rule ϕ, the function ε_ϕ is upper semicontinuous.*

Proof. Fix an arbitrary scoring rule ϕ, and take an arbitrary sequence $(K_t)_{t=1}^\infty$ in $\mathcal{K}$ converging to some $K_0 \in \mathcal{K}$ in the topology induced by the sup norm.

Step 1: By $K_t - K_0 = \psi_t - \psi_0$, we obtain $\psi_t \to \psi_0$, where as usual $\psi_t := K_t + \phi$. That is, for every $\delta > 0$ there exists some $t_\delta \in \mathbb{N}$ such that $\|\psi_t - \psi_0\|_\infty := \sup_{\mu\in[0,1]} |\psi_t(\mu) - \psi_0(\mu)| < \delta$ for every $t > t_\delta$. Now, for every $t \in \mathbb{N}$, take the concave closure $\overline{\psi}_t(\mu) := \sup_{\pi\in\Pi(\mu)} \langle \psi_t, \pi\rangle$, and observe that

$$\begin{aligned}
\left|\overline{\psi}_t(\mu) - \overline{\psi}_0(\mu)\right| &= \left| \sup_{\pi\in\Pi(\mu)} \langle \psi_t, \pi\rangle - \sup_{\pi\in\Pi(\mu)} \langle \psi_0, \pi\rangle \right| \\
&\leq \left| \sup_{\pi\in\Pi(\mu)} \langle \psi_t - \psi_0, \pi\rangle \right| \\
&\leq \sup_{\pi\in\Pi(\mu)} \left|\langle \psi_t - \psi_0, \pi\rangle\right| \\
&\leq \sup_{\pi\in\Pi(\mu)} \left\langle |\psi_t - \psi_0|, \pi\right\rangle.
\end{aligned}$$

Then, by the definition of the sup norm, we obtain

$$\begin{aligned}
\|\overline{\psi}_t - \overline{\psi}_0\|_\infty &\leq \sup_{\mu\in[0,1]} \sup_{\pi\in\Pi(\mu)} \left\langle |\psi_t - \psi_0|, \pi\right\rangle \\
&\leq \sup_{\mu\in[0,1]} |\psi_t - \psi_0| \\
&= \|\psi_t - \psi_0\|_\infty,
\end{aligned}$$

implying that $\overline{\psi}_t \to \overline{\psi}_0$.

STEP 2: Fix an arbitrary $\mu \in (0,1)$. For each $t \in \mathbb{N}$, define $[a_t(\mu), b_t(\mu)] := J_{\psi_t}(\mu)$ like in (3). Then we define $a_t^*(\mu) := r_\phi^-(a_t(\mu))$ and $b_t^*(\mu) := r_\phi^+(b_t(\mu))$, i.e., the interval $[a_t^*(\mu), b_t^*(\mu)]$ is the smallest closed interval that contains all the posteriors that can be optimally reported by the subject when his prior is μ. In other words, $a_t^*(\mu)$ and $b_t^*(\mu)$ are the worst-case scenarios, in terms of distance from μ. Throughout Step 2, for notation simplicity, we omit writing the prior μ.

We will now show that

$$a_0^* \leq \liminf a_t^* \leq \limsup b_t^* \leq b_0^*. \tag{B.1}$$

For every $t \in \mathbb{N}$, the (most dispersed) optimal attention strategy at μ is denoted by $\pi_t \in \Pi(\mu)$ and is distributed over $\{a_t, b_t\}$. Hence, by continuity of ψ_t which follows from continuity of K_t, we obtain $\overline{\psi}_t(\mu) = \langle \psi_t, \pi_t \rangle$ and, therefore, by Step 1,

