Honest binary choice: the two player case

Jakša Cvitanić[*], Dražen Prelec[†], Sonja Radas[‡]and Hrvoje Šikić [§]

November 20, 2024

Abstract

In Bayesian mechanisms of the 'truth serum' type, respondents declare their type via a multiple choice question and predict the answers of other respondents. The center or planner does not know the information structure. In spite of many positive results, the important 2×2 case (two respondents and two types) remains basically unsolved. It has been shown recently that the 2×2 problem is unsolvable when the two players' types (as random variables) are not exchangeable, and no mechanism that is incentive compatible for arbitrary exchangeable cases has yet been proposed. Here we conduct a comprehensive analysis of the exchangeable 2×2 case, and uncover a diverse set of mechanisms with different properties. Included in our analysis is a method for implementing desired relationships between expected payoffs and the information structure.

Key words: proper scoring rules, Bayesian Truth Serum, eliciting truthful responses, incentive compatible surveys

JEL codes: C11, D82, D83, M00

[*] Division of the Humanities and Social Sciences, Caltech. E-mail: cvitanic@hss.caltech.edu. Research supported in part by NSF grant DMS–1810807.

[†] MIT, Sloan School of Management, Department of Economics, Department of Brain and Cognitive Sciences. E-mail: dprelec@mit.edu. Thanks are due to All Souls College, Oxford, for Visiting Fellowships during Michaelmas 2020, Hillary 2021, and Michaelmas 2022 terms.

[‡] The Institute of Economics, Zagreb, and MIT, Sloan School of Management. E-mail: sradas@eizg.hr; sradas@mit.edu. This work was made as part of the project "Determinants of Strengthening Technological Capabilities of Different Sectors" at the Institute of Economics, Zagreb and funded within the National Recovery and Resilience Plan 2021-2026 – NextGenerationEU.

[§] University of Zagreb, Faculty of Science, Department of Mathematics. E-mail: hsikic@math.hr. Research supported in part by the Croatian Science Foundation under the project HRZZ-IP-2022-10-5116 (FANAP).

1 Introduction

Bayesian mechanisms of the 'truth serum' type have been developed to reward honest answers to questions about beliefs, behaviors, preferences or intentions, as might be elicited in a social science survey or lab experiment (Prelec 2004). Respondents provide information (their 'signal' or 'type') by answering a multiple choice question, and also predict the distribution of answers of other respondents. A scoring system computes a payoff for each one based on these inputs. Respondents are presumed indifferent about truth: they do not have an explicit desire to deceive, but need strict incentives to answer honestly. The mechanism is deemed incentive-compatible (IC) if honest answers and honest predictions are in a strict Bayes-Nash equilibrium.

A notable feature of the literature is that the simplest case — two participants and two types — has received little theoretical attention. With one exception (Radanovic and Faltings 2014), mechanisms required either an infinite population (Prelec 2004), or some minimal finite number greater than two, e.g., Radanovic and Faltings (2013), Witkowski and Parkes (2012), Baillon (2017), Cvitanić et al. (2019). [1]

The omission is notable because the two player case is easy to set up and arises often in applications, for example when pairs of raters classify images. In that case, the relevant distribution is defined by a 2 × 2 matrix,

$$\begin{array}{c|cc} & T^s = Y & T^s = N \\ \hline T^r = Y & a & b \\ T^r = N & c & d \end{array}$$

with three degrees of freedom as $a+b+c+d=1$. Two players, r and s, can each be two types, Y or N, associated with random variables T^r, T^s. We regard Y and N as signals indicating the honest answer to a dichotomous question. An honest prediction for $T^r = Y$ would be the conditional distribution $(a/(a+b), b/(a+b))$, etc.. Importantly, players know the matrix but the planner who seeks information does not.

An IC mechanism elicits type declaration and predictions such that honest declarations and honest predictions by one player are, in expectation, a strict best response to honesty by the other player.

[1] Other papers on the topic include Miller, Resnick, and Zeckhauser (2005), Witkowski (2014), Waggoner and Chen (2013), Zhang and Chen (2014), Dasgupta and Ghosh (2013), Frongillo and Witkowski (2017). Cvitanić et al (2018), Cvitanić et al (2020), Prelec (2021).

A well-known necessary condition, termed 'stochastic relevance,' is that signals are informative, i.e., $a/(a + b) \neq c/(c + d)$. However, stochastic relevance is not sufficient as shown recently (Prelec, 2023). For all distributions a, b, c, d there exists an alternative distribution a', b', c', d', such that if honest type declaration is a strict best response to honesty given a, b, c, d and mechanism M, then dishonest declaration is strict best response given a', b', c', d' and that same mechanism M. Because the planner does not know which distribution is in effect, it follows that there is no such M for the general stochastically relevant 2×2 case.

Given this negative result, our focus shifts to the case where random variables are exchangeable, and the joint distribution symmetric:

$$\begin{array}{ccc} & T^s = Y & T^s = N \\ T^r = Y & a & b \\ T^r = N & b & c \end{array}$$

with $a + 2b + c = 1$. The only mechanism in the literature that (partially) addresses this case is the 'divergence based truth serum' of Radanovic and Faltings (2013), which assumes that: $ac > b^2$. However, honest expected payoffs are dominated by a pooling equilibrium, in which players declare a particular type irrespective of their signal. This creates a potential robustness concern, as players could profit by being uninformative.

In this paper, we attempt a comprehensive analysis of IC mechanisms for the exchangeable 2×2 case (without the restriction that $ac > b^2$). Going beyond individual solutions, we show which expected payoffs are implementable in a strict Bayesian equilibrium of some mechanism. Being able to specify how expected payoffs should depend on types and a, b, c (without knowing types or actual values of a, b, c), is useful in applications. For example, one might want to punish noisy responding by relating expected payoffs to Shannon information. Alternatively, one might desire a neutral mechanism that delivers constant expected payoffs irrespective of a, b, c, or a mechanism that assigns a bonus to rare respondent types.

A main result is that nearly all expected payoff functions that clear certain consistency requirements are implementable, including the cases of Shannon information and constant expected payoffs. We then examine other desirable mechanism properties, such as robustness, simplicity, boundedness and frugality. Here, the conclusions are more mixed, as not all such properties may obtain simultaneously. The planner thus faces tradeoffs, which we illustrate through specific examples. The overall theoretical landscape is surprisingly rich given the simplicity of the setup.

We define the model in Section 2, then, we define incentive compatibility and state basic properties of IC mechanisms in Section 3. We present results on implementation of expected payoffs in Section 4. Section 5 introduces desirable properties of IC mechanisms, offers alternative ways of constructing them, and presents a number of examples. We provide a discussion of the main results, including an extension to the case of more than two respondents in Section 6. All the proofs are provided in Appendix.

2 The Model

In this paper the focus is on dyads, with the set $\mathcal{R}$ of players (respondents) consisting of two players. Each player is asked by a planner to select (declare) one of $M = 2$ possible answers (types). We will denote the two types sometimes as Y(es) and N(o), and sometimes as 1 and 0, with Y corresponding to 1 and N corresponding to 0.

