Elicitability*

Y. Azrieli† C. Chambers‡ P.J. Healy* N.S. Lambert§

September 2023

Abstract

An analyst is tasked with producing a statistical study. The analyst is not monitored and is able to manipulate the study. He can receive payments contingent on his report and trusted data collected from an independent source, modeled as a statistical experiment. We describe the information that can be elicited with appropriately shaped incentives, and apply our framework to a variety of common statistical models. We then compare experiments based on the information they enable us to elicit. This order is connected to, but different from, the Blackwell order. Data preferred for estimation are also preferred for elicitation, but not conversely. Our results shed light on how using data as incentive generator in payment schemes differs from using data for statistical inference.

*This version: September 2023. We thank seminar participants at Caltech, the University of Chicago, Cornell, Duke, Harvard/MIT, the University of Iowa, Northwestern, the UK Seminars in Economic Theory, USC, Yale, Zurich, and the audiences of the PARIS Symposium on AI Society and of the Institute for Mathematical and Statistical Innovation.

†Dept. of Economics, Ohio State University, azrieli.2@osu.edu and healy.52@osu.edu.

‡Dept. of Economics, Georgetown University, christopher.chambers@georgetown.edu.

§Dept. of Economics, University of Southern California, lambertn@usc.edu.

1 Introduction

Imagine a principal who tasks an analyst with producing a statistical study. The analyst is able to manipulate the study without being detected and, absent additional incentives, has an interest to do so. For example, the analyst may work with data that are private or not verifiable, may work with proprietary algorithms, and may have a conflict of interest with the principal. Suppose the principal has access to independent, trusted data, which may be in too small quantity to replicate the study, but that can be utilized to make contingent payments for the purpose of incentive provision. Our main goal is to understand the extent to which we can incentivize the truthful reporting of information with such data.

To understand the issues, consider the following example. A regulatory agency asks a pharmaceutical company to assess the efficacy of a new drug as part of an evaluation process. This assessment is done by clinical trials to be performed by the firm itself, and could be falsified or fabricated due to a lack of good monitoring practice.[1] Let the fraction of the population on which the new drug is effective be θ. This fraction is what the agency cares to learn, and what the firm has evaluated. Generally, the firm is not sure about θ. Its knowledge, or belief about θ, takes the form of a probability distribution. If the agency was eventually able to observe θ perfectly and in due time, then classical methods could be employed to reward the firm for telling all it knows about θ and punish every deviation from the truth. It can be done, for example, with the performance score of Matheson and Winkler (1976).

In the present paper, the relevant 'state of nature' is a parameter not directly observed. Instead, the principal collects data that correlate imperfectly with the parameter. As a result, the principal may be unable to solicit the full information the analyst possesses about this parameter. Returning to our example, suppose the agency performs one clinical trial on its own, and by doing so, is able to assess the efficacy of the new drug on just one random individual. Evidently, this piece of data is insufficient to estimate θ with reasonable precision. It is also not enough to incentivize the firm to reveal all it knows about θ: Payments must depend, at most, on the result of the single trial, and any choice from a menu of such contingent payments does not

[1] Although it is hard to detect, there are numerous documented cases of fraud with clinical trials, particularly in cancer medication markets, including fictitious patients, altered laboratory data, fabricated results, or intentional biases in randomization procedures. See, for example, George and Buyse (2015) for a recent account.

reveal any information beyond the believed likelihood of a positive trial. However, it is enough to incentivize the firm to reveal a point estimate of θ. For example, if $x = 1$ for a positive trial and $x = 0$ for a negative one, and if the firm who reports μ gets $1 - (\mu - x)^2$, then the firm maximizes expected payoffs exactly when it reports the truth about the assessed mean of θ.

Our first objective is to describe the information for which truthful reports can be incentivized, to an arbitrary magnitude, in a general environment. We call such information *elicitable*. The relationship between observations and the parameter of the underlying statistical model is specified by an arbitrary statistical experiment (Blackwell, 1951). Our results show that the elicitable information varies widely as a function of the amount of data. In our example, one clinical trial enables the agency to elicit the mean of θ, and with two trials, it also elicits the variance. However, the mode and median, that are other typical point estimates, cannot be elicited with any finite number of trials. The elicitable information also depends critically on the statistical model. Suppose that the outcome of a clinical trial is not binary but real valued, that we account for certain covariates such as age and weight, and that we postulate a Gaussian linear model that links covariates to outcomes. Then, we can incentivize the reporting of the full information by adequately creating payments contingent on a single observation. This ability is lost if we work with a semiparametric linear model that does not specify the shape of the error term. Nonetheless, in this case, with the data contained in four observations, we can still incentivize the reporting of the mean coefficients and their variance-covariance matrix. In addition, the data requirements do not depend on the number of covariates in the model.

Our second objective is to compare datasets—specifically, the experiments that produce them—and investigate the conceptual differences between working with data to estimate model parameters, and working with data to create incentives that elicit information. The Blackwell order is a classical benchmark, whereby an experiment A is more informative than an experiment B if B is a garbling of A. This means that posteriors obtained from the data A supplies are mean-preserving spreads of posteriors obtained from the data B supplies. The Blackwell order is the relevant mean of comparison when the experiment is carried out by a statistician who estimates model parameters and wants to minimize risks (Le Cam, 1996). On the other hand, in an elicitation context, A is preferred to B when the data it produces makes it possible to elicit least as much information than the data issued from B. Although the elicitation

order and the Blackwell order apply to different contexts, our results suggest they are closely related: A preference for A over B in the sense of Blackwell implies the same preference in the sense of elicitation. The converse is false, because the elicitation order makes it possible to do many more comparisons than the Blackwell order: Some type of noise in the data does not affect our ability to offer incentives, while it always impedes learning. In addition, the power of incentives can be maintained when switching from B to A: We demonstrate that the elicitation order coincides with an 'incentive order' that ranks A above B when the incentives that can be implemented with B can also be implemented with A. However, to do so, it may be necessary to expand the range of possible payments. If the range of possible payments is constrained to be nonnegative or bounded within a given interval, then the resulting incentive order is stronger than than the elicitation order but remains weaker than the Blackwell order.

The paper proceeds as follows. The remainder of this section discusses the related literature. Section 2 presents the model. Section 3 provides three main results to describe the information that can be elicited when the principal has access to some given arbitrary experiment, and puts these results to work in several common statistical models. Section 4 investigates the comparison of experiments from the viewpoint of information elicitation and contrasts it with the classical Blackwell comparison of experiments. The appendices include the proofs omitted from the main text.

Related Literature

To understand the connection with the literature, it is helpful to consider the following framework. An agent has a private type $t \in T$. A principal observes a signal $s \in S$ about the agent's type. The type and signal spaces, T and S, are arbitrary, they depend on the context. The joint distribution over types and signals is common knowledge. The agent is first asked to report his type, then the signal realizes and the principal makes a transfer to the agent, $\varphi(\hat{t}, s)$, as a function of the report $\hat{t}$ and the signal s. In incentive compatible transfer schemes, the agent, assumed to be risk neutral, is best off reporting honestly. In the abstract, this framework fits many models of the literature. Our model is also an instance of this framework, whereby the agent is the analyst and his type is the analyst's assessment of the parameter of interest, while the signal represents the data collected by the principal.

The voluminous literature on belief and probability elicitation goes back to Brier

(1950), Good (1952), McCarthy (1956), De Finetti (1962) and Savage (1971) (for a recent survey, see Gneiting and Raftery, 2007). As in our work, the problem being dealt with can be viewed as an instance of the framework above, where the agent's type is a probability distribution over states, and the principal's signal is a state realization drawn according to the agent's type. The main focus of this literature is the design of transfer schemes that offer strict incentives to report truthfully. The fundamental difference with our work is that, in our setup, the parameter of the underlying statistical model, which is the analog of the state in the probability elicitation literature, is not known and cannot be inferred from the principal's signal. For this reason, it is often not possible to find a scheme that offer strict incentives for truthfully reporting the agent's type. A recent strand of literature simplifies communication by restricting the set of feasible reports.[2] In the framework above, this is saying that the message space used to report the type is smaller than the entire type space. In this case, it is not always possible to implement strict incentives for honest reports. These works examine the message spaces for which transfer schemes with strict incentives continue to exist and study these schemes. However, they continue to assume that the full state is observed. The inability to design adequate incentives stems from the impossibility for the agent to express his belief rather than the imperfect knowledge of the principal. Notably, the concept of elicitation in these works differ from ours because we do not constrain the agent's communication. Other works deal with groups of individuals, as in Miller et al. (2005). The object of interest is a random signal privately observed by each member. Incentives are provided by exploiting a form of correlation among signals. Once the correlation structure becomes known, these mechanisms reduce to standard probability elicitation methods.

In the mechanism design paradigm, belief elicitation is central to the problem of surplus extraction. In a seminal paper, Crémer and McLean (1988) provide necessary and sufficient conditions under which an auction can be designed that extracts the full surplus from the bidders when their valuations are correlated in the private value environment. The auction of Crémer and McLean is composed of a belief elicitation part, where each bidder reports, indirectly through her bid, her belief on the distribution over other bidders' valuations, and an allocation part. The key feature is the control of the payments in the belief elicitation part, a point explicitly

[2]See, in particular, Lambert et al. (2008), Gneiting (2011), Abernethy and Frongillo (2012), Frongillo and Kash (2015).

made by McAfee and Reny (1992), who interpret the problem of surplus extraction in general mechanisms as an instance of the framework above.[3] The agent's type is the characteristic of a participant to some mechanism, and the principal's signal can be the characteristic of some other participant assumed to have communicated truthfully. McAfee and Reny argue that the property needed for full surplus extraction is much stronger than the strict incentive compatibility called for in the probability elicitation literature.[4] We differ in our objective and in the richness of our type and signal spaces (this fact is not merely technical but also substantive). Most importantly, the conditions on the joint distribution of (t, s) that are necessary for full surplus extraction are not satisfied in the context of our paper. Indeed, this condition asks that the family of conditional distributions over signals, given the type, *excludes* all convex combinations. On the contrary, in our work this family *includes* all convex combinations. Exporting the conditions of Crémer and McLean or McAfee and Reny to our setup is making the assumption that the principal can always elicit the full information from the analyst—which in most environments mean the principal eventually observes the model parameter—and also rules out the possibility that the analyst may have varying degrees of uncertainty about the parameter—essentially, it demands that the analyst knows the parameter of the underlying statistical model exactly. The latter observation relates to the impossibility of extracting full surplus in auctions whose bidders receive signals of varying informativeness, as demonstrated by Strulovici (2017).

Finally, another line of research examines the possibility to test the knowledge of forecasters (see, in particular, Dawid, 1982, Foster and Vohra, 1998, Olszewski and Sandroni, 2008, Shmaya, 2008). A state—typically, an infinite stream of binary data—is drawn from an unknown data generating process, and a forecaster communicates

[3]For recent developments on the possibility or impossibility of full surplus extraction, see Neeman (2004), Heifetz and Neeman (2006), Barelli (2009), Chen and Xiong (2013), Rahman (2012), and Lopomo et al. (2019). Notably, Fu et al. (2021) extend the original setting of Crémer and McLean to allow the seller to receive signals on the bidders' valuations and remark that these signals are used to construct contingent payments and not for statistical inference.

[4]McAfee and Reny argue that the problem of full surplus extraction reduces to showing that any real function of the agent's type matches the ex-interim expected transfer of some incentive compatible transfer scheme. It is the reason why Crémer and McLean did not use a standard scheme such as a classical quadratic scoring rule, because these schemes offer very limited control on expected transfers. When the property identified by McAfee and Reny is satisfied, a mechanism designer can extract the full rent of the participant by setting a participation fee that exactly offsets his gains. This participation fee is set by a transfer scheme whose ex-interim expected transfers are equal to the negative of the participant's profit function in the original mechanism.

some data generating process. Here the goal is not to provide incentives but rather to design, prove or disprove the existence of statistical tests that, based on the state realization, distinguish the forecaster who knows the true data generating process from those who don't. In most applications we deal with, the statistical model is identified and generates i.i.d. samples, which makes it possible to infer exactly the data generating process from an infinite data stream. Al-Najjar et al. (2010) provide positive results for the case of non-i.i.d. samples.

2 Model

The model features an analyst and a principal. The analyst (e.g., a researcher, an expert, or a firm) performs a statistical study whose content is summarized by the distribution over the parameters of a statistical model. We refer to this distribution as the analyst's belief over the model parameters. The principal (e.g., a manager, a firm, or an agency) would like to elicit information on the analyst's belief. We follow the convention of using the female pronoun for the principal and the male pronoun for the analyst.

The analyst's belief captures his uncertainty about possible parameter values. The statistical model is fixed exogenously and assumed well-specified. Throughout the paper, Θ denotes the set of parameters of interest. The process by which the analyst arrives at a belief is irrelevant and not part of the model; to fix ideas, the analyst may be a Bayesian statistician who starts from some prior over parameters and forms a posterior belief based on laboratory data.