$$\langle \psi_t, \pi_t \rangle \to \langle \psi_0, \pi_0 \rangle. \tag{B.2}$$

Now suppose—contrary to what we want to show—that $\liminf a_t^* < a_0^*$ or $\limsup b_t^* > b_0^*$. Since $\inf_{k \geq t} a_k^*$ is increasing and $\sup_{k \geq t} b_k^*$ is decreasing in k, there exists a subsequence of (a_t^*, b_t^*), identified by a countable subset $T \subseteq \mathbb{N}$, such that for each $t \in T$ it is the case that $a_t^* < a_0^* - \delta^*$ or $b_t^* > b_0^* + \delta^*$, for some $\delta^* > 0$. The latter implies that $I_\phi(a_t) \cap I_\phi(a_0) = \varnothing$ or $I_\phi(b_t) \cap I_\phi(b_0) = \varnothing$ for all $t \in T$. Let us now consider two cases:

(i) $a_0 > a_0^*$, i.e., a_0 is in the interior or at the upper bound of $I_\phi(a_0)$: Then there is some $\delta_a > 0$ such that $a_t < a_0 - \delta_a$. Likewise, if $b_0 < b_0^*$, there exists $\delta_b > 0$ such that $b_t > b_0 + \delta_b$.

(ii) $a_0 = a_0^*$, i.e., a_0 is at the lower bound of $I_\phi(a_0)$: Then, there exists a neighborhood $B_{\delta_a}(a_0)$ such that ϕ is strictly convex in the left part of the neighborhood, viz., in $\{\nu \in B_{\delta_a}(a_0) : \nu \leq a_0\}$; otherwise, $a_0^* < a_0$ and we are back to the previous case. Hence, $a_t < a_0 - \delta_a$. Likewise, if $b_0 = b_0^*$, there is some $\delta_b > 0$ such that $b_t > b_0 + \delta_b$.

In either case, there exists some strictly positive $\delta < \min\{\delta_a, \delta_b\}$ such that $a_t < a_0 - \delta$ or $b_t > b_0 + \delta$. Now, define the attention strategy $\pi \in \Pi(\mu)$ with $\operatorname{supp}(\pi) = \{a_0 - \delta, b_0 + \delta\}$, and observe that $\langle \psi_0, \pi \rangle < \langle \psi_0, \pi_0 \rangle$, while $\langle \psi_t, \pi \rangle = \langle \psi_t, \pi_t \rangle$ for all $t \in T$. Since, $\langle \psi_t, \pi \rangle \to \langle \psi_0, \pi \rangle$, the latter obviously contradicts (B.2), thus proving (B.1).

STEP 3: Define first $d_t^*(\mu) := b_t^*(\mu) - a_t^*(\mu)$, and subsequently $\varepsilon_t := \sup_{\mu \in [0,1]} d_t^*(\mu)$. Then it is straightforward to verify that $\varepsilon_t = \varepsilon_\phi^{K_t}$. Then, by Step 2,

$$\begin{aligned} d_0^*(\mu) &= b_0^*(\mu) - a_0^*(\mu) \\ &\geq \limsup b_t^*(\mu) - \liminf a_t^*(\mu) \\ &\geq \limsup\big(b_t^*(\mu) - a_t^*(\mu)\big) \\ &= \limsup d_t^*(\mu). \end{aligned}$$

Therefore, it follows directly that $\limsup \varepsilon_t \leq \varepsilon_0$, which completes the proof. □

LEMMA B.2. *Fix a sequence of continuous functions $(\psi_k)_{k=1}^{\infty}$ in $[0,1]$ such that (a) $\psi_{k+1}-\psi_k$ is concave, and (b) there exists some strictly concave ψ_0 such that $\psi_k \to \psi_0$ in the topology induced by the sup norm. For any belief $\mu \in [0,1]$, it is the case that $a_{\psi_k}(\mu) \uparrow \mu$ and $b_{\psi_k}(\mu) \downarrow \mu$.*

PROOF. First, we introduce the simpler notation, $a_k := a_{\psi_k}(\mu)$ and $b_k := b_{\psi_k}(\mu)$. For every $k \in \mathbb{N} \setminus \{0\}$, by definition, there exists some linear function $L_k : [0,1] \to \mathbb{R}$, such that

(Λ_1) $L_k(a_k) = \psi_k(a_k)$ and $L_k(b_k) = \psi_k(b_k)$, and

(Λ_2) $L_k(\nu) \geq \psi_k(\nu)$ for all $\nu \in [0,1]$, with strict inequality for all $\nu \notin [a_k, b_k]$.