Player r receives a signal, the outcome of a random variable T^r with values in $\{Y, N\}$. We assume symmetry, for $r \neq s$:

$$P(T^r = Y, T^s = N) = P(T^r = N, T^s = Y) \ . \tag{2.1}$$

In our, 2×2 case, symmetry is equivalent to exchangeability. We also assume the *full-mixture condition*, i.e., for $r \neq s$, and for every $k, \ell \in \{Y, N\}$,

$$P(T^r = k, T^s = \ell) > 0 \ . \tag{2.2}$$

In addition to declaring their type, the planner asks the players to provide their prediction of the distribution of the declared types. The true predictions (beliefs) are represented by the *X-matrix of metapredictions* defined by $X = [x_{\ell k}]_{\ell \in \{0,1\}, k \in \{0,1\}}$, where for $r \neq s$,

$$x_{\ell k} := P(T^s = k \mid T^r = \ell) \ . \tag{2.3}$$

Note that (2.2) implies that $x_{\ell k}$ is always well-defined and (2.1) implies that (2.3) does not depend on the choice of r and s (as long as $r \neq s$). Since X is a stochastic matrix, we can simplify the notation by denoting

$$x := x_{11}, y := x_{01} \tag{2.4}$$

Hence, X is of the form

$$\begin{pmatrix} x & 1-x \\ y & 1-y \end{pmatrix} \tag{2.5}$$

where $0 < x, y < 1$. We shall often identify the X-matrix with the corresponding pair of values (x, y).

The planner rewards the players based on their declared types and predictions. She would like the reward mechanism to be incentive compatible (IC), that is, such that it is an equilibrium for the players to declare the types and predictions truthfully.

In the case $x = y$, for type 1 to have incentives to declare his type truthfully gives exactly the opposite incentives to type 0, so that there are no IC mechanisms. Thus, we impose the following standard assumption on the X-matrix.

Assumption 2.6 *The X-matrix satisfies stochastic relevance:*

$$x \neq y \; . \tag{2.7}$$

We denote the set of all such X-matrices by $\mathcal{X}$.

In principle, one could have a theory based on any choice of a subset $\mathcal{Y}$ of $\mathcal{X}$. In this paper we focus on three subsets of $\mathcal{X}$, the set $\mathcal{X}$ itself, and its subsets, denoted by $\mathcal{X}^+$ and $\mathcal{X}^-$, defined below. The X-matrix belongs to the set $\mathcal{X}^+$ if it satisfies the following property :

$$x > y \; . \tag{2.8}$$

Notice that (2.8) is implied by exchangeability if one assumes that the random variables generating the two players' signals are taken from an infinite sequence of exchangeable random variables. In that case, by de Finetti's classical theorem, there exists a latent 'state of nature' variable Ω such that signals of players are independent, identically distributed conditional on that Ω. With two players only, this is also equivalent to $x > y$. The interpretation is natural in applications if the two players are sampled from a large, undifferentiated population.

Set $\mathcal{X}^-$ is defined as the complement of $\mathcal{X}^+$ within $\mathcal{X}$. Hence, the X-matrix is in $\mathcal{X}^-$ if and only if $x < y$.

Finally, in practice, one can disallow metapredictions very close to 1 or 0. With this in mind, let $\varepsilon \in (0, \frac{1}{2})$ and consider the set $B(\varepsilon) = [\varepsilon,\ 1-\varepsilon] \times [\varepsilon,\ 1-\varepsilon]$ instead of $(0,1) \times (0,1)$. For any $\mathcal{Y} \subseteq \mathcal{X}$ we define

$$\mathcal{Y}_\varepsilon = \mathcal{Y} \cap B(\varepsilon) \quad . \tag{2.9}$$

3 Incentive Compatibility

As indicated above, each respondent r provides a response in the form (U^r, Z^r), where U^r is the declaration of the type and Z^r is the declaration of the beliefs about the other player's signal. From the planner's viewpoint, U^r is a random variable with values in $\{Y, N\}$ (or $\{1, 0\}$), while Z^r is a random variable with the values in the set of distributions $\{(z, 1-z) : z \in (0,1)\}$ on $\{$ 0,1 $\}$ (or $\{Y, N\}$). Since we have only two types, we shall identify $(z, 1-z) \sim z$. Hence $(U^r, Z^r) \in \{0,1\} \times (0,1)$. Furthermore, we denote by $-r$ the other player. The reward function chosen by the planner depends on all the reports, and the one for player r is denoted $f(u^r, z^r, u^{-r}, z^{-r})$. Observe that $u^r, u^{-r} \in \{1, 0\}$, and if we write $f(u^r, z^r, u^{-r}, z^{-r})$ in the form $f_{u^r,u^{-r}}(z^r, z^{-r})$, then f is completely described via four functions of two variables

$$f_{11}, f_{10}, f_{01}, f_{00} : (0,1) \times (0,1) \longrightarrow \mathbb{R} \tag{3.1}$$

We say that reward function f is *incentive compatible* (IC) if truth-telling is (strictly) an equilibrium, that is, assuming risk-neutral players,

$$E[f(r\ true; -r\ true) \mid T^r] > E[f(r\ deceives; -r\ true) \mid T^r] \ . \tag{3.2}$$

Respondent r is telling the truth if $U^r = T^r$ and Z^r is given via the X-matrix of metapredictions. In our notation it means that the true response is either $(1; x)$ or $(0; y)$. It follows that the left side of (3.2) is, under the condition $T^r = 1$, equal to

$$f_{11}(x, x)x + f_{10}(x, y)(1-x), \tag{3.3}$$

and, under the condition $T^r = 0$, equal to

$$f_{01}(y, x)y + f_{00}(y, y)(1-y). \tag{3.4}$$

3.1 IC Property

We allow deception of any form, as is standard in the literature. For example, if the true answer is $(1; x)$, then a deception can be either $(0; z),\ z \in (0,1)$, or $(1; z),\ z \in (0,1) \setminus \{x\}$. Let $\mathcal{Y}$ be a subset of $\mathcal{X}$.

Definition 3.5 *We say that f is $(\mathcal{Y}, \text{IC})$ if and only if for every $(x,y) \in \mathcal{Y}$ and for every $z, w \in (0,1)$,*

$$[f_{11}(x,x) - f_{01}(z,x)]x + [f_{10}(x,y) - f_{00}(z,y)](1-x) > 0 \ , \tag{3.6}$$

and

$$[f_{01}(y,x) - f_{11}(w,x)]y + [f_{00}(y,y) - f_{10}(w,y)](1-y) > 0 \ , \tag{3.7}$$

and, for every $(x,y) \in \mathcal{Y}$ and for every $z, w \in (0,1)$ such that $z \neq x$ and $w \neq y$,

$$[f_{11}(x,x) - f_{11}(z,x)]x + [f_{10}(x,y) - f_{10}(z,y)](1-x) > 0 \ , \tag{3.8}$$

and

$$[f_{01}(y,x) - f_{01}(w,x)]y + [f_{00}(y,y) - f_{00}(w,y)](1-y) > 0 \ . \tag{3.9}$$

As mentioned above, if we allowed X matrices with $x = y$, then there would be no f which is IC. This is because, in the case $x = y$, (3.6) with $z = y$, and (3.7) with $w = x$ provide exactly the opposite inequalities. It follows that $\mathcal{X}$ is the maximal set of matrices for which we can expect to have the IC rule, and We will, indeed, show that there are IC mechanisms that cover all of $\mathcal{X}$. We now state some basic properties. The following are the direct consequences of the definition above.