The principal has access to an independent dataset. The data it includes are randomly generated and their distribution is a function of the parameter of the underlying statistical model. This relation is formalized by the concept of statistical experiments, following Blackwell (1951) and Le Cam (1996). A *statistical experiment* is a pair (Y, π); Y is the set of possible outcomes that captures the observables, and π is a Markov kernel that associates, to every parameter value θ, a distribution over outcomes, $\pi(\cdot|\theta)$.[5] Thus, the outcome randomly generated by the experiment includes all the data observed by the principal. Parameter and outcome spaces are assumed to be standard Borel spaces.[6] The goal beneath this assumption is to encompass a

[5]By definition of a Markov kernel, $\theta \mapsto \pi(A|\theta)$ is measurable for every A.

[6]A standard Borel space is a separable completely metrizable topological space—i.e., a Polish

wide range of statistical models, from finite-dimensional parametric models to fully nonparametric ones. The intuition behind our results is best described with finite spaces, which trivially satisfy this assumption.

It is useful to recognize three classes of experiments. An experiment (Y, π) is *categorical* if its outcome space, Y, is finite. It is *identified* when the underlying statistical model is identified, which means the parameter can be inferred from the outcome distribution: $\pi(\cdot|\theta) \neq \pi(\cdot|\theta')$ if $\theta \neq \theta'$. Intuitively, infinitely repeated draws from an identified experiment make it possible to infer perfectly the true parameter. Finally, it is *complete* if, by analogy to statistical theory, the experiment is so that every probability distribution over outcomes can be induced by some belief over model parameters. Formally, for every distribution over outcomes $\mu \in \Delta(Y)$, there exists a belief $p \in \Delta(\Theta)$ such that for all measurable sets of outcomes $A \subseteq Y$,

$$\mu(A) = \int_{\Theta} \pi(A|\theta) p(\mathrm{d}\theta).$$

The principal offers incentives in the form of contingent payments, whereby the analyst is paid as a function of his report and the data observed by the principal after the analyst sent the report. Alternatively, we can think of the principal designing a performance score that the analyst wants to maximize on average. Formally, for a given experiment (Y, π), an elicitation mechanism, or simply *mechanism*, is a mapping $\varphi : \mathcal{R} \times Y \to \mathbf{R}$ that takes as input a report from a set $\mathcal{R}$, together with an outcome randomly generated by the experiment, and outputs a payoff. The analyst cares about expected payoffs. In a direct mechanism the set of reports is the set of possible beliefs over parameters, $\mathcal{R} = \Delta(\Theta)$. A direct mechanism is *incentive compatible* when

$$\mathrm{E}_p[\varphi(p, y)] \geq \mathrm{E}_p[\varphi(q, y)] \qquad \forall p, q \in \Delta(\Theta),$$

where y is the outcome randomly generated by the experiment, and the notation E_p refers to the expectation operator in the case of p being the parameter distribution. There are, implicitly, two levels of randomness in the expectations above, first, a randomization over parameters according to p, and second, a randomization over outcomes according to π, given the draw of the parameter. To simplify notation, when there is no ambiguity, the same symbol (e.g., y or θ) is used for a random element

space—endowed with its Borel σ-algebra.

and a possible realization.

Incentive compatibility is a weak requirement trivially satisfied. The literature typically restricts attention to mechanisms that satisfy strict incentive compatibility: Reporting one's true belief must be the only best response. In the present context, this criterion is not achievable in general, but we can still incentivize the truthful reporting of *some* information. We formalize what we learn from the analyst by means of information partitions (Aumann, 1976). An *information partition* of distributions, $\mathcal{P}$, or simply information for short, is a partition of the space $\Delta(\Theta)$. Two parameter distributions that belong to different members of the partition are said to be *distinguishable under* $\mathcal{P}$. They are indistinguishable otherwise. The concept of elicitability is then defined as follows:

Definition 1. *An incentive-compatible mechanism* φ elicits information $\mathcal{P}$ *when, for any* p, q *distinguishable under* $\mathcal{P}$*, the strict inequality* $\mathrm{E}_p[\varphi(p, y)] > \mathrm{E}_p[\varphi(q, y)]$ *obtains.*

An information partition $\mathcal{P}$ is *elicitable* with a given experiment when a mechanism for this experiment can be designed to elicit $\mathcal{P}$. Of course, a strict inequality alone cannot serve as a constraint on the power of incentives, but once established, it becomes possible to generate incentives of any desired magnitude by shifting and scaling the payoffs accordingly.

3 Elicitable Information

The results of this section deal with elicitable information when the principal carries out a fixed arbitrary experiment. We begin with the general case. In Sections 3.1 and 3.2 we refine the experiment respectively by considering the case of multiple independent observations, and by accounting for observable covariates in the data.

Let us consider an arbitrary experiment (Y, π) whose random outcome is y. Any belief over parameters, $p \in \Delta(\Theta)$, induces a distribution over outcomes, $\lambda_p \in \Delta(Y)$, defined by

$$\lambda_p(A) = \int_\Theta \pi(A|\theta) p(\mathrm{d}\theta),$$

for each measurable set of outcomes A. The value $\lambda_p(A)$ is the mean probability $\mathrm{E}_p[\pi(A|\theta)]$ that $y \in A$. We refer to λ_p as the *mean outcome distribution* associated to belief p. We denote by $\mathcal{P}^\star$ the information partition that captures the mean outcome

distribution, meaning that two parameter distributions p, q are indistinguishable under $\mathcal{P}^\star$ if (and only if) they induce the same outcome distribution, i.e., $\lambda_p = \lambda_q$.

The result below characterizes the information an arbitrary experiment enables us to elicit.

Theorem 1. *An information partition $\mathcal{P}$ is elicitable with (Y, π) if, and only if, $\mathcal{P}$ is coarser than $\mathcal{P}^\star$.*

By coarser and finer, we always mean weakly coarser and weakly finer, respectively. In the sequel, we shall refer to $\mathcal{P}^\star$ as the *maximal information elicitable with* (Y, π).

The proof, in Appendix A, is simple. When payment contingencies are on experiment outcomes, the analyst only cares to assess the outcome distribution, all other information is superfluous. Since this distribution is itself determined by the model parameter, which the analyst is typically unsure about, the analyst's belief over outcomes becomes the mean outcome distribution that results from these two layers of randomization. Eliciting this information is well-known in special cases of sets Y such as finite and finite-dimensional sets. While not possible for arbitrary sets Y, eliciting this information in the more general cases allowed under our assumptions on Y remains possible by a mechanism whose functioning is equivalent to randomizing over a large enough collection of quadratic scoring rules.

One implication of this theorem is that if, for each index i of an arbitrarily rich set, some information partition $\mathcal{P}_i$ is elicitable with experiment (Y, π), possibly using a specific mechanism for each value of i, then the join information is elicitable as well. In other words, the information partitions $\mathcal{P}_i$ are elicitable all at once with a unique mechanism. This observation is instrumental to many examples of our paper.[7]

An information of common interest is point estimates. The two corollaries below give necessary and sufficient conditions for the elicitation of the mean estimate of some real function of the parameter. In these results, and throughout the paper, $\mathcal{G}$ refers to the set of bounded measurable functions from Θ to $\mathbf{R}$. The focus on bounded functions is identical to asking that the expected value of $g(\theta)$ is finite under all beliefs about θ, and so ensures that the task of elicitation is well defined. Saying the mean

[7]This observation owes to the structure we impose on the outcome space. For example, consider an arbitrary space of states of the world and a family of events E_i with $i \in [0, 1]$ whose possible occurrence is eventually observed. We can of course elicit the probability of each E_i separately (e.g., with a quadratic scoring rule) but it is not possible to elicit the probabilities of all E_i's at once with the same mechanism in general. For a detailed discussion of the issues involved, see Chambers and Lambert (2021, p. 398).

of $g(\theta)$ is elicitable with (Y, π) is formally saying that the information partition that captures the mean of $g(\theta)$ is coarser than $\mathcal{P}^\star$, i.e., that $\mathrm{E}_p[g(\theta)] = \mathrm{E}_q[g(\theta)]$ whenever p, q are indistinguishable under $\mathcal{P}^\star$.

Corollary 1. *If $g \in \mathcal{G}$ and there exists $w : Y \to \mathbf{R}$ with $g(\theta) = \int_Y w(y)\pi(\mathrm{d}y|\theta)$, then the mean of $g(\theta)$ is elicitable with (Y, π).*[8]

This corollary is a direct implication of the law of iterated expectations, sometimes referred to as the martingale property. A short formal proof is in Appendix A. The converse does not hold in general, but does hold when the experiment is categorical.

Corollary 2. *If (Y, π) is categorical, $g \in \mathcal{G}$, and the mean of $g(\theta)$ is elicitable with (Y, π), then there exists $w : Y \to \mathbf{R}$ with $g(\theta) = \sum_y w(y)\pi(y|\theta)$.*

Hence, in the categorical case, the mean of $g(\theta)$ is elicitable precisely when there exists an unbiased estimator for $g(\theta)$ taking as input the principal's observations. This second corollary, proved in Appendix A, is an application of the Fundamental Theorem of Duality. The examples below illustrate these results.

Example 1 (The German tank problem). Consider a population that includes an unknown number of units numbered consecutively starting from 1, together with the experiment that consists in observing just one unit drawn at random from this population (each unit has the same probability of a draw).[9] In this example, the underlying statistical model is the uniform discrete distribution. The parameter θ is the population size, the outcome is the numbered unit, thus $\Theta = Y = \{1, 2, \dots\}$, and the Markov kernel is

$$\pi(k|\theta) = \begin{cases} 1/\theta & \text{if } k \leq \theta, \\ 0 & \text{if } k > \theta. \end{cases}$$

With this single data point, it is possible to elicit the full information on the analyst's belief about θ. To see this, we apply Corollary 1 with the functions $g_m : \Theta \to \mathbf{R}$ defined as

$$g_m(\theta) = \sum_{k=1}^{\infty} w_m(k)\pi(k|\theta)$$

[8] We implicitly assume that w is integrable with respect to $\pi(\cdot|\theta)$ for every θ.

[9] The German tank problem is a classical example in statistics about the estimation of the maximum of a discrete uniform distribution, whose original application is the estimation of the production of tanks and other military items manufactured during World War II (Goodman, 1954).

for $m \in \mathbf{N}_+$ and

$$w_m(k) = \begin{cases} 1 & \text{if } k \leq m, \\ -m & \text{if } k = m+1, \\ 0 & \text{if } k > m+1. \end{cases}$$

Corollary 1 says that the mean of every $g_m(\theta)$ is elicitable, and by Theorem 1 these means are elicitable all at once with a same mechanism. Observe that $\mathrm{E}[g_m(\theta)] = \Pr[\theta \leq m]$, since $g_m(\theta) = 0$ if $\theta \geq m+1$ and $g_m(\theta) = 1$ if $\theta \leq m$. Therefore, the c.d.f. of θ, which captures the full information on the analyst's belief, is elicitable.

Example 2 (Poisson model). Consider a population of individuals who borrow money from a bank or an investor. There is a risk of default, and the number of defaults over a reference time period is modeled as a Poisson distribution, whose parameter θ captures the default rate, presumed constant for the term of interest. In this context, the experiment that reveals the number of loan defaults over the reference period has the outcome set $Y = \{0, 1, 2, \dots\}$ and the Markov kernel

$$\pi(k|\theta) = \frac{\theta^k}{k!} e^{-\theta}.$$

With such data, it is also possible to elicit the full information on the analyst's belief. To see this, let us define for each $t \in \mathbf{R}$ the function on outcomes $w_t(k) = (1 + \mathbf{i}t)^k$ where $\mathbf{i}$ is the imaginary unit,[10] and then let

$$g_t(\theta) = \sum_{k=0}^{\infty} w_t(k)\pi(k|\theta) = e^{-\theta} \sum_{k=0}^{\infty} \frac{(\theta + \mathbf{i}t\theta)^k}{k!} = e^{\mathbf{i}t\theta}.$$

We apply Corollary 1 and Theorem 1, and get that the means of all $g_t(\theta)$ are elicitable at once. Observing that $t \mapsto \mathrm{E}[e^{\mathbf{i}t\theta}]$ is the characteristic function of θ, the above experiment elicits the full distribution over θ—and so the full information on the analyst's belief.

[10] Alternatively, we can leverage results on the identification of mixtures of distributions. The fact that mixtures of Poisson distributions are identified in the statistical sense (see p. 245 of Teicher, 1961) implies that one can infer exactly the distribution over Poisson parameters from the induced distribution over outcomes, which the mechanism used in the proof of Theorem 1 elicits indirectly. We will make use of this fact in our example on Gaussian linear models.

Example 3 (Bernoulli model). Let us return to the example raised in the Introduction. There is a large population of individuals—a continuum—and a new drug is effective on an unknown fraction of the population. This fraction is the parameter of interest, and the underlying statistical model is the Bernoulli model. We look at the experiment that consists in performing a single clinical trial on a random individual from the population.

Here, the set of possible parameters is $\Theta = [0, 1]$, and the experiment is described by the binary outcome set $Y = \{0, 1\}$ (outcome 1 if the drug is effective on the individual and outcome 0 otherwise) with the Markov kernel

$$\pi(1|\theta) = \theta,$$
$$\pi(0|\theta) = 1 - \theta.$$

By Theorem 1, the elicitable information is precisely the mean fraction of the population on which the vaccine is effective. No further information can be elicited with a single observation.