STEP 1: We will first prove that a_k is increasing and b_k is decreasing in k. Obviously, if $a_{k+1} = b_{k+1} = \mu$, then it is trivially the case that $a_k \leq \mu \leq b_k$. So, let $a_{k+1} < \mu < b_{k+1}$, and assume—contrary to what we want to prove—that $a_{k+1} < \mu \leq b_k < b_{k+1}$, implying that there exists some $\theta \in (0,1)$ such that $b_k = \theta a_{k+1} + (1-\theta) b_{k+1}$.

By (Λ_1)–(Λ_2) applied for $k+1$, we obtain

$$\psi_{k+1}(b_k) - \big(\theta \psi_{k+1}(a_{k+1}) + (1-\theta)\psi_{k+1}(b_{k+1})\big) \leq 0. \tag{B.3}$$

By concavity of $\psi_{k+1} - \psi_k$, we obtain

$$(\psi_{k+1}-\psi_k)(b_k) - \big(\theta(\psi_{k+1}-\psi_k)(a_{k+1}) + (1-\theta)(\psi_{k+1}-\psi_k)(b_{k+1})\big) \geq 0. \tag{B.4}$$

Combining (B.3) and (B.4) yields

$$\psi_k(b_k) - \big(\theta \psi_k(a_{k+1}) + (1-\theta)\psi_k(b_{k+1})\big) \leq 0. \tag{B.5}$$

But then, by (Λ_1)–(Λ_2) applied for k, we obtain

$$\psi_k(b_k) - \big(\theta \psi_k(a_{k+1}) + (1-\theta)\psi_k(b_{k+1})\big) > 0,$$

which contradicts (B.5). Hence, b_k is decreasing, and likewise we show that a_k is increasing.

STEP 2: For each $k \in \mathbb{N} \setminus \{0\}$, define

$$\Psi_k := \psi_k(\mu) - \big(\theta_k \psi_k(a_k) + (1-\theta_k)\psi_k(b_k)\big),$$

where $\mu = \theta_k a_k + (1-\theta_k) b_k$, also noticing that $\Psi_k \leq 0$ for every $k \in \mathbb{N} \setminus \{0\}$ (by (Ψ_1)–(Ψ_2)). Observe that Ψ_k is increasing in k:

$$\begin{aligned} \Psi_k &\leq \psi_k(\mu) - \big(\theta_{k+1}\psi_k(a_{k+1}) + (1-\theta_{k+1})\psi_k(b_{k+1})\big) && \text{(B.6)} \\ &\leq \psi_{k+1}(\mu) - \big(\theta_{k+1}\psi_{k+1}(a_{k+1}) + (1-\theta_{k+1})\psi_{k+1}(b_{k+1})\big) && \text{(B.7)} \\ &= \Psi_{k+1}, \end{aligned}$$

where (B.6) follows from $[a_{k+1}, b_{k+1}] \subseteq [a_k, b_k]$ (Step 1) combined with (Λ_1)–(Λ_2), while (B.7) follows from $\psi_{k+1} - \psi_k$ being concave.