Lemma 3.10 *If $\mathcal{Y}_1 \subseteq \mathcal{Y}_2$ and f is ($\mathcal{Y}_2$, IC), then f is ($\mathcal{Y}_1$, IC).*

Lemma 3.11 *If either f_{00} or f_{11} is a constant function, then f is neither ($\mathcal{X}^+$, IC) nor ($\mathcal{X}^-$, IC).*

Lemma 3.12 *There is no ($\mathcal{X}$, IC) mechanism such that $f_{10} \equiv 0 \equiv f_{01}$.*

The following result is a straightforward consequence of the respective definitions. It shows that IC functions form a convex cone and that without loss of generality one can study functions such that either the diagonal values $f_{00}(z,z)$ and $f_{11}(z,z)$ or the diagonal values of $f_{10}(z,z)$ and $f_{01}(z,z)$ are identically zero.

Proposition 3.13 *(a) If f and g are ($\mathcal{Y}$, IC) and $\alpha, \beta > 0$ are positive real numbers, then $\alpha f + \beta g$ is ($\mathcal{Y}$, IC).*

(b) If f is ($\mathcal{Y}$, IC) and $g, h : (0,1) \to \mathbb{R}$ are any real functions, then the following is ($\mathcal{Y}$, IC):

$$\begin{bmatrix} f_{11}(z_1, z_2) + h(z_2) & f_{01}(z_1, z_2) + h(z_2) \\ f_{10}(z_1, z_2) + g(z_2) & f_{00}(z_1, z_2) + g(z_2) \end{bmatrix}$$

Remark 3.14 Let us list several other direct consequences of previous results and definitions.

(i) Prop. 3.13(b). implies that if f is ($\mathcal{Y}$, IC) and $A \in \mathbb{R}$ is a constant, then $f + A$ is ($\mathcal{Y}$, IC).

(ii) If f depends only on the first variable, i.e. $f(z_1, z_2) = f(z_1)$, then f is not ($\mathcal{X}^+$, IC).

(iii) If f depends only on the second variable, i.e. $f(z_1, z_2) = f(z_2)$, then f is not ($\mathcal{X}^+$, IC).

Here is an example of a (continuous) function f which is ($\mathcal{X}^+$, IC), but such that $f_{10} \equiv 0 \equiv f_{01}$, so that it cannot be ($\mathcal{X}$, IC).

Example 3.15 The following is a ($\mathcal{X}^+$, IC) mechanism, but it is not a ($\mathcal{X}$, IC) mechanism:

$$\begin{bmatrix} f_{11}(z_1, z_2) & f_{10}(z_1, z_2) \\ f_{01}(z_1, z_2) & f_{00}(z_1, z_2) \end{bmatrix} = \begin{bmatrix} (1 - z_2) - \mid z_1 - z_2 \mid & 0 \\ 0 & z_2 - \mid z_1 - z_2 \mid \end{bmatrix}$$

Players receive a non-zero score only if they declare the same type. Because honest predictions should then be identical, honest predictions are incentivized by an absolute value penalty on prediction differences. Honest type declarations are incentivized through a term controlled by the opposing player, i.e., z_2 for Player 1. This term rewards 'improbability', in that declaring type 1 yields $1 - z_2$, while declaring type 2 yields z_2. The expected score in equilibrium is $x(1 - x)$ for type 1 and $y(1 - y)$ for type 2.

Note that if players collude to declare type 1 and predict 50-50 irrespective of true type, they are guaranteed to receive the maximum possible expected score of 0.25. The mechanism is therefore vulnerable to pooling if one or the other type is a focal point. Pooling is less plausible if players are not known to each other, and play the game only once. In that case, it is not clear how they might choose between pooling on type 1 or type 0.

4 Implementing expected payoffs

Suppose that the planner has another goal - not just the type revelation, but also the specification of the expected payoff, either for both types or at least for the type which is of special interest (for example, the one whose response to a highly sensitive question is ”Yes”). More precisely, the planner wants to implement bounded and continuous expected payoff functions $V_i(x, y)$ for type i, $i = 0, 1$. What functions $V_i(x, y)$ can be implemented in strict IC equilibrium?

It is straightforward to verify that there are (many) pairs (V_0, V_1) that cannot be implemented. However, we will show that, if one imposes some reasonable conditions on functions V_i, implementation is possible both within $\mathcal{X}^-$ and within $\mathcal{X}^+$ frameworks. The general intuition behind the implementation is the following. If the players declare different types then the planner knows matrix X (assuming honesty). The planner then also knows that types 1 and 2 assign probabilities $1-x$ and y , respectively, to this outcome. The desired $V_i(x, y)$ can then be implemented (roughly) by requiring an equilibrium score of zero when type declarations are the same, and scores $V_1(x, y)/(1 - x)$ and $V_2(x, y)/y$ when type declarations are different.

Suppose we are given two real functions $V_1(z, w)$ and $V_0(w, z)$, where $z, w \in (0, 1)$, and an environment $\mathcal{Y} \subseteq \mathcal{X}$. Here, $V_1(z, w)$ represents the expected payoff of type 1 if type 1 declares z and type 0 declares w, and similarly for $V_0(w, z)$. Thus, $V_1(x, y)$ ($V_0(y, x)$) are the expected payoffs in the case the respondents honestly declare their type. In particular, this means that we seek an IC mechanism f such that, for every $(x, y) \in \mathcal{Y}$,

$$V_1(x, y) = xf_{11}(x, x) + (1 - x)f_{10}(x, y) \tag{4.1}$$

and

$$V_0(y, x) = yf_{01}(y, x) + (1 - y)f_{00}(y, y)\ . \tag{4.2}$$

As mentioned above, we can always transform any given mechanism so that $f_{00}(y, y) = 0 = f_{11}(x, x)$ by using Proposition 3.13 (b). Thus, we will be looking to implement such mechanisms.

Definition 4.3 *We say that a pair (V_0, V_1) satisfies the weak IC condition if, for every $x, y \in (0, 1)$, $x \neq y$,*

$$\frac{y}{x}V_1(x, y) > V_0(y, x) > \frac{1 - y}{1 - x}V_1(x, y) \tag{4.4}$$

It is not difficult to see that in the case $f_{00}(y,y)=0=f_{11}(x,x)$, this condition is necessary for the IC property (and equivalent to the weak ($\mathcal{X}$,IC) property below). Observe that in the case $(x,y)\in\mathcal{X}^+$ the condition implies $V_1(x,y)<0$ and $V_0(y,x)<0$, while in the case $(x,y)\in\mathcal{X}^-$ it implies $V_1(x,y)>0$ and $V_0(y,x)>0$. We will aim to implement nonnegative functions V_i. First, we state

Lemma 4.5 *Let $\mathcal{Y}\subseteq\mathcal{X}$. Suppose that functions V_1 and V_0 satisfy the following properties:*

(i) V_1 and V_0 are bounded;

(ii) V_0 and V_1 satisfy the weak IC condition (4.4);

(iii) For every $z,w\in(0,1)$,

$$\sup_{x\in(0,1)}\frac{V_1(x,z)-V_1(x,w)}{1-x}<\infty\ ;$$

(iv) For every $z,w\in(0,1)$,

$$\sup_{y\in(0,1)}\frac{V_0(y,z)-V_0(y,w)}{y}<\infty\ .$$

Then there exists a function $K(z,w)$, where $z,w\in(0,1)$, such that $K(z,z)=0$, for every $z\in(0,1)$, $K(z,w)>0$, for every $z\neq w$, and the following f is ($\mathcal{Y}$, IC) with expected payoffs V_1 and V_0:

$$\begin{bmatrix} f_{11}(z_1,z_2) & f_{10}(z_1,z_2) \\ f_{01}(z_1,z_2) & f_{00}(z_1,z_2) \end{bmatrix} = \begin{bmatrix} -\frac{K(z_1,z_2)}{(1-z_1)(1-z_2)z_1z_2} & \frac{V_1(z_1,z_2)}{1-z_1} \\ \frac{V_0(z_1,z_2)}{z_1} & -\frac{K(z_1,z_2)}{(1-z_1)(1-z_2)z_1z_2} \end{bmatrix}$$

Remark 4.6 Note that conditions (iii) and (iv) and similar conditions below are satisfied if the planner only allows responses x and y which are bounded away from zero and one, that is, if we work with $\mathcal{Y}_\varepsilon$.