3.1 Data Made of Multiple Observations

Suppose the data the principal collects is composed of independent observations from one or more experiments. Specifically, consider n arbitrary experiments (Y_i, π_i), $i = 1, \ldots, n$, which may be identical, and let y_i denote the outcome randomly and independently generated by (Y_i, π_i), conditionally on the model parameter. The principal observes the vector $y = (y_1, \ldots, y_n) \in Y_1 \times \cdots \times Y_n$. The experiment that generates y is a compound experiment, which is the product of the n elementary experiments, written $(Y_1, \pi_1) \otimes \cdots \otimes (Y_n, \pi_n)$.

Of course, Theorem 1 continues to apply to this compound experiment, but it can be more convenient to work with the individual experiments that make the product. Our second result, below, provides sufficient conditions for mean information to be elicitable overall given knowledge of information elicitable with these individual experiments.

Theorem 2. *If, for $i = 1, \ldots, n$, the experiment (Y_i, π_i) elicits the mean of $g_i(\theta)$ with $g_i \in \mathcal{G}$, then the product experiment elicits the mean of $g(\theta)$ with $g = g_1 \times \cdots \times g_n$.*

The proof of Theorem 2 is in Appendix A.

Several implications are worth noting. First, recall that by Theorem 1, with one observation from an experiment, we can elicit the mean probability of each set of outcomes. Combining this fact with Theorem 2, we conclude that collecting multiple independent observations from the same experiment enables us to elicit the higher order moments. Second, for any experiment and any $g \in \mathcal{G}$, if the mean of $g(\theta)$ is elicitable with n observations from this experiment, then the variance of $g(\theta)$ is elicitable with $2n$ observations: Since the variance is written $\mathrm{E}[g(\theta)^2] - \mathrm{E}[g(\theta)]^2$, this fact follows by applying Theorem 2 on the product of the two experiments that each supply n observations. Similarly, if we have two experiments (Y, π_Y) and (Z, π_Z) with which we can elicit the mean of $g_Y(\theta)$ and $g_Z(\theta)$, respectively, then their covariance is elicitable with one observation from each experiment.

The examples below illustrate simple applications of Theorem 2. In Section 3.2, we put the theorem to work for several statistical models.

Example 4 (Elicitation of the expertise level). Suppose the principal collects two independent observations from an arbitrary experiment (Y, π) that is identified. The composite experiment that produces the principal's data is the product $(Y, \pi) \otimes (Y, \pi)$.

We claim that this experiment enables the principal to learn whether the analyst knows the parameter and, when the analyst does know, to elicit the true parameter value. By knowing the parameter, we mean that the analyst is absolutely certain about the parameter value. Put formally, this means that under the information partition that captures the maximal information elicitable by this experiment, the probability measure that puts full mass on a single parameter value is distinguishable from any other belief in $\Delta(\Theta)$. The argument goes as follows.

In the proof of Theorem 1, we have shown the existence of a countable collection of measurable subsets of Y, $\{E_i : i \in \mathbf{N}\}$, such that for any two probability distributions over Y, μ and ν, we have $\mu = \nu$ if and only if for all i, $\mu(E_i) = \nu(E_i)$; that is, outcome distributions are identified on this countable collection of sets. And since the experiment is identified, any parameter value θ_0 is uniquely identified by the values $\pi(E_i|\theta_0)$, $i \in \mathbf{N}$.

Let $g_i(\theta) = \pi(E_i|\theta)$. By Theorem 1, experiment (Y, π) elicits the means of all $g_i(\theta)$. Thus one observation enables the principal to learn the mean of every $g_i(\theta)$, and by Theorem 2, the two observations makes it possible to elicit the variance of every $g_i(\theta)$.

Let $p \in \Delta(\Theta)$ be the analyst's belief. To assess if the analyst knows the true

parameter, the principal can examine the variances elicited. If $\mathrm{var}_p[g_i(\theta)] = 0$ for all i, then, letting $\gamma_i = \mathrm{E}_p[g_i(\theta)]$, we have that, for every i, with p-probability 1, $g_i(\theta) = \gamma_i$. Since there are countably many indices i, we may reverse the order of the quantifiers and get that with p-probability 1, $g_i(\theta) = \gamma_i$ for all i. Hence, there exists θ_0 such that $p(\theta = \theta_0) = 1$: The analyst believes he knows the parameter. And conversely, if the analyst believes the parameter value is θ_0 for sure, then clearly $\mathrm{var}_p[g_i(\theta)] = 0$.

Next, if, after examining the variances, the principal concludes that the analyst knows the true parameter parameter value, here referred to as θ_0, the principal infers $\pi(E_i|\theta_0)$ from the values of the elicited means $\mathrm{E}_p[g_i(\theta)]$. The property of identification mentioned above guarantees that θ_0 may be deduced from the values $\pi(E_i|\theta_0)$, $i \in \mathbf{N}$.

Example 5 (Elicitation of probability density functions). When the analyst is not sure about the parameter but may still possess valuable information, the principal can work with more than two observation to approximate the analyst's estimated density function over parameters.

Consider a categorical experiment (Y, π) that is identified, and let us enumerate the elements of Y from 1 to m. We may, without loss of generality, ask that the parameter $\theta = (\theta_1, \ldots, \theta_m)$ represents the outcome distribution and take for Θ the $(m-1)$-simplex of $\mathbf{R}^m$ (the corresponding model is sometimes referred to as the multinoulli model).

Suppose the researcher's belief on θ is captured by a Lipschitz-continuous probability density function f. We will demonstrate the following claim. *With n independent observations, the principal can elicit from the analyst a function $\hat{f}$ that approximates the true density of the analyst according to*

$$\int_\Theta |f(\theta) - \hat{f}(\theta)|^2 \,\mathrm{d}\theta = O(1/n).$$

Notice that the left hand side is the mean integrated squared error commonly used in statistics for the estimation of probability densities.

The proof of this claim builds once again on Theorem 2. We begin by defining two classes of functions, $\mathcal{C}_D$ will be the class of all Lipschitz-continuous densities over Δ^{m-1}, and $\mathcal{C}_P$ will be the class of all polynomials of $m-1$ variables of degree n. Let $P_1, \ldots, P_K$ be an orthonormal basis of $\mathcal{C}_P$ with respect to the inner product on L^2;

that is, a basis that satisfies

$$\int_\Theta P_i P_j = \mathbb{1}\{i = j\},$$

where we use the shorthand integral notation. Such bases are easily constructed using the Gram–Schmidt process of orthonormalization, or even more simply using the tensor product basis from Legendre polynomials.

From Theorem 2, we know that with n independent observations, we can elicit the mean, c_k, of every $P_k(\theta_1, \ldots, \theta_m)$ for $k = 1, \ldots, K$. With this information, we can compute the polynomial

$$P = \sum_k c_k P_k \quad \text{with} \quad c_k = \int_\Theta P_k f.$$

We observe that this polynomial is the orthogonal projection of f on $\mathcal{C}_D$, hence,

$$\int_\Theta (f - P)^2 = \min_{Q \in \mathcal{C}_P} \int_\Theta (f - Q)^2.$$

By the multidimensional version of Jackson's inequality (Theorem 2 of Ganzburg, 1981), there exists a constant γ that depends on f such that

$$\min_{Q \in \mathcal{C}_P} \int_\Theta (f - Q)^2 \leq \gamma / n.$$

We conclude that, from the information that can be elicited truthfully from the researcher, we can infer an approximation $\hat{f}$ of the density f that satisfies

$$\int_\Theta |f - \hat{f}|^2 = O(1/n)$$

where n is the size of the principal's dataset.

3.2 Data with Covariates

Suppose the data observed by the principal includes features or characteristics of the units of observation. The principal now observes a pair (x, y) where $y \in Y$ continues to be called the outcome and $x \in X$ is a covariate, in a general sense, the set

of possible covariate values being arbitrary. The distribution of covariates is known and part of the model. Both X and Y are standard Borel spaces.

The experiment that generates the random pair (covariate,outcome) can be represented as a compound experiment made of two parts. First, a covariate x is randomly generated. Second, an outcome y is randomly generated by an experiment (Y, π_x), where the subscript x captures the possible dependence of the outcome on the covariates. These compound experiments are mixtures of elementary experiments. Formally, a mixture of elementary experiments (Y, π_x), $x \in X$, is an experiment $(X \times Y, \pi)$ that satisfies

$$\pi(A \times B) = \int_A \pi_x(B|\theta)\mu(\mathrm{d}x),$$

where $x \mapsto \pi_x(B|\theta)$ is measurable, and where μ is the probability distribution over covariates.

Of course, Theorem 1 continues to apply, but it can be more convenient to work with the elementary experiments that make the mixture. Our second result links the information elicited by elementary experiments to information that can be elicited by the mixture. It provides sufficient conditions on the information elicited when the principal has access to covariate data.

In the theorem below, we consider any mixture experiment of the form described above, and let $\mathcal{P}(x)$ be an information partition elicitable with (Y, π_x).

Theorem 3. *Let $\mathcal{P}_0$ be any information partition and $\{x_i\}$ be a finite or countably infinite collection of covariates independently drawn from μ. If, with nonzero probability, the join of the partitions in $\{\mathcal{P}(x_i)\}$ is finer than $\mathcal{P}_0$, then $\mathcal{P}_0$ is elicitable with the mixture experiment.*

The proof of Theorem 3 is in Appendix A. The examples that follow illustrate the use of both Theorems 2 and 3 in the context of various statistical models.

Example 6 (Gaussian linear model). This example considers the Gaussian linear model that relates a real-valued dependent variable y (the outcome) to a covariate vector $x = (x^{(1)}, \ldots, x^{(K)}) \in \mathbf{R}^K$ by

$$y = \beta_0 + \beta_1 x^{(1)} + \cdots + \beta_K x^{(K)} + \sigma\varepsilon,$$

where ε is a standard normal error. The model parameter is the $(K+2)$-dimensional vector composed of the vector of coefficients $\beta = (\beta_0, \ldots, \beta_K) \in \mathbf{R}^{K+1}$ and the

variance of the error term, $\sigma^2 \in \mathbf{R}_+$. Let $\theta = (\beta, \sigma^2) \in \Theta$ and let μ be the probability distribution over covariates. Assume the support of μ has a nonempty interior and $\Theta = I^{K+2}$ where I is a compact interval of $\mathbf{R}$.

We claim that under the Gaussian linear model just described, the full information of the analyst is elicitable with just one observation.

The proof of this claim is composed of two simple steps, noting that the relevant experiment is a mixture experiment that generates the data point (x, y). First, fix an arbitrary $x \in X$ and write $h(x) = \beta_0 + \beta_1 x^{(1)} + \cdots + \beta_K x^{(K)}$. A belief (probability distribution) over θ induces a belief over $(h(x), \sigma^2)$. Mixtures of Gaussian distributions are identifiable under the condition that the mean and variance of the individual Gaussians are in compact set (Bruni and Koch, 1985). Therefore, the distribution over y allows the inference of the analyst's belief over $(h(x), \sigma^2)$, and, for every $a, b \in \mathbf{R}$, of the distribution of $ah(x) + b\sigma^2$—that is, we learn the distribution of every one-dimensional projection of $(h(x), \sigma^2)$. Let $\mathcal{P}(x)$ be the information partition that captures the knowledge of each of these distributions. By Theorem 1, this information partition is elicited from the elementary experiment that generates y conditionally on x following the Gaussian model just described.

Second, let $\mathcal{P}_0$ be the information partition associated with the full information on the analyst's belief, $\mathcal{P}_0 = \{\{p\} : p \in \Delta(\Theta)\}$, and let $x_1, x_2, \ldots,$ be an infinite sequence of covariates independently drawn according to μ. By the assumption imposed on μ, with positive probability, this sequence is dense in some open set $O \subset \mathbf{R}^K$. Any finite-dimensional distribution is uniquely determined by its one-dimensional projections (Cramér and Wold, 1936) and hence, the density of the sequence of covariates in O together with the fact that a multivariate c.d.f. is right-continuous implies that with positive probability, the join of the partitions in $\{\mathcal{P}(x_i)\}$ is finer than $\mathcal{P}_0$. The claim then follows from a direct application of Theorem 3.

Example 7 (Semiparametric linear model). This example extends the model of Example 6. We continue to postulate the linear relationship

$$y = \beta_0 + \beta_1 x^{(1)} + \cdots + \beta_K x^{(K)} + \sigma\varepsilon$$

but we no longer restrict the shape of the error term ε. The model parameter θ is now composed of the finite-dimensional vector $(\beta_1, \ldots, \beta_K) \in I^K$ (where I continues

to be a compact interval of $\mathbf{R}$) and a real distribution θ_ε with $\theta_\varepsilon = \beta_0 + \sigma\varepsilon$ such that $\beta_0, \sigma^2 \in I$. It is convenient to work with the parameter space $\Theta = I^K \times \Theta_\varepsilon$, where the parametric part of the model, I^K, is the space of possible values for the coefficients of the independent variables, and the nonparametric part of the model, Θ_ε, is the space of all real distributions whose mean and variance belong to I. This space is rich but satisfies the assumption of Section 2.[11] Let μ be the covariate distribution and assume that for every $\alpha \in \mathbf{R}^K$, with nonzero probability, the covariate x satisfies $\alpha_1 x^{(1)} + \cdots + \alpha_K x^{(K)} \notin \{0, 1\}$. This assumption is weaker than the assumption on μ made in the preceding example.

We claim that under the semiparametric linear model just described, and with one observation, the principal can elicit from the analyst the mean belief of $\beta_0, \ldots, \beta_K$.