STEP 3: Since both sequences $(a_k)_{k=1}^{\infty}$ and $(b_k)_{k=1}^{\infty}$ are monotonic in $[0,1]$, they converge to the respective limits $a_k \uparrow a_0^*$ and $b_k \downarrow b_0^*$. Assume—contrary to what we want to prove—that $a_0^* < b_0^*$. Without loss of generality let $\mu \in (a_0^*, b_0^*)$. Then, by $a_k \leq a_0^* < b_0^* \leq b_k$ (Step 1), we obtain

$$\Psi_k \geq \psi_k(\mu) - \big(\theta_0^* \psi_k\big(a_0^*\big) + \big(1-\theta_0^*\big)\psi_k\big(b_0^*\big)\big), \tag{B.8}$$

where $\mu = \theta_0^* a_0^* + (1-\theta_0^*) b_0^*$. Then it is the case that

$$\begin{aligned} 0 &\geq \lim_{k\to\infty} \Psi_k && \text{(B.9)}\\ &\geq \psi_0(\mu) - \big(\theta_0^* \psi_0\big(a_0^*\big) + \big(1-\theta_0^*\big)\psi_0\big(b_0^*\big)\big) && \text{(B.10)}\\ &> 0 && \text{(B.11)} \end{aligned}$$

with (B.9) following from $\Psi_k \leq 0$ combined with the fact that the limit exists (Step 2), (B.10) follows from (B.8) combined with $\psi_k \to \psi_0$, while (B.11) follows from strict concavity of ψ_0. This obviously contradicts ψ_0 being strictly concave, thus completing the proof. □

PROOF OF THEOREM 2. STEP 1: Fix a strictly decreasing sequence $(\lambda_t)_{t\in\mathbb{N}}$ in $(0,1)$ such that $\lambda_t \downarrow 0$, and for each $t \in \mathbb{N}$ define the scoring rule $\phi_t := \lambda_t \phi$. Notice that since ϕ is $\tilde{\varepsilon}$-proper, ϕ_t is also $\tilde{\varepsilon}$-proper for every $t \in \mathbb{N}$. Fix arbitrary $\varepsilon > \tilde{\varepsilon}$ and $K \in \mathcal{K}$, and define

$$T_K := \min\{t \in \mathbb{N} : \phi_t \text{ is } \varepsilon\text{-robust given } K\}.$$

Let us first prove that T_K is well-defined. In particular, we will show that there exists some $T \in \mathbb{N}$ such that ϕ_t is ε-robust given K for all $t > T$.

For each $t \in \mathbb{N}$ and each $\mu \in [0,1]$, define $[a_t(\mu), b_t(\mu)] := J_{\psi_t}(\mu)$ as in (3). Take $\varepsilon^* := \varepsilon - \tilde{\varepsilon}$, and consider some $N \in \mathbb{N}$ such that $1/N < \varepsilon^*$. Then, by Lemma B.2, for every $n \in \{0,\dots,N-1\}$, there exists some $t_n \in \mathbb{N}$ such that

$$b_t\left(\frac{n}{N}\right) < a_t\left(\frac{n+1}{N}\right)$$

for all $t > t_n$. Then define $T^* := \max\{t_n | n \in \{0,\dots,N-1\}\}$, and observe that

$$b_t(\mu) - a_t(\mu) < \frac{1}{N}$$

for all $t > T^*$ and all $\mu \in [0,1]$. The latter implies that

$$\mu - \varepsilon^* < a_t(\mu) \leq b_t(\mu) < \mu + \varepsilon^*. \tag{B.12}$$

By definition, if $\pi \in \arg\max_{\rho\in\Pi(\mu)} V_{\phi_t}(\rho)$ and $\nu \in \operatorname{supp}(\pi)$, then

$$b_t(\mu) \leq \nu \leq a_t(\mu).$$

Moreover, since ϕ_t is $\tilde{\varepsilon}$-proper, it is the case that

$$b_t(\mu) - \tilde{\varepsilon} \leq \nu - \tilde{\varepsilon} \leq \nu + \tilde{\varepsilon} \leq a_t(\mu) + \tilde{\varepsilon}. \tag{B.13}$$

Hence, by combining (B.12) with (B.13), we obtain $I_{\phi_t}(\nu) \subseteq B_\varepsilon(\mu)$, implying that ϕ_t is ε-robust for every $t > T^*$ given K, as required.