This lemma already guarantees implementation in the case $\mathcal{Y}=\mathcal{X}^-$, even under the desirable restriction $V_1(x,y)\geq 0$ and $V_0(y,x)\geq 0$, since in that case we necessarily have $V_1(x,y)>0$ and $V_0(y,x)>0$. We state this result precisely in the following

Theorem 4.7 *Let $V_1(x,y)>0$ and $V_0(y,x)>0$, $(x,y)\in\mathcal{X}^-$, be bounded functions such that, for every $(x,y)\in\mathcal{X}^-$,*

$$\frac{y}{x}V_1(x,y)>V_0(y,x)>\frac{1-y}{1-x}V_1(x,y)\ .$$

Then there exists f which is ($\mathcal{X}^-$, IC), such that $f_{00}(y,y)=0=f_{11}(x,x)$, and $V_1(x,y)$ is the expected payoff for type 1 and $V_0(y,x)$ is the expected payoff for type 0, i.e, for every $(x,y)\in\mathcal{X}^-$,

$$V_1(x,y)=f_{10}(x,y)(1-x)\quad\text{and}\quad V_0(y,x)=f_{01}(y,x)y\ .$$

Let us also note that in the previous and in the next theorem, the existence of the IC-functions f is established through explicit constructions; for details see the proofs in the appendix.

In the remainder of the section, we want to establish implementation for non-negative V_i on $\mathcal{X}^+$. In the previous theorem V_1 and V_0 are only given for $0 < x < y < 1$, and the idea is to modify V_1, V_0, and condition (4.4) properly to $(0,1) \times (0,1)$, after which we will apply Lemma 4.5. However, the price to pay will be less flexibility in choosing one of the functions, say V_0.

Theorem 4.8 *Let $V_1(x,y) > 0$, $(x,y) \in \mathcal{X}^+$, be a bounded and strictly positive function such that*

$$\sup_{(x,y)\in\mathcal{X}^+} \frac{V_1(x,y)}{x(1-x)y} < \infty \quad .$$

Then, there exists f which is ($\mathcal{X}^+$, IC), such that $V_1(x,y)$ is the expected payoff for type 1, i.e., for every $(x,y) \in \mathcal{X}^+$,

$$V_1(x,y) = f_{11}(x,x)x + f_{10}(x,y)(1-x) \ .$$

We also have the following version of Theorem 4.8 for $\mathcal{X}_\varepsilon^+$.

Corollary 4.9 *Let $V_1(x,y) > 0$, where $(x,y) \in \mathcal{X}_\varepsilon^+$, be an arbitrary bounded and strictly positive function. Then there exists f which is ($\mathcal{X}_\varepsilon^+$, IC), such that $V_1(x,y)$ is the expected payoff for type 1, i.e, for every $(x,y) \in \mathcal{X}_\varepsilon^+$,*

$$V_1(x,y) = f_{11}(x,x)x + f_{10}(x,y)(1-x) \ .$$

It can also be shown that an analogous corollary holds in the case both V_1 and V_0 are fixed in advance, as long as they satisfy a reduction of the weak IC condition (4.4), i.e., the one valid on $B(\varepsilon)$.

Remark 4.10 Consider now implementation of payoffs V_1 and V_0 on the whole $\mathcal{X}$, restricting, without loss of generality, our discussion to the case $f_{00}(y,y) = 0 = f_{11}(x,x)$.

Weak IC condition (4.4) implies that for every $x, y \in (0,1), x \neq y$, we have

$$y(1-x)V_1(x,y) > x(1-y)V_1(x,y).$$

This implies $(y-x)V_1(x,y) > 0$. Hence, on the set where $y > x$, we must have $V_1(x,y) > 0$, while on the set where $y < x$, we must have $V_1(x,y) < 0$. By weak IC condition (4.4), the same then holds for V_0. In other words, it is not possible to implement a positive payoff on the entire domain $\mathcal{X}$. However, if we allow negative payoffs, we can implement the payoffs as in Lemma 4.5, by simply taking $\mathcal{Y} = \mathcal{X}$.

4.0.1 Examples

Information Gain. Consider $\mathcal{X}^+$ and suppose we want the expected payoff for player r to be equal to information gain, defined as

$$P\left(T^s = 1 \mid T^r\right) \log \frac{P\left(T^s = 1 \mid T^r\right)}{P\left(T^s = 1\right)} + P\left(T^s = 0 \mid T^r\right) \log \frac{P\left(T^s = 0 \mid T^r\right)}{P\left(T^s = 0\right)}$$

It is straightforward to verify that for this implementation we need to have

$$\begin{aligned} V_1(x, y) &= x \log \frac{x}{y/(1-x+y)} + (1-x) \log \frac{1-x}{(1-x)/(1-x+y)} \\ &= x \log \frac{x}{y} + \log(1-x+y) > 0 \\ V_0(x, y) &= y \log \frac{y}{y/(1-x+y)} + (1-y) \log \frac{1-y}{(1-x)/(1-x+y)} \\ &= (1-y) \log \frac{1-y}{1-x} + \log(1-x+y) > 0 \end{aligned}$$

Using Lemma 4.5, we can choose the following mechanism to implement information gain:

$$\begin{bmatrix} f_{11}(z_1, z_2) & f_{10}(z_1, z_2) \\ f_{01}(z_1, z_2) & f_{00}(z_1, z_2) \end{bmatrix}$$

$$= \begin{bmatrix} -\frac{K(z_1,z_2)}{(1-z_1)(1-z_2)z_1z_2} & \frac{1}{1-z_1}\left(z_1 \log \frac{z_1}{z_2} + \log\left(1 - z_1 + z_2\right)\right) \\ \frac{1}{z_2}\left((1 - z_2) \log \frac{1-z_2}{1-z_1} + \log\left(1 - z_1 + z_2\right)\right) & -\frac{K(z_1,z_2)}{(1-z_1)(1-z_2)z_1z_2} \end{bmatrix}$$

In contrast to Example 3.15, now the players can receive a positive score only by declaring different types. Declaring the same type either yields zero or a negative score. Considering Player 1, the expected score is:

$$-x \frac{K(z_1, x)}{(1-z_1)(1-x)z_1x} + (1-x) \frac{1}{1-z_1}\left(x \log \frac{z_1}{y} + \log\left(1 - z_1 + y\right)\right)$$

One potential problem from the standpoint of implementation is that, when the players know these functions, they might be tempted to transform their true probabilities x, y into probabilities $\phi(x), \phi(y)$ such that $V_i(\phi(x), \phi(y)) > V_i(x, y)$. We would still have $f_{11}(\phi(x), \phi(x)) = f_{00}(\phi(y), \phi(y)) = 0$, and we would have an equilibrium that dominates in expected score the honest equilibrium. Of course, both players would have to switch to this equilibrium; if only one switches they would lose.