To prove the claim, we first take any $x \in X$ and let $h(x) = \beta_0 + \beta_1 x^{(1)} + \cdots + \beta_K x^{(K)}$. Corollary 1 shows that the mean of $h(x)$ is elicitable with the experiment that generates y randomly conditionally on x (this is obtained by using $w(y) = y$). Let $\mathcal{P}(x)$ be the corresponding information partition.

Second, let $\mathcal{P}_0$ be the information partition associated to the mean belief of $\beta_0, \ldots, \beta_K$, and let $x_1, \ldots, x_{K+1}$ be drawn independently at random from μ. For any given belief p over θ, the means of $\beta_0, \ldots, \beta_K$ satisfy the equalities

$$\mathrm{E}_p[\beta_0] + \mathrm{E}_p[\beta_1] x_i^{(1)} + \cdots + \mathrm{E}_p[\beta_K] x_i^{(K)} = \mathrm{E}_p[h(x_i)|x_i] \qquad \forall i = 1, \ldots, K+1.$$

The assumption imposed on μ implies that with positive probability, the matrix

$$\begin{pmatrix} 1 & x_1^{(1)} & \ldots & x_1^{(K)} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{K+1}^{(1)} & \ldots & x_{K+1}^{(K)} \end{pmatrix}$$

has full rank, which, in turn, implies that with positive probability, one can infer the mean of each β_i from knowledge of $\mathrm{E}_p[h(x_i)|x_i]$. And hence, with positive probability, the join of $\mathcal{P}(x_1), \ldots, \mathcal{P}(x_{K+1})$ is finer $\mathcal{P}_0$. We can then apply Theorem 3 to get the claim.

Example 8 (Nonparametric classification). We now look at a nonparametric

[11] Specifically, the parameter space is a Polish space as a product of the Euclidean space I^K and the space Θ_ε endowed with the Wasserstein distance between distributions.

model of classification in which labels take binary values 0 or 1. The set of possible vectors of features, X, is a compact subset of $\mathbf{R}^K$ for $K \geq 1$. In this model, the parameter θ is the likelihood function that maps feature vectors to the range $[0, 1]$, so $\theta(x)$ is the probability of observing label 1 for a unit with feature vector x. We assume $\theta \in \Theta$, with Θ the set of all continuous functions from X to $[0, 1]$. This space satisfies the assumption in Section 2.[12]

Suppose the distribution over covariates, μ, has full support, and consider once again the experiment that generates just one random observation (x, y), where x is the feature and y is the label of a unit randomly drawn. We claim that *under the nonparametric classification model just described, with one observation, the principal can elicit from the analyst the mean label likelihood function.*

First, observe that for an arbitrary dense subset D of X, if for every $x \in D$ the mean of $\theta(x)$ is identical when θ is distributed according to two different beliefs over parameters, p and q, then by continuity of the mapping $x \mapsto \int_\Theta \theta(x)p(\mathrm{d}\theta)$, the mean of $\theta(x)$ continues to be identical for every $x \in X$. That the mapping is continuous follows from the Dominated Convergence Theorem. Second, Example 3 shows that the experiment that generates, at random, a label associated with a given feature vector x elicits the mean of $\theta(x)$. Let $\mathcal{P}(x)$ be the information partition that captures this information.

Finally, let $\mathcal{P}_0$ be the information partition associated with the mean label likelihood function and $x_1, x_2, \ldots,$ be an infinite sequence of covariates independently drawn according to μ. With probability 1, this sequence is dense in X, so by the observation above, with probability 1, the join of the partitions $\mathcal{P}(x_1), \mathcal{P}(x_2), \ldots,$ is a refinement of $\mathcal{P}_0$. We conclude with an application of Theorem 3.

3.3 Discussion

We conclude this section with a discussion on various aspects of our framework.

3.3.1 On Data Requirements for Complete Elicitation

We have argued that the full information on the analyst's belief is generally not elicitable, and we also have given examples when it is. The ability to elicit the full

[12]The space of continuous functions from a compact metric space to a Polish space is Polish by Theorem 4.19 of Kechris (1995).

information depends on the relative size of the space of the analyst's beliefs and the space of the possible outcome distributions.

Suppose the principal collects independent observations from a categorical, identified experiment. We make two claims.

The first claim is that if the parameter space is infinite and the principal gets finitely many data points, it is never possible to elicit the complete information of the analyst (his full belief). To elicit the full information, we need to have an infinite set of observations, and this also makes it possible to infer the true value of the parameter by statistical inference. In terms of data requirements, complete elicitation is as demanding as perfect parameter estimation.

The second claim is that if, on the contrary, the parameter space is finite of size n, then the principal never needs more than $n-1$ observations to elicit the analyst's full information.

We begin by proving the first claim.

The product experiment that generates the principal's data is categorical. Let m be the size of its outcome set, which we enumerate implicitly. Suppose the parameter space is finite of size n with $n > m$, so that $\Theta = \{\theta_1, \ldots, \theta_n\}$. Of course, the argument carries over to all larger parameter spaces. Consider the following linear operator from $\mathbf{R}^n$ to $\mathbf{R}^m$:

$$L(p_1, \ldots, p_n) = \sum_i p_i \pi(\cdot|\theta_i),$$

where a distribution over outcomes takes values in the $(m-1)$-simplex and so is identified with a vector of $\mathbf{R}^m$. This operator transforms a parameter distribution into the mean outcome distribution, for the composite outcome generated by the product experiment.

The maximal information that the product experiment elicits is precisely the mean outcome distribution (Theorem 1). Thus, if we could elicit the full parameter distribution, then any pair of distinct beliefs over parameters would induce two different mean outcome distributions, which is impossible because the dimension of the domain of L exceeds the dimension of the codomain.

To demonstrate the second claim, we note that since the experiment is both identified and categorical, there exists a one-to-one function $g \in \mathcal{G}$ such that the mean of $g(\theta)$ is elicitable. Let $\Theta = \{\theta_1, \ldots, \theta_n\}$. By Theorem 2, with $n-1$ data points, we can elicit the mean of $g(\theta)^k$ for $k = 1, \ldots, n-1$. Consider the following linear

operator from $\mathbf{R}^n$ to $\mathbf{R}^n$:

$$L(p_1, \ldots, p_n) = \sum_i p_i \begin{pmatrix} 1 \\ g^1(\theta_i) \\ \vdots \\ g^{n-1}(\theta_i) \end{pmatrix}.$$

This operator transforms a belief over parameters to the mean distribution of the vector $(1, g^1(\theta), \ldots, g^{n-1}(\theta))$. The transformation is one-to-one (it is represented by a full rank Vandermonde matrix), and thus the analyst's belief over parameters is fully determined by the mean of each $g(\theta)^k$, so that the full belief is elicitable.

3.3.2 On the Elicitation of Point Estimates

Mean point estimates appear to play a special role in our framework. Our first result states that one can always elicit the mean of the outcome distribution, and we have argued by example that the mean belief on some parameters can, sometimes, be elicited with just a few data points. It turns out the other two standard point estimates that are the mode and the median are difficult to elicit in that they demand more data for incentive provision—in fact, they demand as much data as for the elicitation of the full information.

At a general level, this fact is a consequence of the geometry of the maximal information elicited, $\mathcal{P}^\star$. This partition takes the form of parallel, or translated, linear spaces. The information partition of the mean is also composed of parallel linear spaces, but not that of the mode or median. We illustrate the case of the mode below, the case of the median is similar. To simplify matters, we focus on a finite, real parameter space: $\Theta = \{\theta_1, \ldots, \theta_n\} \subset \mathbf{R}$ with $\theta_1 < \cdots < \theta_n$.

Take an arbitrary experiment. We claim that if the mode of θ is elicitable with this experiment, then the full parameter distribution is also elicitable.

By contradiction, suppose the experiment does not elicit the full parameter distribution but that it elicits the mode. Suppose p and p' are two beliefs over θ are indistinguishable under $\mathcal{P}^\star$:

$$\sum_i p(\theta_i)\pi(\cdot|\theta_i) = \sum_i p'(\theta_i)\pi(\cdot|\theta_i).$$

Let $v = p' - p$. Note that $\sum_i v(\theta_i) = 0$. Then note that for all beliefs q and q' such that $q' = q + \alpha v$ for some α, we also have

$$\sum_i q(\theta_i)\pi(\cdot|\theta_i) = \sum_i q'(\theta_i)\pi(\cdot|\theta_i),$$

so that q and q' are also indistinguishable under $\mathcal{P}^\star$.

Thus, if the experiment elicits a mode, then q and q' must share a mode in common. Let q be the uniform distribution, and let $I = \arg\max_i v(\theta_i)$ and $J = \arg\min_j v(\theta_j)$. Note that $I \cap J = \emptyset$. If α is small enough, then $q + \alpha v \in \Delta(\Theta)$. In addition, if $\alpha > 0$ then $q + \alpha v$ has all modes in I, and if $\alpha < 0$ then $q + \alpha v$ has all modes in J. Hence a contradiction.

4 Comparison of Experiments

So far we have focused on one fixed experiment. When data take multiple forms, can be collected in different ways, or simply when various amounts of data can be gathered, the principal has a choice to make about which experiment to carry out.

A standard way to compare experiments is Blackwell's informativeness order. An experiment (Y, π_Y)—that generates random outcome y—is more informative than an experiment (Z, π_Z)—that generates random outcome z—if we can write $z = h(y, \varepsilon)$ for some function h and some independent random noise ε (and where equality is in distribution); that is, the second experiment is a garbling of the first. Intuitively, we can emulate the data supplied by (Z, π_Z) by transforming the data supplied by (Y, π_Y), possibly with the addition of random noise. Blackwell's comparison is relevant when data is used for statistical inference, because the statistician who cares to minimize expected losses will be better off with an experiment that is more informative in the sense of Blackwell.

Rather than estimate model parameters, we utilize experiments as incentive generator to elicit information on these parameters. This goal motivates two alternatives to compare experiment. Our first basis of comparison, explored in Section 4.1, is the information elicitable with a given experiment. Our second basis of comparison, detailed in Section 4.2, is the power of incentives that can be implemented. We will see that these two approaches are essentially equivalent.

Throughout this section, the focus is on experiments that are categorical. By

'experiment' we always mean categorical experiment. We also assume that the outcomes associated with these experiment are enumerated, and leave the enumeration implicit. For any two (enumerated) sets of outcomes Y and Z, a Y-by-Z matrix M stands for a matrix with $|Y|$ rows and $|Z|$ columns, $M(y,z)$ is the entry in row y and column z. When M is a Markov matrix—a matrix with nonnegative entries whose rows sum to 1—we also use the familiar notation $M(z|y)$, interpreted as the transition probability from y to z.

4.1 Comparison based on Elicitability

Let us say that *an experiment dominates another experiment in the sense of elicitation* if every information partition elicitable with the latter is also elicitable with the former. This domination relation defines an 'elicitation' order on experiments, just like Blackwell's informativeness order.[13] Although Blackwell's informativeness order and the elicitation order serve very different purposes, they are closely related.

To see this, consider two experiments (Y, π_Y) and (Z, π_Z). Recall that the Blackwell order can be restated as follows: (Y, π_Y) is more informative than (Z, π_Z) if, and only if, there exists a Y-by-Z Markov matrix M such that

$$\pi_Z(z|\theta) = \sum_{y \in Y} M(z|y)\pi_Y(y|\theta) \qquad \forall z \in Z, \theta \in \Theta. \tag{1}$$

On the other hand, (Y, π_Y) dominates (Z, π_Z) in the sense of elicitation if, and only if, there exists a Y-by-Z matrix M (which need not be Markovian) such that

$$\pi_Z(z|\theta) = \sum_{y \in Y} M(y,z)\pi_Y(y|\theta) \qquad \forall z \in Z, \theta \in \Theta. \tag{2}$$

Indeed, the mean distribution of the outcomes generated by (Z, π_Z) is elicitable with (Z, π_Z) (Theorem 1), and so it is also elicitable with (Y, π_Y) if (Y, π_Y) dominates (Z, π_Z). Corollary 2 then yields the existence of a Y-by-Z matrix M that satisfies Equation (2). Of course, if Equation (2) holds for some matrix M, then (Y, π_Y) dominates (Z, π_Z) in the sense of elicitation.

In the sequel we abuse notation slightly and summarize Equations (1) and (2) by $\pi_Z = \pi_Y M$, even though π_Z and π_Y are not matrices when the parameter space is

[13] Strictly speaking, these are quasi-orders.

infinite. The next lemma summarizes the discussion above.

Lemma 1. *The experiment (Y, π_Y) dominates the experiment (Z, π_Z) in the sense of elicitation if, and only if, $\pi_Z = \pi_Y M$ for some Y-by-Z matrix M.*

Remarkably, the only difference between Blackwell dominance and dominance by elicitation is that M is a Markov matrix in the former, while it is a general matrix in the latter. As an immediate consequence, any experiment that dominates another in the sense of Blackwell also dominates in the sense of elicitation, but not the opposite: The elicitation order is "more complete" than the informativeness order.