Step 2: For each $t \in \mathbb{N}$, define

$$\mathcal{K}_t := \{K \in \mathcal{K} : T_K \leq t\},$$

trivially noticing that $\mathcal{K}_t \subseteq \mathcal{K}_{t+1}$. Moreover, since T_K exists for every $K \in \mathcal{K}$ (by Step 1), it is obviously the case that $\mathcal{K}_t \uparrow \mathcal{K}$. Therefore, by Billingsley (1995, Theorem 2.1), we obtain $P(\mathcal{K}_t) \uparrow p$, implying that for every $\delta > 1 - p$ there is some $T_\delta \in \mathbb{N}$ such that $P(\mathcal{K}_t) \geq 1 - \delta$ for all $t \geq T_\delta$, thus completing the proof. □

Appendix C: Proofs of Section 6

Proof of Proposition 1. Sufficiency. Let C be posterior separable. Then it is obvious that $C(\hat{\mu}) = 0$ for every $\mu \in [0, 1]$, thus proving (C_1). By strict concavity of K, it follows that $C(\pi) = K(\mu) - \langle K, \pi\rangle > 0$ for every $\pi \in \hat{\Pi}(\mu)$ and every $\mu \in [0, 1]$, thus proving (C_2). For an arbitrary $\pi \in \Pi(\mu)$,

$$\begin{aligned} C(\pi) + \mathbb{E}_\pi\big(C \circ \pi^*\big) &= K(\mu) - \langle K, \pi\rangle + \mathbb{E}_\pi\big(K - \langle K, \pi^*\rangle\big) \\ &= K(\mu) - \langle K, \pi\rangle + \langle K, \pi\rangle - \langle K, \pi^*_\mu\rangle \\ &= C\big(\pi^*_\mu\big), \end{aligned}$$

with the first and the third equation following from posterior-separability, and the second one following from the linearity of the expectation $(\mathbb{E}_\pi)$ and the inner product $(\langle K, \cdot\rangle)$. Hence, (C_3) is also proven.

Necessity. Assume that C satisfies (C_1)–(C_3), and let $K : [0, 1] \to \mathbb{R}$ be the cost of learning the state with certainty, i.e., $K(\mu) := C(\pi^*_\mu)$ for each $\mu \in [0, 1]$. Now, for an arbitrary $\pi \in \Pi(\mu)$,

$$\begin{aligned} C(\pi) &= C\big(\pi^*_\mu\big) - \mathbb{E}_\pi\big(C \circ \pi^*\big) \\ &= K(\mu) - \langle K, \pi\rangle, \end{aligned} \tag{C.1}$$

with the first equation following directly from rearranging (C_3), and the second one following from the definition of K. Hence, it suffices to prove that K is strictly concave. Take arbitrary $0 \leq \mu_1 < \mu_2 \leq 1$ and $\theta \in (0, 1)$, and let $\pi_0 \in \hat{\Pi}(\theta\mu_1 + (1 - \theta)\mu_2)$ be the attention strategy that assigns probability θ to μ_1 and probability $1 - \theta$ to μ_2. Then

$$\begin{aligned} K\big(\theta\mu_1 + (1 - \theta)\mu_2\big) &= C(\pi_0) + \theta K(\mu_1) + (1 - \theta)K(\mu_2) \\ &> \theta K(\mu_1) + (1 - \theta)K(\mu_2), \end{aligned}$$

with the equation above following from (C.1), and the inequality following from (C_2). Hence, K is strictly concave, thus completing the proof. □

Proof of Proposition 2. (C_4) Take $\pi, \rho \in \Pi(\mu)$ such that $\pi \trianglerighteq \rho$. Then, by definition since $-K$ is convex, we obtain $-\langle K, \pi\rangle \geq -\langle K, \rho\rangle$. The latter implies $K(\mu) - \langle K, \pi\rangle \geq$

$K(\mu) - \langle K, \rho\rangle$ and, therefore, by posterior-separability, $C(\pi) \geq C(\rho)$, which completes the proof.