Robustness to metapredictions. By this, we mean that V_1 and V_0 satisfy condition (4.4) and that, for all $(x, y) \in \mathcal{Y}$ and $(z, w) \in \mathcal{Y}$ we have:

$$V_1(z, w) > V_1(x, y) \text{ if and only if } V_0(w, z) < V_0(y, x).$$

With $\mathcal{Y} = \mathcal{X}_\varepsilon^+$, here is an example that is robust to metapredictions:

$$V_0(y,x) = \frac{1}{V_1(x,y)} \ , \quad V_1(x,y) = -\sqrt{\frac{x(1-x)}{y(1-y)}}.$$

The above mechanisms are not bounded if not restricted to $\mathcal{X}_\varepsilon^+$, and may have other inconvenient features. We explore mechanisms with desirable properties in what follows.

Remark 4.11 What we have provided here is only one way to implement feasible expected payoffs. However, there are usually other ways to implement them. For example, we will provide three distinct mechanisms that implement constant expected payoffs. One of them is not bounded, once adjusted to have $f_{ii}(z,z) = 0$, and another is not continuous.

5 IC Mechanisms with Desirable Properties

5.1 Desirable properties

Given that it turns out there are many IC mechanisms, the question is whether we can narrow down the choices by imposing some desirable properties. We propose to focus on these properties:

- 1. **Simplicity.** This is important so that the reward mechanism can be explained to the respondents. Henceforth, all our examples are built on the principle of simplicity.

-2. **Robustness.** Ideally, we would like to have a mechanism for which there is no non-honest equilibrium in which the expected payoff of both respondents is higher than in the honest equilibrium. Less stringently, a mechanism would be robust to a certain kind of deception. For example, we say that a mechanism is robust to pooling if it is not advantegeous for both of the players to report the same type when they are actually of different types.

- 3. **Continuity and boundedness.** These may also be desirable properties, e.g., for controlling the level of the realized cost to the planner.

-4. **Frugality.** The planner may want the expected payoff to the respondents to be as low as possible. Suppose there is a participation constraint, as typically assumed in contract theory: the expected payoff of the players of both types in the truthful equilibrium has to be no lower than a constant $C > 0$. That is, we need to have

$$f_{11}(x,x)x + f_{10}(x,y)(1-x) \geq C \tag{5.1}$$

and

$$f_{01}(y,x)y + f_{00}(y,y)(1-y) \geq C. \tag{5.2}$$

If there is an IC mechanism for which these two inequalities are equalities, such a mechanism minimizes the expected cost of the planner in honest equilibrium.

We will see below that there are examples that satisfy one or more of these properties for various sets of X-matrices.

In this section we study ways of constructing ($\mathcal{X}^+$, IC) and ($\mathcal{X}$, IC) mechanisms. In general, the family of ($\mathcal{X}^+$, IC) mechanisms is richer, but there are also examples of ($\mathcal{X}$, IC) mechanisms having some of the desirable properties. The theory of ($\mathcal{X}^-$, IC) is to some extent analogous to the theory of ($\mathcal{X}^+$, IC) mechanisms.

5.2 Examples of ($\mathcal{X}$, IC) mechanisms

We start by providing two relatively simple examples of mechanisms on the whole domain, that is, requiring only that $x \neq y$. A couple of more examples of this sort are also obtained later below, as extensions of ($\mathcal{X}^+$, IC) mechanisms.

Example 5.3 Robust (somewhat) to pooling of types, bounded from above and continuous ($\mathcal{X}$, IC) mechanism.

The following is a ($\mathcal{X}$, IC) mechanism that implements

$$V_1(x,y) = x(1-x)(y-x).$$

$$\begin{bmatrix} f_{11}(z_1,z_2) & f_{10}(z_1,z_2) \\ f_{01}(z_1,z_2) & f_{00}(z_1,z_2) \end{bmatrix} = \begin{bmatrix} -2^{\frac{|z_1-z_2|}{z_2}} & (z_2-z_1)z_1 \\ (z_1-z_2)(1-z_1) & -2^{\frac{|z_1-z_2|}{1-z_2}} \end{bmatrix}.$$

If all $x \neq y$ are allowed, this mechanism has a dominating equilibrium, in which the types, but not metapredictions, are reported truthfully. Indeed, suppose the players declare the types truthfully, and, the players of type 1 declare metaprediction z, while the players of type 0 declare metaprediction w. It is not hard to check that an equilibrium choice which maximize the players' expected payoffs is $(z,w) = (1/3, 2/3)$. This equilibrium, declaring honest types and metapredictions $(1/3, 2/3)$ dominates all the equilibria in which the types pool together and all the equilibria in which the players declare the types honestly.

The mechanism is robust to pooling of types in the sense that the dominating equilibrium involves honestly declaring types, although it misreports metapredictions. If we assume that the players will choose one of the two natural equilibria – the honest one, or the dominating one – then they will declare types honestly. However, if we assume that pooling of types is also a natural possible equilibrium, even if dominated, and since pooling may dominate the honest equilibrium, then the mechanism is not robust to pooling.

The following is a $(\mathcal{X}, \text{IC})$ mechanism that is bounded both from above and below and is everywhere continuous.[2]

Example 5.4 Bounded, robust to pooling of types and continuous $(\mathcal{X}$, IC$)$ mechanism.

The following mechanism has the just mentioned desirable properties. It implements

$$V_1(x, y) = (1 - x)y(y^2 - x^2)[1 - y^2 + 1 - x^2].$$

$$\begin{bmatrix} f_{11}(z_1, z_2) & f_{10}(z_1, z_2) \\ f_{01}(z_1, z_2) & f_{00}(z_1, z_2) \end{bmatrix}$$

$$= \begin{bmatrix} -4\,[\mid z_1 - z_2 \mid + \mid z_1^2 - z_2^2 \mid + \mid z_2^4 - z_1^4] \mid & (z_2^2 - z_1^2)\,[1 - z_1^2 + 1 - z_2^2]\, z_2 \\ (z_1^2 - z_2^2)\,[1 - z_1^2 + 1 - z_2^2]\,(1 - z_2) & -4\,[\mid z_1 - z_2 \mid + \mid z_1^2 - z_2^2 \mid + \mid z_1^4 - z_2^4 \mid] \end{bmatrix}.$$

If all $x \neq y$ are allowed, this mechanism has a dominating equilibrium, in which the types, but not metapredictions, are reported truthfully. Indeed, suppose the players declare the types truthfully, and, the players of type 1 declare metaprediction z, while the players of type 0 declare metaprediction w. Computing expected payoffs, we get

$$V_1(z, w) = (1 - x)w(w^2 - z^2)\left[(1 - z^2) + (1 - w^2)\right].$$

This is maximized at $z = 0$. We also get

$$V_0(w, z) = (1 - y)(1 - w)(z^2 - w^2)\left[(1 - w^2) + (1 - z^2)\right].$$

[2] It took us some time to construct this example, even if, in hindsight, it looks relatively simple.