To illustrate the difference, let us go back to the Bernoulli model evoked in Example 3. Let $(\{0,1\}, \pi)$ be the experiment that corresponds to a single clinical trial, as in the original example: $\pi(1|\theta) = \theta$, $\pi(0|\theta) = 1-\theta$. In addition, let $(\{0,1\}, \pi')$ be the experiment that adds noise to the clinical trial: $\pi'(1|\theta) = .05 + .9\theta$, $\pi'(0|\theta) = .95 - .9\theta$. In words, with 90% chance, the outcome the experiment generates is the true result of the clinical trial, and with 10% chance, the outcome generated is 0 or 1 with equal probability and is thus entirely uninformative.[14] Clearly, $(\{0,1\}, \pi)$ dominates $(\{0,1\}, \pi')$ in the sense of Blackwell's informativeness, and thus also in the sense of elicitation. The corresponding Markov matrix is

$$M = \begin{pmatrix} .95 & .05 \\ .05 & .95 \end{pmatrix}.$$

Observe that $(\{0,1\}, \pi')$ dominates $(\{0,1\}, \pi)$ in the sense of elicitation, because $\pi = \pi' M^{-1}$, while it is not more informative than $(\{0,1\}, \pi)$—and indeed it is easily verified that M^{-1} is not Markov.

This fact turns out to be quite general: The main difference between elicitation and Blackwell informativeness is that, when data is used in contingent payments as incentive generator, certain types of noise in the data do not impact the ability to offer incentives, whereas in a context of estimation, the statistician cares to avoid every type of noise. Of course, the incentive designer is not immune to all noises. In the example above, no information can be elicited with the experiment that, with 50% chance, reveals the true result of the clinical trial, and with the complementary probability reveals the opposite of the true result.

[14] The observation continues to be the final outcome. It does not include information on whether the outcome comes from the clinical trial or the roulette lottery.

To formalize this idea, we introduce the concept of uniform garbling. An experiment (Y, π') is a *uniform garbling* of another experiment (Y, π) that shares the same outcome space when, for some $\varepsilon \in [0, 1)$, (Y, π') reveals the same outcome[15] as (Y, π) with probability $1 - \varepsilon$, and reveals an outcome drawn uniformly from Y with the complementary probability: $\pi'(y|\theta) = \varepsilon/|Y| + (1 - \varepsilon)\pi(y|\theta)$. We then prove the following result:

Proposition 1. *An experiment (Y, π_Y) dominates another experiment (Z, π_Z) in the sense of elicitation if, and only if, (Y, π_Y) is more informative than a uniform garbling of (Z, π_Z).*

Importantly, this result states that the elicitation order is simply the transitive closure of Blackwell's informativeness order and the order induced by uniform garblings. We may view the elicitation order as refining the Blackwell order with a lot of indifference. The proof of Proposition 1 is in Appendix B.

The uniform distribution is enough to obtain the result, but not at all essential. For example, fix an arbitrary distribution μ on every outcome space, and look at μ-garblings instead, where (Y, π') is a μ-garbling of (Y, π) if π' is as in the case of a uniform garbling except that, instead of the uniform draw, the outcome is drawn according to μ. Then, $\pi' = \pi M$ with M the Markov matrix defined by

$$M(z|y) = \begin{cases} 1 - \varepsilon & \text{if } y = z, \\ \varepsilon\mu(z) & \text{otherwise,} \end{cases}$$

and M is invertible.[16] Hence, (Y, π) and (Y, π') elicit the same information, and it can be verified that the arguments of Proposition 1 continue to hold when μ-garblings are used in place of uniform garblings. [17]

[15] Strictly speaking, outcomes are always assumed to be conditionally independent across experiments, so by 'same' outcome we mean an outcome with an identical distribution conditionally on the parameter.

[16] Observe that $M = (1 - \varepsilon)I + \varepsilon P$ where I is the identity matrix, and P is a square Markov matrix whose rows are identical. Since P is idempotent, it is easily verified that the inverse is $(1 - \varepsilon)^{-1}(I - \varepsilon P)$.

[17] Also note that uniform garblings (or μ-garblings) are not the only types of noisy transformations the principal is immune to, since any experiment (Y, π') that is a garbling of (Y, π), and so satisfies $\pi' = \pi M$ for a Markov matrix M, elicits the same information as (Y, π) as long as M has full rank.

4.2 Comparison based on the Power of Incentives

The order just defined does not deal with the power of incentives. We may wish that, in addition to the ability to elicit more information, the experiment also enables us to keep the same incentives; that is, we may wish to make comparisons based on incentives rather than just the information elicited. To make this comparison, let us say that a (general) mechanism $\varphi : \mathcal{R} \times Y \to \mathbf{R}$ for an experiment (Y, π_Y) is *payoff-equivalent* to another mechanism $\psi : \mathcal{R} \times Z \to \mathbf{R}$ for an experiment (Z, π_Z) if, for all beliefs $p \in \Delta(\Theta)$, and all reports $r \in \mathcal{R}$, we have $\mathrm{E}_p[\varphi(r, y)] = \mathrm{E}_p[\psi(r, z)]$.

An 'incentive' order can then be defined, according to which an experiment (Y, π_Y) dominates another, (Z, π_Z), if the incentives that can be implemented using (Z, π_Z) can also be implemented using (Y, π_Y). Specifically, to any mechanism for (Z, π_Z), there corresponds a payoff-equivalent mechanism for (Y, π_Y). This comparison demands that not only we can elicit the same information with the dominating experiment, but we can do so with the same incentives. It turns out that this incentive order coincides with the elicitation order.

Indeed, recall that if (Y, π_Y) dominates (Z, π_Z) in the sense of elicitation, then there exists a Y-by-Z matrix M with $\pi_Z = \pi_Y M$. Consider a mechanism ψ for (Z, π_Z) and construct a mechanism φ for (Y, π_Y) by

$$\varphi(r, y) = \sum_z \psi(r, z) M(y, z). \tag{3}$$

For every belief over parameters, $p \in \Delta(\Theta)$,

$$\begin{aligned} \mathrm{E}_p[\varphi(r, y)] &= \int_\Theta \sum_y \sum_z \psi(r, z) \cdot M(y, z) \cdot \pi_Y(y|\theta) \cdot p(\mathrm{d}\theta) \\ &= \int_\Theta \sum_z \psi(r, z) \cdot \pi_Z(z|\theta) \cdot p(\mathrm{d}\theta) \\ &= \mathrm{E}_p[\psi(r, z)], \end{aligned}$$

hence, φ and ψ are payoff-equivalent, which proves the next proposition:

Proposition 2. *If (Y, π_Y) dominates (Z, π_Z) in the sense of elicitation, then given any mechanism for (Z, π_Z), there exists a payoff-equivalent mechanism for (Y, π_Y).*

In particular, whenever an experiment strictly dominates another in the sense

of elicitation, a strictly richer set of incentives can be implemented. However, if we normalize the payoffs of the truthful analyst, it is not possible to use this flexibility to augment the power of incentives by a stronger punishment of deviations. Returning to Example 3, with a single clinical trial we elicit the mean parameter value with the incentive compatible mechanism $\varphi(p, y) = 2\mu y - \mu^2$ where $\mu = \mathrm{E}_p[\theta]$ is the mean parameter according to the reported parameter distribution p. The analyst who is offered this mechanism and reports his belief p truthfully earns on average $\mathrm{E}_p[\theta]^2$, and the cost of deviating to some other report q is $(\mathrm{E}_p[\theta] - \mathrm{E}_q[\theta])^2$. If instead we observe the outcomes of 10 clinical trials and continue to elicit the mean parameter, we cannot leverage the additional data to generate a higher cost of deviation if we insist on paying on average $\mathrm{E}_p[\theta]^2$ to the truthful analyst. This observation is formalized in the proposition below. We call a direct mechanism φ continuous if $p \mapsto \varphi(p, y)$ is continuous in the total variation distance.

Proposition 3. *Consider two continuous direct incentive compatible mechanisms, φ and ψ (either for one same experiment or for two different experiments). If φ and ψ generate the same expected payoffs to the truthful analyst, then the expected payoff of an analyst who believes p and reports q is identical under both mechanisms.*

This result owes to the Envelope Theorem. A short proof is in Appendix B.

While it is always possible to maintain expected payoffs with a dominating experiment, the range of the possible payments may need to be enlarged. Intuitively, it may be necessary to compensate for the presence of additional noise in the data that the dominating experiment generates.

In particular, if the principal has a limited liability constraint, so that the payments to the agent have to be nonnegative, then domination in the sense of elicitation does not imply in general that expected payments can be maintained. This can be easily seen in our example of Section 4.1: $(\{0, 1\}, \pi')$ dominates $(\{0, 1\}, \pi)$ in the sense of elicitation, but consider a mechanism ψ for the π experiment in which one of the reports r is such that $\psi(r, 1) = 1$ and $\psi(r, 0) = 0$. For any belief p we have that $\mathrm{E}_p[\psi(r, z)] = \mathrm{E}_p[\theta]$, so in particular $\mathrm{E}_p[\psi(r, z)] = 0$ if p assigns probability 1 to state $\theta = 0$. If a π'-mechanism φ is to be payoff-equivalent to ψ, then for this belief p we must have $0 = \mathrm{E}_p[\varphi(r, y)] = .05\varphi(r, 1) + .95\varphi(r, 0)$, which by nonnegativity implies that $\varphi(r, 1) = \varphi(r, 0) = 0$. But then $\mathrm{E}_p[\varphi(r, y)] \neq \mathrm{E}_p[\psi(r, z)]$ for any other belief p.

In our next result we characterize when it is possible to maintain expected payments, and hence incentives, when limited liability is required.

Proposition 4. *Let (Y, π_Y) and (Z, π_Z) be two experiments. Then the following two statements are equivalent:*
(i) Given any mechanism for (Z, π_Z) with nonnegative payoffs there exists a payoff-equivalent mechanism for (Y, π_Y) also with nonnegative payoffs.
(ii) There exists a nonnegative Y-by-Z matrix M such that $\pi_Z = \pi_Y M$.

Notice how condition (ii) of the proposition, that $\pi_Z = \pi_Y M$ for some nonnegative M, is situated between the conditions that characterize the elicitation order defined in Section 4.1 and Blackwell's informativeness order.[18] In the former M can be any matrix, while in the latter it needs to be a Markov matrix. We have already demonstrated with the above example that condition (ii) is strictly more demanding than the former; we now give an example showing that it is strictly less demanding than the latter.

Let $\Theta = \{\theta_1, \theta_2, \theta_3\}$, $Y = \{1, 2, 3, 4\}$, $Z = \{1, 2, 3\}$. Consider the following Markov kernels in their matrix representations:

$$\pi_Y = \begin{pmatrix} .5 & 0 & 0 & .5 \\ 0 & .5 & 0 & .5 \\ 0 & 0 & .5 & .5 \end{pmatrix} \qquad \text{and} \qquad \pi_Z = \begin{pmatrix} .5 & .5 & 0 \\ .5 & 0 & .5 \\ 0 & .5 & .5 \end{pmatrix}.$$

Observe that (Y, π_Y) is not more informative than (Z, π_Z), yet $\pi_Z = \pi_Y M$ with

$$M = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 0 \end{pmatrix}.$$

However, when the dominating experiment is complete—as defined in Section 2—the limited liability constraints become stronger and the incentive order reduces precisely to Blackwell's informativeness order. This is the content of the next corollary.

Corollary 3. *Let (Y, π_Y) and (Z, π_Z) be two experiments, and suppose that (Y, π_Y) is complete. Then the following two statements are equivalent:*
(i) Given any mechanism for (Z, π_Z) with nonnegative payoffs there exists a payoff-

[18]The condition $\pi_Z = \pi_Y M$ for some nonnegative M also appears in the work of Lehrer and Shmaya (2008). There, it is used to characterize the case where π_Y gives positive expected payoff whenever π_Z does in the presence of an outside option.

equivalent mechanism for (Y, π_Y) *also with nonnegative payoffs.*
(ii) (Y, π_Y) *is more informative than* (Z, π_Z).

The proof of Corollary 3 is in Appendix B.

We now add the further restriction that the range of payments in the payoff-equivalent mechanism for the dominating experiment be contained in the range of payments of the original mechanism. This restriction is relevant when mechanism payoffs are not final payments but rather performance scores taking values on an exogenous scale, or probabilities of getting a fixed reward. In the latter case, realized payoffs are associated to expected payments, or expected utilities. A well-known property of these mechanisms, often used in experimental economics, is to solicit the truth from general expected utility maximizers, independently of the underlying utility function.[19]

Notice first that if (Y, π_Y) is more informative than (Z, π_Z), so that $\pi_Z = \pi_Y M$ for some Markov matrix M, then the payments $\varphi(r, y)$ defined in Equation (3) are convex combinations of $\psi(r, z)$, $z \in Z$, and hence, the range of the possible payments is no larger under φ than under ψ. Therefore, being more informative is a sufficient condition for being able to generate the same incentives without increasing the range of payments; surprisingly, it is not a necessary condition.