(C_5) Take $\pi, \rho \in \Pi(\mu)$ and $\lambda \in (0,1)$. Then we obtain

$$\begin{aligned} C\big(\lambda\pi + (1-\lambda)\rho\big) &= K(\mu) - \big\langle K, \lambda\pi + (1-\lambda)\rho\big\rangle \\ &= K(\mu) - \lambda\langle K, \pi\rangle - (1-\lambda)\langle K, \rho\rangle \\ &= \lambda C(\pi) + (1-\lambda)C(\rho), \end{aligned}$$

implying an even stronger result, i.e., C is linear in $\Pi(\mu)$. □

Proof of Proposition 4. Define $[a_K(\mu), b_K(\mu)]$ as the largest interval of μ where K is linear. Then fix an arbitrary $\varepsilon > \hat{\varepsilon}_K$ and a strictly decreasing sequence $(\lambda_t)_{t\in\mathbb{N}}$ in $(0,1)$ such that $\lambda_t \downarrow 0$, and for each $t \in \mathbb{N}$ define the scoring rule $\phi_t := \lambda_t \phi$. Notice that since ϕ is proper, ϕ_t is also proper for every $t \in \mathbb{N}$. Define $[a_t(\mu), b_t(\mu)] := J_{\psi_t}(\mu)$ like in (3), and observe that ϕ_t is ε-robust if

$$\big[a_t(\mu), b_t(\mu)\big] \subseteq \big[a_K(\mu) - \varepsilon, b_K(\mu) + \varepsilon\big] \tag{C.2}$$

for all $\mu \in (0,1)$.

Note that the sequence of continuous functions $(\phi_t)_{t=1}^{\infty}$ satisfies (a) $\psi_{t+1} - \psi_t$ is concave, and (b) there exists some weakly concave ψ_0 such that $\psi_t \to \psi_0$ in the topology induced by the sup norm. Then, for any belief $\mu \in [0,1]$ it is the case that $a_t(\mu) \uparrow a_K(\mu)$ and $b_t(\mu) \downarrow b_K(\mu)$. The proof follows proceeds exactly the same as the one of Lemma B.2. Finally, similar to the proof of Theorem 2, we obtain that there exists some $t \in \mathbb{N}$ satisfying (C.2) for every $\mu \in (0,1)$, thus completing the proof. □

References

Alaoui, Larbi, Janezic Katharina, and Penta Antonio (2019), "Reasoning about others' reasoning." Barcelona GSE Working Paper 1003. [957]

Alaoui, Larbi and Antonio Penta (2016), "Endogenous depth of reasoning." *Review of Economic Studies*, 83, 1297–1333. [957]

Alaoui, Larbi and Antonio Penta (2018), "Cost-benefit analysis in reasoning." Unpublished paper, Universitat Pompeu Fabra and Barcelona GSE. [957]

Allais, Maurice (1953), "Le comportement de l'homme rationnel devant le risque: critique des postulats et axiomes de l'école américaine." *Econometrica*, 21, 503–546. [959]

Aumann, Robert J. and Michael Maschler (1995), *Repeated Games With Incomplete Information*. MIT Press. [965]

Billingsley, Patrick (1995), *Probability and Measure*. Wiley, New York. [983]

Blackwell, David (1953), "Equivalent comparisons of experiments." *Annals of Mathematical Statistics*, 24, 265–272. [974]

Brier, Glenn W. (1950), "Verification of forecasts expressed in terms of probability." *Monthly Weather Review*, 78, 1–3. [955, 959, 960]

Cabrales, Antonio, Olivier Gossner, and Roberto Serrano (2013), "Entropy and the value of information to investors." *American Economic Review*, 103, 360–377. [957]

Cabrales, Antonio, Olivier Gossner, and Roberto Serrano (2017), "A normalized value for information purchases." *Journal of Economic Theory*, 170, 266–288. [957]