This is maximized at $w = 1$. This equilibrium, declaring honest types and metapredictions $(0, 1)$ dominates all the equilibria in which the types pool together and all the equilibria in which the players declare the types honestly. The same discussion as in the previous example applies to robustness with respect to pooling of types.

In summary, if misreporting meta-predictions is not a concern, either because it is not likely, or the planner does not mind it, the last two mechanisms are appealing in the case the whole domain $x \neq y$ is allowed.

5.3 Examples of and methods for constructing ($\mathcal{X}^+$, IC) mechanisms

In this section we provide ways of constructing IC mechanisms that are often alternative to the method of implementation of expected payoffs that we introduced in Section 4. It will be useful to first define a condition weaker than the IC property. If $\mathcal{Y} = \mathcal{X}$, then the following notion is exactly the weak IC condition defined earlier.

Definition 5.5 *We say that f is weak ($\mathcal{Y}$,IC) if f satisfies (3.6), with $z = y$, and (3.7), with $w = x$.*

5.3.1 Equal payoffs to different types

There are no ($\mathcal{X}$, IC) mechanisms such that "the competition between different types is always a draw", i.e., such that $f_{10} \equiv 0 \equiv f_{01}$ (see Lemma 3.13). However, there are ($\mathcal{X}^+$, IC) such mechanisms. They can be fully characterized as follows.

Theorem 5.6 *Suppose $f_{10} \equiv 0 \equiv f_{01}$. Then, f is ($\mathcal{X}^+$, IC) if and only if f is weak ($\mathcal{X}^+$, IC) and, for every $0 < y < x < 1$ and for every $z, w \in (0, 1)$ such that $z \neq x, w \neq y$, we have*

$$f_{11}(z, x) < f_{11}(x, x) \text{ and } f_{00}(w, y) < f_{00}(y, y) \ . \tag{5.7}$$

The theorem raises the question of characterizing all weak ($\mathcal{X}^+$, IC) mechanisms with $f_{10} \equiv 0 \equiv f_{01}$. We have the following characterization.

Theorem 5.8 *Suppose $f_{10} \equiv 0 \equiv f_{01}$, and function $y \to f_{00}(y, y) =: g(y)$, $y \in (0, 1)$, is left-continuous. Then f is weak ($\mathcal{X}^+$, IC) if and only if g is strictly positive, strictly increasing and continuous, and we have, for every $0 < x < 1$, $f_{11}(x, x) = f_{00}(x, x)\frac{1-x}{x}$, and function $x \to f_{11}(x, x)$ is strictly decreasing.*

Remark 5.9 This leads to the question of finding continuous functions $g : (0,1) \to (0,+\infty)$ which are strictly increasing and such that $\frac{1-x}{x}g(x)$ is strictly decreasing. Here are some examples:

(i) $g(x) = x^\alpha, 0 < \alpha \leq 1$;

(ii) $g(x) = x^2$ does not satisfy the requirements, but $g(x) = x^2 + a$, where $a > \frac{1}{27}$, does;

(iii) $g(x) = e^x$;

(iv) $g(x) = \frac{1}{1-x}$.

Example 5.10 Frugal, robust (somewhat) to pooling, and continuous ($\mathcal{X}^+$, IC) mechanism. We now define a mechanism that implements $V_1(x,y) = (1-x)g(x)$. In particular, with $g(x) = \frac{C}{1-x}$, it implements a constant payoff. Using Theorem 5.8 and Theorem 5.6, define [3]:

$$\begin{bmatrix} f_{11}(z_1,z_2) & f_{10}(z_1,z_2) \\ f_{01}(z_1,z_2) & f_{00}(z_1,z_2) \end{bmatrix} = \begin{bmatrix} g(z_2)\frac{1-z_2}{z_2} - k|z_2 - z_1| & 0 \\ 0 & g(z_2) - k|z_2 - z_1| \end{bmatrix}.$$

Here, $k > 0$ and $g(z)$ is strictly positive and strictly increasing and $h(z) := g(z)(1-z)/z$ is strictly positive and strictly decreasing, so the highest values are attained at 1 and 0, respectively. We also take g such that $g(1) = h(0)$, for example $g(z) = 1/(1-z)$.

The mechanism is continuous. As for robustness to pooling, suppose we consider $\mathcal{X}^+_\varepsilon$ for some $\varepsilon > 0$. If the players' belief is such that when trying to pool, the probability that the metapredictions will be different is strictly positive, then, if k is large enough, the expected value when trying to dishonestly pool would be lower than when declaring the true type, and the mechanism would be robust to pooling. (However, see below for pooling equilibria.)

We can also make this mechanism maximally frugal. If C is the minimal expected value the players would accept, we can set

$$g(z) = \frac{C}{1-z} ,$$

and have the players' expected values equal to C in honest equilibrium. In other words, this mechanism implements constant (equal) expected payoffs.

Note, however, that this mechanism is not bounded. Moreover, it is not robust to metapredictions – it can be verified that any pair of metapredictions (z_1, z_2) that satisfes $z_1 = z_2$ constitutes an equilibrium. If the players declare $z_1 = z_2 = 1/L$, we can check that the expected payoffs of type 1 and

[3] The choice of $f_{00}(w,y)$, $f_{11}(z,x)$ can be something else, as long as it is negative enough when $y \neq w$ and $x \neq z$.

type 0 are LxC and $\frac{L}{L-1}(1-y)C$, respectively, which are strictly larger than C when $y < \frac{1}{L} < x$. A natural equilibrium of this type, then, might be setting $1/L$ in the middle, $\frac{1}{L} = \frac{1}{2}(x+y)$, or choosing L so that the expected payoffs to both types are equal. Note also that the players could achieve infinite payoff if they both declare type 1 and metapredictions $z_1 = z_2 = 0$, or if they both declare type 0 and metapredictions $z_1 = z_2 = 1$. However, it would be hard for the players to agree on a specific dishonest equilibrium without collusion, or without repeating the game many times, with feedback.

Remark 5.11 A potential problem with choosing a large k in this example and some examples below is, if the players are actually of the same type, they might prefer to declare different types, fearing they might be far off the diagonal with their metapredictions.

5.3.2 Decomposable mechanisms

We describe now so-called decomposable mechanisms, as defined in Radanovic and Faltings (2013). That is, we consider reward functions which can be divided into a sum of two contributions, one based on u^r, u^{-r}, z^{-r}, and the other on z^r, u^{-r}, z^{-r}. More precisely, we say that f is *decomposable* if there exist six functions $h_{00}, h_{01}, h_{10}, h_{11} : (0,1) \to \mathbb{R}$ and $g_0, g_1 : (0,1) \times (0,1) \to \mathbb{R}$ such that, for every $j, k \in \{0,1\}$ and for every $z_1, z_2 \in (0,1)$,

$$f_{kj}(z_1, z_2) = h_{kj}(z_2) + g_j(z_1, z_2) \ . \tag{5.12}$$

There are many decomposable IC mechanisms and we narrow the choice further by assuming that g functions depend only on the first variable, i.e.,

$$g_0(z_1, z_2) = g_0(z_1), \ \ g_1(z_1, z_2) = g_1(z_1). \tag{5.13}$$

Under this assumption we have the following results, one positive and one negative.