To see why, let us go back to the last example with

$$\pi_Y = \begin{pmatrix} .5 & 0 & 0 & .5 \\ 0 & .5 & 0 & .5 \\ 0 & 0 & .5 & .5 \end{pmatrix} \qquad \text{and} \qquad \pi_Z = \begin{pmatrix} .5 & .5 & 0 \\ .5 & 0 & .5 \\ 0 & .5 & .5 \end{pmatrix}.$$

Recall that π_Y is not more informative than π_Z. Consider any mechanism $\psi(r, z)$ with payoffs in $[0, 1]$. We now define a payoff-equivalent mechanism $\varphi(r, y)$ also with payoffs in $[0, 1]$. Fix r and let us write ψ_z instead of $\psi(r, z)$ for any $z \in Z$, and φ_y instead of $\varphi(r, y)$ for any $y \in Y$. By symmetry, we may assume without loss that $\psi_1 \geq \psi_2 \geq \psi_3$. Define $\varphi_4 = 0$ if $\psi_1 + \psi_2 \leq 1$, and $\varphi_4 = \psi_1 + \psi_2 - 1$ otherwise. Next, define $\varphi_1 = \psi_1 + \psi_2 - \varphi_4$, $\varphi_2 = \psi_1 + \psi_3 - \varphi_4$, and $\varphi_3 = \psi_2 + \psi_3 - \varphi_4$.

We leave it to the interested reader to verify that $0 \leq \varphi_y \leq 1$ for every $y \in Y$. To see that $\varphi(r, \cdot)$ gives the same expected payoff as $\psi(r, \cdot)$ for any belief p, suppose

[19] This point is made in Savage (1971) and exploited notably by Roth and Malouf (1979), Grether (1981) and Karni (2009).

first that p assigns probability 1 to state θ_1. The expected payoff under ψ is then $.5(\psi_1 + \psi_2)$, while under φ it is $.5(\varphi_1 + \varphi_4) = .5(\psi_1 + \psi_2)$, i.e., they are equal. Similar calculations apply for states θ_2 and θ_3, implying that expected payoffs are identical under any belief p.

What property then characterizes the case where the range of payoffs in the dominating experiment need not increase to generate the same incentives? It turns out that the answer requires us to think about transitions from signals in the dominating experiment π_Y to *sets of signals* in the dominated experiment π_Z, instead of transitions from signals to signals as in Blackwell's order. We let N stand for a Y-by-2^Z matrix with rows corresponding to signals in Y and columns corresponding to events (subsets) in Z. We then have the following.

Proposition 5. *Let (Y, π_Y) and (Z, π_Z) be two experiments. Then the following two statements are equivalent:*
(i) Given any mechanism for (Z, π_Z) with payoffs in $[0, 1]$ there exists a payoff-equivalent mechanism for (Y, π_Y) also with payoffs in $[0, 1]$.
(ii) There exists a Y-by-2^Z matrix N with entries in $[0, 1]$ such that, for every $A \subseteq 2^Z$ and every θ,

$$\sum_{z \in A} \pi_Z(z|\theta) = \sum_y \pi_Y(y|\theta) N(y, A). \tag{4}$$

The proof of Proposition 5 is in Appendix B. Note that if π_Y is more informative than π_Z, $\pi_Z = \pi_Y M$ for some Markov matrix M, then by defining $N(y, A) = \sum_{z \in A} m(z|y)$ we obtain the equality in (4).

A Proofs of Section 3

A.1 Proof of Theorem 1

We first remark that there exists a sequence of measurable sets of outcomes, $E_1, E_2, \ldots$, such that outcome distributions are identified on the family; that is, for any probability measures μ, ν on Y, $\mu = \nu$ if and only if $\mu(E_i) = \nu(E_i)$ for every $i \in \mathbf{N}_+$. Indeed, since Y is a Polish space, it has a countable base for its topology. The Borel σ-algebra on Y is thus generated by a countable collection of events $\{A_1, A_2, \ldots\}$. For each n, the algebra G_n generated by the truncated collection $\{A_1, \ldots, A_n\}$ is finite, so the algebra G generated by the full collection $\{A_1, A_2, \ldots\}$, which is equal to the union of all G_n, is countable. By the Carathéodory extension theorem, any probability measure on Y is uniquely defined by the probability assigned to every event on G.

Let y be the random outcome generated by (Y, π). A belief p on parameters induces a belief λ_p on outcomes as follows:

$$\lambda_p(A) = \int_\Omega \pi(A|\omega) p(\mathrm{d}\omega) = \mathrm{E}_p[\pi(A|\omega)].$$

Aside from the reported belief, the only other input to any mechanism for (Y, π) is the experiment outcome, so any information that the mechanism elicits must be coarser than the information associated with the full information on the outcome distribution induced by the belief, $\mathcal{P}^\star$.

We describe below a simple mechanism that elicits $\mathcal{P}^\star$.

Fix an arbitrary full-support distribution over the positive integers and consider the following protocol. The analyst first reports a parameter distribution. Then, a positive integer i is drawn at random from the full-support distribution, and the analyst is paid

$$1 - \left(\lambda_p(E_i) - \mathbb{1}_{E_i}(x)\right)^2, \tag{5}$$

where $\mathbb{1}_E$ is the indicator function of the set E.

Suppose the analyst's true belief is captured by the parameter distribution p. A parameter distribution q is a best response if and only if the analyst maximizes the expected value of Equation (5) for every positive integer i—since every i is drawn with positive probability and Equation (5) is the classical Brier score (Brier, 1950). This is equivalent to having $\lambda_p(E_i) = \lambda_q(E_i)$ for every i, which, in turn, is equivalent

to $\lambda_p = \lambda_q$ by the remark above. Hence, any strict best response includes full information on the outcome distribution induced by the true belief. Since the analyst maximizes expected payoffs, the randomized protocol just mentioned is equivalent to the deterministic mechanism φ defined as

$$\varphi(p, x) = 1 - \sum_i \Pr[i] \cdot (\lambda_p(E_i) - \mathbb{1}_{E_i}(x))^2.$$

We conclude there exists a mechanism φ for experiment (Y, π) that elicits $\mathcal{P}^\star$.

A.2 Proof of Corollary 1

Consider the product space $\Theta \times Y$ with its product σ-algebra. Let μ_p be the probability measure over that space induced by a belief p and the experiment (Y, π): For measurable sets $A \subseteq \Theta$ and $B \subseteq Y$,

$$\mu_p(A \times B) = \int_A \pi(B|\theta) p(\mathrm{d}\theta).$$

We continue to denote by λ_p the outcome distribution induced by belief p. Let g and w be as in the statement of Corollary 1, and (θ, y) be the random pair (parameter,outcome) generated by μ_p. By definition, $g(\theta)$ is (a version of) the conditional expectation of $w(y)$ given θ, so by the law of iterated expectations,

$$\mathrm{E}_p[g(\theta)] = \int_\Theta g(\theta) p(\theta) = \int_{\Theta \times Y} w(y) \mu_p(\mathrm{d}(\theta, y)) = \int_Y w(y) \lambda_p(\mathrm{d}y).$$

For any two beliefs p and q that are indistinguishable under $\mathcal{P}^\star$, $\lambda_p = \lambda_q$, so $\mathrm{E}_p[g(\theta)] = \mathrm{E}_q[g(\theta)]$ which implies that the mean of $g(\theta)$ describes a coarser information than $\mathcal{P}^\star$. By Theorem 1, we conclude that the mean of $g(\theta)$ is elicitable with (Y, π).

A.3 Proof of Corollary 2

Let $g \in \mathcal{G}$. As (Y, π) is categorical we can write $Y = \{y_1, \ldots, y_n\}$. By Theorem 1, the maximal information $\mathcal{P}^\star$ that (X, π) elicits is the partition induced by the means of $g_1(\theta), \ldots, g_n(\theta)$ defined by $g_i(\theta) = \pi(y_i|\theta)$.

Let $\mathcal{M}$ be the vector space of finite signed measures on Ω, and let Γ_i be the linear

functional on $\mathcal{M}$ defined by

$$\Gamma_i(\mu) = \int_\Theta g_i(\theta)\mu(\mathrm{d}\theta),$$

and Γ be the linear functional defined by

$$\Gamma(\mu) = \int_\Theta g(\theta)\mu(\mathrm{d}\theta).$$

Let us show the implication

$$\Gamma_1(\mu) = 0, \ldots, \Gamma_n(\mu) = 0 \quad \Longrightarrow \quad \Gamma(\mu) = 0 \tag{6}$$

for all $\mu \in \mathcal{M}$.

The implication is immediate if $\mu = 0$, so suppose $\mu \neq 0$. By the Jordan decomposition theorem, $\mu = \mu^+ - \mu^-$ where μ^+, μ^- are positive measures. If $\int_\Theta g_i(\theta)\mu(\mathrm{d}\theta) = 0$ for all i then $\mu^+(\Theta) = \mu^-(\Theta)$ because $\sum_i g_i = 1$, and so $\mu = \alpha p - \alpha q$ for $p, q \in \Delta(\Theta)$ and $\alpha = \mu^+(\Theta) = \mu^-(\Theta) > 0$. Hence, $\Gamma_i(\mu) = 0$ for all i implies $\mathrm{E}_p[g_i(\theta)] = \mathrm{E}_q[g_i(\theta)]$ for all i. Since $\mathcal{P}^\star$ is the most refined information partition elicitable with (Y, π), if $\mathrm{E}_p[g_i(\theta)] = \mathrm{E}_q[g_i(\theta)]$ for all i, then $\mathrm{E}_p[g(\theta)] = E_q[g(\theta)]$, and so $\Gamma(\mu) = 0$. Therefore, Equation 6 holds for all $\mu \in \mathcal{M}$.

By Equation 6 and the Fundamental Theorem of Duality (see, for example, Theorem 5.91 of Aliprantis and Border, 2006), Γ is in the linear span of $\Gamma_1, \ldots, \Gamma_n$. Therefore, $\Gamma = \sum_i w(y_i)\Gamma_i$ for some function $w : Y \to \mathbf{R}$. Considering an arbitrary parameter value θ and letting μ_θ be the Dirac measure centered on the point θ, it follows that

$$g(\theta) = \Gamma(\mu_\theta) = \sum_i w(y_i)\Gamma_i(\mu_\theta) = \sum_i w(y_i)g_i(\theta).$$

A.4 Proof of Theorem 2

Recall that Θ is a Polish topological space endowed with the Borel σ-algebra, and $\mathcal{G}$ is the set of bounded measurable real-valued functions on Θ. We continue to denote by $\mathcal{M}$ the set of finite signed measures on Θ.

For $\mu \in \mathcal{M}$ and $g \in \mathcal{G}$, consider the bilinear functional $\langle g, \mu \rangle = \int_\Theta g(\theta)\mu(\mathrm{d}\theta)$. Note that with this bilinear functional, $\langle \mathcal{G}, \mathcal{M} \rangle$ forms a dual pair (in the sense of Definition 5.90 of Aliprantis and Border, 2006). We endow $\mathcal{G}$ with the weak topology associated

with this dual pair, whereby the set of continuous linear functionals on $\mathcal{G}$ coincides with $\mathcal{M}$, in the sense that Γ is a continuous linear functional on $\mathcal{G}$ if and only if there exists $\mu \in \mathcal{M}$ such that $\Gamma(g) = \langle g, \mu \rangle$ for all $g \in \mathcal{G}$ (Aliprantis and Border, 2006, Theorem 5.93).

Theorem 2 then follows from Lemmas 2 and 3 below.

In the first lemma, we consider an arbitrary experiment (Y, π). We write $\mathcal{L}$ for the linear span of the functions $\theta \mapsto \pi(A|\theta)$ with $A \subseteq Y$ a measurable set of outcomes, and write $\overline{\mathcal{L}}$ for the closure of $\mathcal{L}$.

Lemma 2. *For any $g \in \mathcal{G}$, the mean of $g(\theta)$ is elicitable with (Y, π) if, and only if, $g \in \overline{\mathcal{L}}$.*

Proof. We first prove that if $g \in \overline{\mathcal{L}}$ then the mean of $g(\theta)$ is elicitable with (Y, π).

Suppose $p, q \in \Delta(\Theta)$ are such that $\mathrm{E}_p[g(\theta)] \neq \mathrm{E}_q[g(\theta)]$. Then $\langle g, p - q \rangle \neq 0$. Because $f \mapsto \langle f, p - q \rangle$ is a continuous linear functional, there exists an open set U that includes g such that for all $f \in U$, $\langle f, p - q \rangle \neq 0$. Since g is in the closure of $\mathcal{L}$, there exists $f \in U$ such that $f \in \mathcal{L}$. Hence, for this function f, $\mathrm{E}_p[f(\theta)] \neq \mathrm{E}_q[f(\theta)]$. By Theorem 1, the mean of $f(\theta)$ is elicitable with (Y, π), thus p and q belong to different members of the maximal information partition $\mathcal{P}^\star$ as defined in Section 3. Hence, the mean of $g(\theta)$ is elicitable.

We now prove the converse, that if the mean of $g(\theta)$ is elicitable, then $g \in \overline{\mathcal{L}}$.

Suppose, by contradiction, that $g \notin \overline{\mathcal{L}}$. The space $\mathcal{G}$ being equipped with the weak topology, it is locally convex—a generalization of normed vector spaces—and by the separating hyperplane theorem for locally convex spaces (Aliprantis and Border, 2006, Theorem 5.79), there exists a continuous linear functional Γ on $\mathcal{G}$ such that $\Gamma(f) = 0$ for all $f \in \overline{\mathcal{L}}$ while $\Gamma(g) \neq 0$. Since the set of continuous linear functionals coincides with $\mathcal{M}$, there is a finite signed measure μ such that for all $f \in \mathcal{G}$,

$$\Gamma(f) = \int_\Theta f(\theta) \mu(\mathrm{d}\theta).$$

Constant functions trivially belong to $\mathcal{L}$ so $\mu(\Theta) = 0$. Then, by the Jordan decomposition theorem, we can write $\mu = \alpha p - \alpha q$, where $p, q \in \Delta(\Theta)$ and $\alpha > 0$. That $\Gamma(g) \neq 0$ implies $\langle g, \alpha p - \alpha q \rangle \neq 0$, or equivalently, $\mathrm{E}_p[g(\theta)] \neq \mathrm{E}_q[g(\theta)]$. And since $\Gamma(f) = 0$ when $f \in \overline{\mathcal{L}}$, we have $\mathrm{E}_p[f(\theta)] = \mathrm{E}_q[f(\theta)]$.