Caplin, Andrew (2016), "Measuring and modeling attention." *Annual Review of Economics*, 8, 379–403. [957, 959]

Caplin, Andrew and Mark Dean (2015), "Revealed preference, rational inattention, and costly information acquisition." *American Economic Review*, 105, 2183–2203. [959, 964, 974, 976]

Caplin, Andrew, Mark Dean, and John Leahy (2019), "Rationally inattentive behavior: Characterizing and generalizing Shannon entropy." Unpublished paper, Department of Economics, Columbia University. [958, 959, 964, 976]

Chambers, Christopher and Nicolas Lambert (2017), "Dynamic belief elicitation." Unpublished paper, Stanford Graduate School of Business. [959]

Chambers, Christopher, Ce Liu, and John Rehbeck (2018), "Costly information acquisition." Unpublished paper. [959, 964, 976]

Clemen, Robert (2002), "Incentive contracts and strictly proper scoring rules." *Test*, 11, 167–189. [959]

Cover, Thomas M. and Joy A. Thomas (2006), *Elements of Information Theory*, second edition. Wiley-Interscience, Hoboken, New Jersey. [973]

De Oliveira, Henrique, Tommaso Denti, Maximilian Mihm, and Kemal Ozbek (2017), "Rationally inattentive preferences and hidden information costs." *Theoretical Economics*, 12, 621–654. [959, 964, 974, 976]

Dean, Mark and Nathaniel Neligh (2019), "Experimental tests of rational inattention." Unpublished paper, Department of Economics, Columbia University. [958, 964]

Ellis, Andrew (2018), "Foundations for optimal inattention." *Journal of Economic Theory*, 173, 56–94. [959, 964, 976]

Gneiting, Tilmann and Adrain E. Raftery (2007), "Strictly proper scoring rules, prediction, and estimation." *Journal of the American Statistical Association*, 102, 359–378. [962]

Good, Irving John (1952), "Rational decisions." *Journal of the Royal Statistical Society, Series B*, 14, 107–114. [955, 959, 960]

Grisley, William and Earl Kellogg (1983), "Farmers' subjective probabilities in northern Thailand: An elicitation analysis." *American Journal of Agricultural Economics*, 65, 74–82. [976]

Hanson, Robin (2003), "Combinatorial information market design." *Information Systems Frontiers*, 5, 107–119. [976]

Harrison, Glenn (2014), "Real choices and hypothetical choices." In *Handbook of Choice Modelling* (Stephane Hess and Andrew Daly, eds.), 236–254, Edward Elgar Publishing, Northhampton, Massachusetts. [955, 976]

Harrison, Glenn, Jimmy Martìnez-Correa, and Todd Swarthout (2013), "Inducing risk neutral preferences with binary lotteries: A reconsideration." *Journal of Economic Behavior and Organization*, 94, 145–159. [972]

Harrison, Glenn, Jimmy Martìnez-Correa, Todd Swarthout, and Eric Ulm (2015), "Eliciting subjective probability distributions with binary lotteries." *Economics Letters*, 127, 68–71. [972]

Harrison, Glenn and Elisabet Rütstrom (2008), "Experimental evidence on the existence of hypothetical bias in value elicitation experiments." In *Handbook of Experimental Economics Results* (Charles Plott and Vernon Smith, eds.), 752–767, Elsevier, Amsterdam, The Netherlands. [955, 976]

Harrison, Glenn W., Jimmy Martínez-Correa, and J. Todd Swarthout (2014), "Eliciting subjective probabilities with binary lotteries." *Journal of Economic Behavior and Organization*, 101, 128–140. [972]

Hébert, Benjamin and Michael Woodford (2017), "Rational inattention with sequential information sampling." NBER Working Paper 23787. [974]

Hossain, Tanjim and Ryo Okui (2013), "The binarized scoring rule." *Review of Economic Studies*, 80, 984–1001. [972]