Proposition 5.14 *If f satisfies (5.12) and (5.13), then f is not ($\mathcal{X}$, IC).*

Theorem 5.15 *Assume that f satisfies (5.12) and (5.13). Then f is ($\mathcal{X}^+$, IC) if and only if, for every $x, z \in (0,1)$, $x \neq z$*

$$[g_0(x) - g_0(z)](1-x) + [g_1(x) - g_1(z)]x > 0, \tag{5.16}$$

and, for every $0 < y < x < 1$,

$$[h_{11}(x) - h_{01}(x)]x + [h_{10}(y) - h_{00}(y)](1-x) > 0 \ ,$$

$$[h_{01}(x) - h_{11}(x)]y + [h_{00}(y) - h_{10}(y)](1-y) > 0 \ . \tag{5.17}$$

An interesting special case is when there exists a function $g : (0,1) \to \mathbb{R}$ such that

$$g_0(x) = g(1-x), \text{ and } g_1(x) = g(x). \tag{5.18}$$

Observe that, under (5.18), the property (5.16) implies that, for every $z, w \in (0,1),\ z \neq w$,

$$g(z)z + g(1-z)(1-z) > g(w)z + g(1-w)(1-z). \tag{5.19}$$

This means that g is a *strictly proper scoring rule*, SPSR, well studied in the literature, as a mechanism for eliciting truthful (meta)predictions from a single respondent. There are infinitely many examples of functions g that satisfy (5.19), for example: $g(z) = \log(z),\ g(z) = 2z - z^2,\ g(z) = \frac{z}{\sqrt{z^2+(1-z)^2}}$.

Observe that property (5.17) is a special case of the weak ($\mathcal{X}^+$, IC) property. We can then apply the Whittaker-type lemma from [CPRS1]: h-functions satisfy (5.17) if and only if, for every $z \in (0,1)$,

$$h_{00}(z) - h_{10}(z) > 0 \text{ and } h_{11}(z) - h_{01}(z) > 0 \ ,$$

and, for every $0 < y < x < 1$,

$$\frac{y}{1-y}(h_{11}(x) - h_{01}(x)) < h_{00}(y) - h_{10}(y) < \frac{x}{1-x}(h_{11}(x) - h_{01}(x)) \ .$$

This last statement leads directly toward a fairly obvious algorithmic procedure to construct all h-functions that satisfy (5.17).

Example 5.20 Decomposable, bounded and continuous ($\mathcal{X}^+$, IC) mechanism. For a given SPSR g, the following mechanism implements

$$V_1(x,y) = x[h_{11}(x) + g(x)] + (1-x)[h_{10}(y) + g(1-x)].$$

$$\begin{bmatrix} f_{11}(z_1,z_2) & f_{10}(z_1,z_2) \\ f_{01}(z_1,z_2) & f_{00}(z_1,z_2) \end{bmatrix} = \begin{bmatrix} h_{11}(z_2) + g(z_1) & h_{10}(z_2) + g(1-z_1) \\ h_{01}(z_2) + g(z_1) & h_{00}(z_2) + g(1-z_1) \end{bmatrix}$$

such that, for some $k > 0$,

$$h_{00}(z) - h_{10}(z) = kz \ , \ h_{11}(z) - h_{01}(z) = k(1-z) \ .$$

We have, for every $0 < y < x < 1$,

$$\frac{y}{1-y}(h_{11}(x) - h_{01}(x)) = k\frac{y}{1-y}(1-x) < ky = h_{00}(y) - h_{10}(y).$$

Moreover,

$$\frac{x}{1-x}(h_{11}(x) - h_{01}(x)) = kx > ky = h_{00}(y) - h_{10}(y).$$

Thus, this is a ($\mathcal{X}^+$, IC) mechanism.

Suppose we take, for some K,

$$h_{11} = K, h_{00} = K$$

so that

$$h_{10}(z) = K - kz \ , \quad h_{01}(z) = K - k(1-z).$$

Since g is a SPSR, it is seen that with this mechanism there are no equilibria in which players would declare their metapredictions dishonestly, among those equlibria in which they declare their types honestly. However, the players might want to pool and declare the same metaprediction $z_1 = z_2 = z^*$, where z^* maximizes $g(z)$ if they declare type 1, and maximizes $g(1-z)$ if they declare type 0. These would dominate the honest equilibrium. On the other hand, it would be hard for the players to agree on a specific dishonest equilibrium without collusion or without repeating the game many times, with feedback.

Let C be the minimal expected payoff required by the players. Thus, we need to have

$$C \le x[h_{11}(x) + g(x)] + (1-x)[h_{10}(y) + g(1-x)] = xg(x) + (1-x)g(1-x) - ky(1-x) + K$$

and

$$C \le y[h_{01}(x) + g(y)] + (1-y)[h_{00}(y) + g(1-y)] = yg(y) + (1-y)g(1-y) - kx(1-y) + K.$$

As an example, take the quadratic SPSR $g(x) = x(2-x)$. Then, we need to have

$$C \le V_1(x,y) = K + x^2(2-x) + (1+x)(1-x)^2 - ky(1-x)$$

$$= 1 + K + x^2 - k(1-x)y - x.$$

This depends on x and y, so the mechanism cannot achieve maximum frugality, that is, the expected payoffs equal to C for all x, y. The minimal payoff for type 1 is attained at $y = x$ and $x = \frac{1}{2}$, and is equal to $1 + K - \frac{1}{4}(1+k)$. The same for type 0. The planner's best choice is to choose K and/or k that make this payoff equal to C.

Remark 5.21 The above example can be extended to the domain $x < y$, in a way that it is decomposable separately on $x < y$ and on $x > y$, but not on the union. We define g and h as above, but we replace h with

$$h(x)\mathbf{1}_{\{x>y\}} - h(x)\mathbf{1}_{\{x<y\}}.$$

Remark 5.22 A question arises whether one can build IC mechanisms based only on SPSR. Let g be any function that satisfies (5.19). Define a reward function f by

$$\begin{bmatrix} f_{11}(z_1,z_2) & f_{10}(z_1,z_2) \\ f_{01}(z_1,z_2) & f_{00}(z_1,z_2) \end{bmatrix} = \begin{bmatrix} g(z_1) & g(1-z_1) \\ g(z_1) & g(1-z_1) \end{bmatrix}.$$

It can be checked that this is a weak ($\mathcal{X}^+$, IC) mechanism, but not a ($\mathcal{X}^+$, IC) mechanism. We discuss next how to modify it to make it IC.

5.3.3 Separation of variables

We now propose a method for constructing reward functions that is somewhat similar in spirit to decomposability. We call it the *separation of variables* method and it is described in the following theorem. The mechanism implements, on $\mathcal{X}^+$,

$$V_1(x,y) = xg_1(x) + (1-x)g_0(y).$$

In particular, with $g_1 \equiv g_0 \equiv C$, it implements constant payoff.