However, when p and q are beliefs such that for every measurable set $A \subseteq Y$, it is the case that $\mathrm{E}_p[\pi(A|\theta)] = \mathrm{E}_q[\pi(A|\theta)]$, then p and q belong to the same partition of

the information partition $\mathcal{P}^\star$ that, by Theorem 1, captures the maximal information elicited by (Y, π), and so we must have $\mathrm{E}_p[g(\theta)] = \mathrm{E}_q[g(\theta)]$, hence a contradiction. $\square$

In this second lemma, for any two linear subspaces of $\mathcal{G}$, V_1 and V_2, we call pointwise product of V_1 and V_2 the set $\{g_1 g_2 : (g_1, g_2) \in V_1 \times V_2\}$ where $g_1 g_2$ simply refers to the pointwise product of g_1 and g_2.

Lemma 3. *Given two linear subspaces of $\mathcal{G}$, the pointwise product of their closures is included in the closure of their pointwise products.*

Proof. For $i = 1, 2$, let V_i be a linear subspace of $\mathcal{G}$ and g_i be in the closure of V_i.

Let $\mathcal{B}$ be the collection of sets of the form

$$\left\{ h \in \mathcal{G} : \forall k, \ \left| \int_\Theta (h(\theta) - g_1(\theta) g_2(\theta)) \nu_k(\mathrm{d}\theta) \right| < \varepsilon \right\}$$

for $\varepsilon > 0$ and finitely many $\nu_1, \ldots, \nu_K \in \mathcal{M}$. Notice that the collection $\mathcal{B}$ forms an open neighborhood base of $g_1 g_2$ by definition of the weak topology. Therefore, showing that $g_1 g_2$ is in the closure of the pointwise product of V_1 and V_2 reduces to showing that for every $B \in \mathcal{B}$, there exists $(f_1, f_2) \in V_1 \times V_2$ such that $f_1 f_2 \in B$.

Fix a set $B \in \mathcal{B}$. First, using the same notation as above and considering the signed measures η_k defined as

$$\eta_k(A) = \int_A g_1(\theta) \nu_k(\mathrm{d}\theta),$$

we can choose $f_2 \in V_2$ such that for all $k = 1, \ldots, K$,

$$|\langle (f_2 - g_2) g_1, \nu_k \rangle| = |\langle f_2 - g_2, \eta_k \rangle| < \frac{\varepsilon}{2},$$

since g_2 is in the closure of V_2. Second, considering the signed measures η_k now defined instead as

$$\eta_k(A) = \int_A f_2(\theta) \nu_k(\mathrm{d}\theta),$$

we can choose $f_1 \in V_1$ such that for all $k = 1, \ldots, K$,

$$|\langle (f_1 - g_1) f_2, \nu_k \rangle| = |\langle f_1 - g_1, \eta_k \rangle| < \frac{\varepsilon}{2},$$

since g_1 is in the closure of V_1.

Hence, for all $k = 1, \ldots, K$,

$$|\langle f_1 f_2 - g_1 g_2, \nu_k \rangle| \leq |\langle (f_1 - g_1) f_2, \nu_k \rangle| + |\langle (f_2 - g_2) g_2, \nu_k \rangle| < \varepsilon,$$

and thus $f_1 f_2 \in B$. □

We now conclude the proof of Theorem 2. We focus on the case $n = 2$, the case $n > 2$ is an immediate generalization. For $i = 1, 2$, let $\mathcal{L}_i$ be the linear span of the functions $\theta \mapsto \pi_i(A|\theta)$ for A a measurable subset of Y_i. Under the assumption of the theorem, the mean of $g_i(\theta)$ is elicitable with (Y_i, π_i), implying by Lemma 2 that g_i is in the closure of $\mathcal{L}_i$. Then, by Lemma 3, $g_1 g_2$ is in the closure of the pointwise product of $\mathcal{L}_1$ of $\mathcal{L}_2$, and so by Lemma 2 again, the mean of $g_1(\theta) g_2(\theta)$ is elicitable with $(Y_1, \pi_1) \otimes (Y_2, \pi_2)$.

A.5 Proof of Theorem 3

For all $x \in X$, let $\psi(\cdot|x)$ be a mechanism for (Y, π_x) that elicits $\mathcal{P}(x)$ and whose payoffs take values in a bounded interval such as $[0, 1]$ (the proof of Theorem 1 includes an instance of such a mechanism). To elicit information with outcomes from the mixture experiment, we consider the compound mechanism φ defined by $\varphi(p, (x, y)) = \psi(p, y|x)$. We show that this mechanism elicits $\mathcal{P}_0$—that is, fixing any two beliefs on parameters, p and q, that are distinguishable under $\mathcal{P}_0$, we show

$$\mathrm{E}_p[\varphi(p, (x, y))] > \mathrm{E}_p[\varphi(q, (x, y))].$$

Let X^∞ be the space of all infinite sequences in X. A generic element of X^∞ is denoted $x^\infty = (x_1, x_2, \ldots)$. (The case of finite sequences is identical and omitted.) We abuse notation and use again the symbol μ for the probability measure over sequences whose elements are drawn independently and identically according to (the original) μ.

Under the assumptions of Theorem 3, there exists $\mathcal{S} \subseteq X^\infty$ with $\mu(\mathcal{S}) > 0$ and such that, for every $(x_1, x_2, \ldots) \in \mathcal{S}$, the join of the partitions in $\{\mathcal{P}(x_i)\}$ is a refinement of $\mathcal{P}_0$.

Observe that

$$\mathrm{E}_p[\varphi(q,(x,y))] = \int_\Theta \int_{X\times Y} \varphi(q,(x,y)) \cdot \pi(\mathrm{d}(x,y)|\theta) \cdot p(\mathrm{d}\theta)$$
$$= \int_{X^\infty} \left(\sum_{i=1}^{\infty} \frac{1}{2^i} \int_\Theta \int_Y \psi(q,y|x_i) \cdot \pi_{x_i}(\mathrm{d}y|\theta) \cdot p(\mathrm{d}\theta) \right) \mu(\mathrm{d}x^\infty)$$

and similarly if p is used in place of q. For every x, by incentive compatibility of $\psi(\cdot|x)$,

$$\int_\Theta \int_Y (\psi(p,y|x) - \psi(q,y|x)) \cdot \pi_x(\mathrm{d}y|\theta) \cdot p(\mathrm{d}\theta) \geq 0,$$

and hence,

$$\mathrm{E}_p[\varphi(p,(x,y))] - \mathrm{E}_p[\varphi(q,(x,y))]$$
$$\geq \int_{\mathcal{S}} \left(\sum_{i=1}^{\infty} \frac{1}{2^i} \int_\Theta \int_Y (\psi(p,y|x_i) - \psi(q,y|x_i)) \cdot \pi_{x_i}(\mathrm{d}y|\theta) \cdot p(\mathrm{d}\theta) \right) \mu(\mathrm{d}x^\infty).$$

Since, for every $(x_1, x_2, \dots) \in \mathcal{S}$, the join of the partitions in $\{\mathcal{P}(x_i)\}$ is a refinement of $\mathcal{P}_0$, there exists j such that $\mathrm{E}_p[\psi(p,y|x_j)] > \mathrm{E}_p[\psi(q,y|x_j)]$ (where the random element in the expectation is y and not x_i). Hence,

$$\sum_{i=1}^{\infty} \frac{1}{2^i} \int_\Theta \int_Y (\psi(p,y|x_i) - \psi(q,y|x_i)) \cdot \pi_{x_i}(\mathrm{d}y|\theta) \cdot p(\mathrm{d}\theta) > 0,$$

and since $\mu(\mathcal{S}) > 0$,

$$\int_{\mathcal{S}} \left(\sum_{i=1}^{\infty} \frac{1}{2^i} \int_\Theta \int_Y \{\psi(p,y|x_i) - \psi(q,y|x_i)\} \pi_{x_i}(\mathrm{d}y|\theta) p(\mathrm{d}\theta) \right) \mu(\mathrm{d}x^\infty) > 0,$$

which implies $\mathrm{E}_p[\varphi(p,(x,y))] - \mathrm{E}_p[\varphi(q,(x,y))] > 0$.

B Proofs of Section 4

B.1 Proof of Proposition 1

Let us start with the 'if' part of the proposition: Suppose (Y, π_Y) is more informative than (Z, π'_Z) in the sense of Blackwell, with (Z, π'_Z) some uniform garbling of (Z, π_Z).

As discussed in Section 4, (Y, π_Y) also dominates (Z, π'_Z) in the sense of in the sense of elicitation, and by transitivity, it is enough to show that (Z, π'_Z) dominates (Z, π_Z) in the sense of elicitation. For some $\varepsilon \in [0, 1)$, $\pi'_Z = ((1-\varepsilon)I + \varepsilon/|Z|J)\pi_Z$, where I is the identity matrix and J is the unit matrix—here both square matrices of dimension $|Z|$. It is immediate that the matrix $((1-\varepsilon)I + \varepsilon/|Z|J)$ is full rank. Thus, (Z, π'_Z) dominates (Z, π_Z) in the sense of elicitation.

The 'only if' part of the proposition makes use of the following lemma.

Lemma 4. *(Y, π_Y) dominates (Z, π_Z) in the sense of elicitation if, and only if, there exists a Y-by-Z matrix M with $\sum_z M(y, z) = 1$ such that $\pi_Z = \pi_Y M$.*

Proof. We have already discussed the 'if' part in Section 4. For the 'only if' part, recall that if (Y, π_Y) dominates (Z, π_Z) in the sense of elicitation, then $\pi_Z = \pi_Y M$ for some Y-by-Z matrix M. Let $M(y) = \sum_z M(y, z)$ and $\widetilde{M}$ by the Y-by-Z matrix with entries $\widetilde{M}(y, z) = M(y, z) + 1 - M(y)$. Note that $\sum_z \widetilde{M}(y, z) = M(y) + 1 - M(y) = 1$. In addition, for all θ,

$$\sum_y \widetilde{M}(y, z)\pi_Y(y|\theta) = \pi_Z(z|\theta) + 1 - \sum_y M(y)\pi_Y(y|\theta) = \pi_Z(z|\theta),$$

because

$$\sum_y M(y)\pi_Y(y|\theta) = \sum_{y,z} M(y, z)\pi_Y(y|\theta) = \sum_z \pi_Z(z|\theta) = 1,$$

and hence, $\pi_Z = \pi_Y \widetilde{M}$. □

Suppose (Y, π_Y) dominates (Z, π_Z) in the sense of elicitation. Let (Z, π_U) be the degenerate experiment that gives an outcome from Z uniformly at random, and let $T = (1-\varepsilon)M + \varepsilon/|Z|J$ for $0 < \varepsilon < 1$, where J is now the unit matrix of dimension $|Y|$-by-$|Z|$. Notice that T has all positive entries if ε is close enough to 1, and notice that all rows of T sum to 1, hence, we can choose T to be Markov. For such ε, we consider the uniform garbling of (Z, π_Z), (Z, π'_Z), for which the probability of the

uniformly drawn outcome is ε. Then, we have

$$\begin{aligned}\pi'_Z &= (1-\varepsilon)\pi_Z + \varepsilon\pi_U \\ &= (1-\varepsilon)\pi_Y M + \varepsilon\pi_Y J \\ &= \pi_Y((1-\varepsilon)M + \varepsilon J) \\ &= \pi_Y T,\end{aligned}$$

and hence, (Y, π_Y) is more informative than (Z, π'_Z).

B.2 Proof of Proposition 3

Fix a mechanism $\varphi : \Delta(\Theta) \times Y \to \mathbf{R}$ for (Y, π_Y), and another mechanism $\psi : \Delta(\Theta) \times Z \to \mathbf{R}$ for (Z, π_Z), both continuous and incentive compatible.

We let V_φ be the value function of φ,

$$V_\varphi(p) = \sup_{q \in \Delta(\Theta)} \mathrm{E}_p[\varphi(q, y)],$$

and similarly for ψ. If the expected payoffs of the truthful analyst are the same under both mechanisms, then $V_\varphi = V_\psi$.

Consider $V : [0,1] \to \mathbf{R}$ defined as

$$V(\alpha) = V_\varphi(\alpha p + (1-\alpha)q) = \sup_{\alpha'} \mathrm{E}_{\alpha p + (1-\alpha)q}[\varphi(\alpha' p + (1-\alpha')q, y)].$$

By the envelope theorem Milgrom and Segal (2002, Theorem 2), V is differentiable and

$$V'(\alpha) = \mathrm{E}_p[\varphi(\alpha p + (1-\alpha)q, y)] - \mathrm{E}_q[\varphi(\alpha p + (1-\alpha)q, y)].$$

The continuity of φ also makes V' continuous, so

$$V'(0) = \mathrm{E}_p[\varphi(q, y)] - \mathrm{E}_q[\varphi(q, y)],$$

and hence,

$$\mathrm{E}_p[\varphi(q, y)] = V_\varphi(q) + \frac{\partial}{\partial \alpha} V_\varphi(\alpha p + (1-\alpha)q).$$

Since $V_\varphi = V_\psi$, it follows that $\mathrm{E}_p[\varphi(q, y)] = \mathrm{E}_p[\psi(q, z)]$.