Kamenica, Emir and Matthew Gentzkow (2011), "Bayesian persuasion." *American Economic Review*, 101, 2590–2615. [965]

Karni, Edi (2020), "A mechanism for the elicitation of second-order belief and subjective information structure." *Economic Theory*, 69, 217–232. [959]

Maccheroni, Fabio, Massimo Marinacci, and Aldo Rustichini (2006), "Dynamic variational preferences." *Journal of Economic Theory*, 128, 4–44. [974]

Maćkowiak, Bartosz, Filip Matějka, and Mirko Wiederholt (2018), "Rational inattention: A disciplined behavioral model." Unpublished paper, CERGE-EI and CEPR. [959]

Manski, Charles (2004), "Measuring expectations." *Econometrica*, 72, 1329–1376. [955]

Martin, Daniel (2017), "Strategic pricing with rational inattention to quality." *Games and Economic Behavior*, 104, 131–145. [964]

Matějka, Filip (2016), "Rationally inattentive seller: Sales and discrete pricing." *Review of Economic Studies*, 83, 1125–1155. [964]

Matějka, Filip and Alisdair McKay (2015), "Rational inattention to discrete choices: A new foundation for the multinomial logit model." *American Economic Review*, 105, 272–298. [964]

McCarthy, John (1956), "Measures of the value of information." *Proceedings of the National Academy of Sciences*, 42, 654–655. [959, 960]

Offerman, Theo, Joep Sonnemans, Gijs van de Kuilen, and Peter Wakker (2009), "A truth serum for non-Bayesians: Correcting proper scoring rules for risk attitudes." *Review of Economic Studies*, 76, 1461–1489. [959]

Savage, Leonard (1971), "Elicitation of personal probabilities and expectations." *Journal of the American Statistical Association*, 66, 783–801. [959, 960]

Schlag, Karl, James Tremewan, and Joël van der Weele (2015), "A penny for your thoughts: A survey of methods for eliciting beliefs." *Experimental Economics*, 18, 457–490. [959, 962]

Schlag, Karl and Joël van der Weele (2013), "Eliciting probabilities, means, medians, variances and covariances without assuming risk neutrality." *Theoretical Economics Letters*, 3, 38–42. [972]

Schotter, Andrew and Isabel Trevino (2014), "Belief elicitation in the laboratory." *Annual Review of Economics*, 6, 103–128. [956, 959]

Selten, Reinhard, Abdolkarim Sadrieh, and Klaus Abbink (1999), "Money does not induce risk neutral behavior, but binary lotteries do even worse." *Theory and Decision*, 46, 211–249. [972]

Shannon, Claude (1948), "A mathematical theory of communication." *Bell System Technical Journal*, 27, 379–423, 623–656. [964]

Sims, Christopher A. (2003), "Implications of rational inattention." *Journal of Monetary Economics*, 50, 665–690. [958, 959, 964]

Sims, Christopher A. (2006), "Rational inattention: Beyond the linear-quadratic case." *American Economic Review*, 96, 158–163. [959, 964]

Steiner, Jakub, Colin Stewart, and Filip Matějka (2017), "Rational inattention dynamics: Inertia and delay in decision-making." *Econometrica*, 85, 521–553. [964]

Tommaso, Denti (2020), "Posterior-separable cost of information." Unpublished paper, Economics Department, Cornell University. [958, 964]

Tsakas, Elias (2019), "Robust scoring rules." Unpublished paper. [973]

Yang, Ming (2020), "Optimality of debt under flexible information acquisition." *Review of Economic Studies*, 87, 487–536. [959, 960, 964]

Zhong, Weijie (2017), "Optimal dynamic information acquisition." Unpublished paper. [958, 964, 973, 974]

Co-editor Thomas Mariotti handled this manuscript.

Manuscript received 6 December, 2018; final version accepted 30 October, 2019; available online 5 November, 2019.