B.3 Proof of Proposition 4

$(i) \implies (ii)$

Consider a nonnegative mechanism ψ for π_Z where the set of possible reports is $\mathcal{R} = Z$, and such that $\psi(z, z') = 1$ if $z = z'$ and $\psi(z, z') = 0$ otherwise. By assumption, there is a nonnegative mechanism φ for π_Y with the same set of reports $\mathcal{R} = Z$ such that $\sum_y \varphi(z, y)\pi_Y(y|\theta) = \sum_{z'} \psi(z, z)\pi_Z(z'|\theta) = \pi_Z(z|\theta)$ for every state θ. Define $M(y, z) = \varphi(z, y) \geq 0$ to get the required matrix M.

$(ii) \implies (i)$

Given a nonnegative mechanism ψ for π_Z, define a mechanism φ for π_Y as in (3). Since M is nonnegative, so is φ.

B.4 Proof of Corollary 3

For $p \in \Delta(\Theta)$, let λ_p be the distribution over Z induced by the belief p: for all $z \in Z$,

$$\lambda_p(z) = \int_\Theta \pi(y|\theta)p(\mathrm{d}\theta).$$

Observe that by the condition in Corollary 3, (Y, π_Y) dominates (Z, π_Z) in the sense of elicitation (for example, one may apply the condition on the mechanism constructed in the proof of Theorem 1). Therefore, as discussed in Section 4, there exists a Y-by-Z matrix M such that $\pi_Z = \pi_Y M$, and by Lemma 4 from the proof of Proposition 2, we choose M such that every row sums to 1.

The null space of π_Y is defined as

$$\ker \pi_Y = \left\{ v \in \mathbf{R}^Y : \forall \theta \in \Theta,\ \sum_y v(y)\pi(y|\theta) = 0 \right\}.$$

Because (Y, π_Y) is complete, $\ker \pi_Y = \{0\}$. Indeed, since $\mathbf{R}^Y$ is the linear span of the $(|Y| - 1)$-simplex (identified with $\Delta(Y)$), it suffices to show that for any v in the simplex,

$$\forall \theta \in \Theta,\ \sum_y v(y)\pi(y|\theta) = 0 \qquad \implies \qquad v = 0.$$

This implication holds because, if $p \in \Delta(\Theta)$ is such that $\lambda_p = v$—which is possible

because (Y, π_Y) is complete—then,

$$\int_\Theta \sum_y v(y)\pi(y|\theta)p(\mathrm{d}\theta) = \sum_y v(y)\lambda_p(y) = 0 \qquad \Longrightarrow \qquad v = 0.$$

Let $Z = \{z_1, \ldots, z_n\}$, and let $\delta_i \in \Delta(\Theta)$ be the distribution over parameters whose induced distribution over Z, λ_{δ_i}, puts full mass on z_i. The completeness of $(Z\,\pi_Z)$ ensures that this distribution exists.

Let ψ be the mechanism

$$\psi(p, z) = \sum_i \big(1 - (\lambda_p(z_i) - \mathbb{1}\{z = z_i\})^2\big).$$

Note that ψ is direct and incentive-compatible, with nonnegative values. Notice that for all $i \neq j$, $\psi(\delta_i, z_i) = 1$ and $\psi(\delta_i, z_j) = 0$.

By the assumption of Corollary 3 , there exists a direct, incentive-compatible mechanism φ with nonnegative payoffs such that for all $p \in \Delta(\Theta)$, and all i,

$$\mathrm{E}_p[\varphi(\delta_i, y)] = \mathrm{E}_p[\psi(\delta_i, z)].$$

In particular, taking for p the parameter distribution that puts full mass on θ, we get

$$\sum_y \varphi(\delta_i, y)\pi_Y(y|\theta) = \sum_z \psi(\delta_i, z)\pi_Z(z|\theta) = \pi_Z(z_i|\theta) = \sum_y M(y, z_i)\pi_Y(y|\theta).$$

So

$$\varphi(\delta_i, \cdot) - M(\cdot, z_i) \in \ker \pi_Y = \{0\},$$

which means that all entries of M are nonnegative.

Since $\sum_z M(y, z) = 1$ for all y, M is a Markov matrix, and thus (Y, π_Y) is more informative than (Z, π_Z).

B.5 Proof of Proposition 5

$(i) \Longrightarrow (ii)$

Consider a mechanism ψ for π_Z with set of reports $\mathcal{R} = 2^Z$ defined by $\psi(A, z) = 1$ if $z \in A$ and $\psi(A, z) = 0$ otherwise. Since the payoffs of ψ are in $[0, 1]$, there is by assumption a payoff-equivalent mechanism $\varphi(A, y)$ for π_Y with payoffs also in $[0, 1]$.

For every state θ and every $A \in 2^Z$ we have that

$$\sum_{z \in A} \pi_Z(z|\theta) = \sum_z \pi_Z(z|\theta)\psi(A,z) = \sum_y \pi_Y(y|\theta)\varphi(A,y),$$

where the first equality is by construction of ψ, and the second by payoff-equivalence for the belief p that assigns probability 1 to state θ. Defining $N(y,A) = \varphi(A,y)$ completes the proof.

$(ii) \Longrightarrow (i)$

Consider any mechanism ψ for π_Z with payoffs in $[0,1]$, and we will construct a payoff-equivalent mechanism φ for π_Y also with payoffs in $[0,1]$. Let $r \in \mathcal{R}$ be any report, and consider the vector $\psi(r,\cdot) \in [0,1]^Z$. This vector can be represented as a convex combination of the extreme point of the hypercube $[0,1]^Z$, which are the indicators $\mathbb{1}_A(\cdot)$ of all subsets $A \in 2^Z$. Thus, there are non-negative numbers $\{\eta_r(A)\}_{A \in 2^Z}$ summing up to 1 such that for all z, $\psi(r,z) = \sum_{A \in 2^Z} \eta_r(A)\mathbb{1}_A(z)$. Let φ be defined by $\varphi(r,y) = \sum_{A \in 2^Z} \eta_r(A)N(y,A)$, and note that $\varphi(r,y) \in [0,1]$ as a convex combination of numbers in $[0,1]$. Also, for every state θ we have

$$\sum_y \pi_Y(y|\theta)\varphi(r,y) = \sum_y \pi_Y(y|\theta) \sum_{A \in 2^Z} \eta_r(A)N(y,A) = \sum_{A \in 2^Z} \eta_r(A) \sum_y \pi_Y(y|\theta)N(y,A) =$$
$$\sum_{A \in 2^Z} \eta_r(A) \sum_{z \in A} \pi_Z(z|\theta) = \sum_z \pi_Z(z|\theta) \sum_{\{A : z \in A\}} \eta_r(A) = \sum_z \pi_Z(z|\theta)\psi(r,z),$$

where the first equality is by construction of φ, the second is just a change in the order of summation, the third is by the assumption of the proposition, the fourth is again a change in the order of summation, and the last is by $\psi(r,z) = \sum_{A \in 2^Z} \eta_r(A)\mathbb{1}_A(z)$. This shows that φ gives the same expected payoff as ψ at any state θ implying that they are payoff-equivalent.

References

Abernethy, J. and R. Frongillo (2012). A characterization of scoring rules for linear properties. In *the 25th Annual Conference on Learning Theory (COLT)*.

Al-Najjar, N., A. Sandroni, R. Smorodinsky, and J. Weinstein (2010). Testing theories with learnable and predictive representations. *Journal of Economic Theory 145*(6), 2203–2217.

Aliprantis, C. D. and K. C. Border (2006). *Infinite Dimensional Analysis: A Hitchhiker's Guide* (third ed.). Springer.

Aumann, R. J. (1976). Agreeing to disagree. *The Annals of Statistics 4*(6), 1236–1239.

Barelli, P. (2009). On the genericity of full surplus extraction in mechanism design. *Journal of Economic Theory 144*(3), 1320–1332.

Blackwell, D. (1951). The comparison of experiments. In J. Neyman (Ed.), *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, pp. 93–102. University of California Press, Berkeley.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review 78*(1), 1–3.

Bruni, C. and G. Koch (1985). Identifiability of continuous mixtures of unknown Gaussian distributions. *The Annals of Probability 13*(4), 1341–1357.

Chambers, C. P. and N. S. Lambert (2021). Dynamic belief elicitation. *Econometrica 89*(1), 375–414.

Chen, Y.-C. and S. Xiong (2013). Genericity and robustness of full surplus extraction. *Econometrica 81*(2), 825–847.

Cramér, H. and H. Wold (1936). Some theorems on distribution functions. *Journal of the London Mathematical Society 1*(4), 290–294.

Crémer, J. and R. P. McLean (1988). Full extraction of the surplus in Bayesian and dominant strategy auctions. *Econometrica 56*(6), 1247–1257.

Dawid, A. (1982). The well-calibrated Bayesian. *Journal of the American Statistical Association 77*(379), 605–610.

De Finetti, B. (1962). Does it make sense to speak of "good probability appraisers"? In I. Good (Ed.), *The Scientist Speculates: An Anthology of Partly-Baked Ideas*, pp. 257–364. Basic Books.

Foster, D. and R. Vohra (1998). Asymptotic calibration. *Biometrika 85*(2), 379–390.

Frongillo, R. and I. Kash (2015). On elicitation complexity. In *Advances in Neural Information Processing Systems 28 (NIPS)*.

Fu, H., N. Haghpanah, J. Hartline, and R. Kleinberg (2021). Full surplus extraction from samples. *Journal of Economic Theory 193*.

Ganzburg, M. I. (1981). Multidimensional jackson theorems. *Siberian Mathematical Journal 22*(2), 223–231.

George, S. L. and M. Buyse (2015). Data fraud in clinical trials. *Clinical Investigation 5*(2), 161.

Gneiting, T. (2011). Making and evaluating point forecasts. *Journal of the American Statistical Association 106*(494), 746–762.

Gneiting, T. and A. E. Raftery (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association 102*(477), 359–378.

Good, I. J. (1952). Rational decisions. *Journal of the Royal Statistical Society, Series B 14*(1), 107–114.

Goodman, L. A. (1954). Some practical techniques in serial number analysis. *Journal of the American Statistical Association 49*(265), 97–112.

Grether, D. (1981). Financial incentive effects and individual decision-making. Manuscript No 401, Division of the Humanities and Social Sciences, California Institute of Technology.

Heifetz, A. and Z. Neeman (2006). On the generic (im)possibility of full surplus extraction in mechanism design. *Econometrica 74*(1), 213–233.

Karni, E. (2009). A mechanism for eliciting probabilities. *Econometrica 77*(2), 603–606.

Kechris, A. S. (1995). *Classical Descriptive Set Theory*. Springer-Verlag.

Lambert, N. S., D. M. Pennock, and Y. Shoham (2008). Eliciting properties of probability distributions. In *Proceedings of the 9th ACM Conference on Electronic Commerce*, pp. 129–138.

Le Cam, L. (1996). Comparison of experiments: A short review. In T. S. Ferguson, L. S. Shapley, and J. B. MacQueen (Eds.), *Statistics, Probability and Game Theory*, pp. 127–138. Institute of Mathematical Statistics.

Lehrer, E. and E. Shmaya (2008). Two remarks on Blackwell's theorem. *Journal of applied probability 45*(2), 580–586.

Lopomo, G., L. Rigotti, and C. Shannon (2019). Detectability, duality, and surplus extraction. Working paper (arXiv 1905.12788).

Matheson, J. E. and R. L. Winkler (1976). Scoring rules for continuous probability distributions. *Management Science 22*(10), 1087–1096.

McAfee, R. P. and P. J. Reny (1992). Correlated information and mechanism design. *Econometrica 60*(2), 395–421.

McCarthy, J. (1956). Measures of the value of information. *Proceedings of the National Academy of Sciences 42*(9), 654–655.

Milgrom, P. and I. Segal (2002). Envelope theorems for arbitrary choice sets. *Econometrica 70*(2), 583–601.

Miller, N., P. Resnick, and R. Zeckhauser (2005). Eliciting informative feedback: The peer-prediction method. *Management Science 51*(9), 1359–1373.

Neeman, Z. (2004). The relevance of private information in mechanism design. *Journal of Economic theory 117*(1), 55–77.

Olszewski, W. and A. Sandroni (2008). Manipulability of future-independent tests. *Econometrica 76*(6), 1437–1466.

Rahman, D. (2012). Surplus extraction on arbitrary type spaces. Working paper.

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

Savage, L. J. (1971). Elicitation of personal probabilities and expectations. *Journal of the American Statistical Association 66*(336), 783–801.

Shmaya, E. (2008). Many inspections are manipulable. *Theoretical Economics 3*(3), 367–382.

Strulovici, B. (2017). On the impossibility of full surplus extraction. Working paper.

Teicher, H. (1961). Identifiability of mixtures. *The Annals of Mathematical Statistics 32*(1), 244–248.