Theorem 5.23 *Consider any functions $h_1, h_0 : (0,1) \to (-\infty, 0)$ and $g_1, g_0 : (0,1) \to (0, \infty)$. Define f by*

$$\begin{bmatrix} f_{11}(z,w) & f_{10}(z,w) \\ f_{01}(w,z) & f_{00}(w,z) \end{bmatrix} = \begin{bmatrix} (-\frac{1-z}{z}h_1(z) + g_1(z))\mathbf{1}_{\{z=w\}} & (h_1(z) + g_0(w))\mathbf{1}_{\{0<w<z<1\}} \\ (g_1(z) + h_0(w))\mathbf{1}_{\{0<w<z<1\}} & (-\frac{w}{1-w}h_0(w) + g_0(w))\mathbf{1}_{\{z=w\}} \end{bmatrix}.$$

Then f is ($\mathcal{X}^+$, IC).

Remark 5.24 The idea behind this method is to work on adjusting "the honest response part" in order to make the mechanism IC.

Example 5.25 Frugal, robust to pooling of types and bounded ($\mathcal{X}^+$, IC) mechanism, continuous outside the diagonal. Consider the mechanism from Theorem 5.23, and let $C > 0$ be the minimal expected payoff to the players. To minimize the planner's expected cost, we need to have

$$C = x[-\frac{1-x}{x}h_1(x) + g_1(x)] + (1-x)[h_1(x) + g_0(y)] = xg_1(x) + (1-x)g_0(y)$$

$$C = y[g_1(x) + h_0(y)] + (1-y)[-\frac{y}{1-y}h_0(y) + g_0(y)] = yg_1(x) + (1-y)g_0(y).$$

Thus, we can take $g_0 = g_1 = C$. Moreover, we can set, for example, $h_0(y) = k(y-1)$ and $h_1(x) = -kx$, to avoid unbounded payoffs. We get

$$\begin{bmatrix} f_{11}(z,w) & f_{10}(z,w) \\ f_{01}(w,z) & f_{00}(w,z) \end{bmatrix} = \begin{bmatrix} (k(1-z)+C)\mathbf{1}_{\{z=w\}} & (-kz+C)\mathbf{1}_{\{0<w<z<1\}} \\ (-k(1-w)+C)\mathbf{1}_{\{0<w<z<1\}} & (kw+C)\mathbf{1}_{\{z=w\}} \end{bmatrix}.$$

This is also robust to pooling of types, if the players' belief is such that, when trying to pool, the probability that the reported metapredictions are different is strictly positive, and if C is large enough. However, see below about pooling equilibria.

This mechanism is not continuous on the diagonal, i.e., when the metapredictions are the same. Moreover, the mechanism is not robust to metapredictions. If the players declare the types honestly and declare $z = w$, we can check that the expected payoffs of type 1 and type 0 are $k(x-z)+C$ and $k(z-y)+C$ respectively. Thus, type 1 wants to declare $z < x$ and as low as possible, and type 2 wants to declare $z > y$ and as high as possible. A natural equilibrium of this type, then, might be setting z in the middle, $z = \frac{1}{2}(x+y)$, or choosing z so that the expected payoffs to both types are equal. Note also that the players could achieve the highest payoff if they both declare type 1 and metapredictions $z = w = 0$, or if they both declare type 0 and metapredictions $z = w = 1$. On a plus side, it would be hard for the players to agree on a specific dishonest equilibrium without collusion or without repeating the game many times, with feedback.

An advantage of this method (vs decomposability) is that a simple modification provides a ($\mathcal{X}$, IC) mechanism.

Example 5.26 Robust to pooling of types and bounded ($\mathcal{X}$, IC) mechanism, continuous on the subdomains. We now modify the above example to make it work on the extended domain, on which the only requirement on (x, y) is $x \neq y$. Let $C, K, k > 0$ be constants, such that $C + K \geq k$. Consider the mechanism

$$\begin{bmatrix} f_{11}(z,w) & f_{10}(z,w) \\ f_{01}(w,z) & f_{00}(w,z) \end{bmatrix}$$

$$= \begin{bmatrix} (k(1-w)+C)\mathbf{1}_{\{z=w\}} - K\mathbf{1}_{\{z\neq w\}} & [-kz+C]\mathbf{1}_{\{z>w\}} + [kw+C]\mathbf{1}_{\{z\leq w\}} \\ (-k(1-z)+C)\mathbf{1}_{\{z<w\}} + (k(1-w)+C)\mathbf{1}_{\{z\geq w\}} & (kw+C)\mathbf{1}_{\{z=w\}} - K\mathbf{1}_{\{z\neq w\}} \end{bmatrix}.$$

This is a ($\mathcal{X}$, IC) mechanism. It is continuous separately on the subdomain on which $z < w$ and on the subdomain on which $z > w$.

5.3.4 Mechanisms based on strictly proper scoring rules

Recall the mechanism defined in Remark 5.22, where g is an SPSR. As mentioned in the remark, the mechanism is weak ($\mathcal{X}^+$, IC) and it satisfies (3.8) and (3.9), satisfies (3.6) when $x \neq z$, and satisfies (3.7) when $w \neq y$. In other words, to be ($\mathcal{X}^+$, IC), it fails only in (3.6) when $x = z$, and in (3.7) when $w = y$, and in both cases the left-hand side is zero. Therefore, it is not difficult to adjust it to obtain a reward function that is ($\mathcal{X}^+$, IC); we just need to add a strictly positive ”penalty” when the respondent lies in such a way that s/he reports an honest metaprediction, but reports her/his type dishonestly.

In particular, the following is a ($\mathcal{X}^+$, IC) reward function:

$$\begin{bmatrix} f_{11}(z_1,z_2) & f_{10}(z_1,z_2) \\ f_{01}(z_1,z_2) & f_{00}(z_1,z_2) \end{bmatrix} = \begin{bmatrix} g(z_1) & g(1-z_1) - \mathbf{1}_{\{z_1=z_2\}} \\ g(z_1) - \mathbf{1}_{\{z_1=z_2\}} & g(1-z_1) \end{bmatrix}.$$

However, this mechanism is not continuous. We now present a version of this approach that allows for continuous mechanisms. A similar idea was used in Radanovic and Faltings (2014), although still proposing a discontinuous mechanism.[4]

[4] More precisely, the mechanism Radanovic and Faltings (2014) proposes is

$$\begin{bmatrix} f_{11}(z_1,z_2) & f_{10}(z_1,z_2) \\ f_{01}(z_1,z_2) & f_{00}(z_1,z_2) \end{bmatrix} = \begin{bmatrix} g(z_1)) - \mathbf{1}_{D(z_1||z_2)>\theta} & g(1-z_1) \\ g(z_1) & g(1-z_1) - \mathbf{1}_{D(z_1||z_2)>\theta} \end{bmatrix}.$$

Here, θ is a parameter, and $D(z_1||z_2)$ is a divergence function associated with SPSR g.

Given a strictly proper scoring rule g, the three mechanisms from the following proposition implement, on $\mathcal{X}^+$, the following expected payoffs:

$$(i) \text{ and } (ii): \; V_1(x,y) = xg(x) + (1-x)g(1-x) + x(1-x).$$

$$(iii): \; V_1(x,y) = xg(x) + (1-x)g(1-x).$$

We will see that for a specific quadratic SPSR the constant payoff can be implemented by the mechanism in (ii) below.

Proposition 5.27 *Let $g : (0,1) \to \mathbb{R}$ be a strictly proper scoring rule. Then, the following are ($\mathcal{X}^+$, IC) mechanisms:*

(i) With $k > 0$

$$\begin{bmatrix} f_{11}(z_1,z_2) & f_{10}(z_1,z_2) \\ f_{01}(z_1,z_2) & f_{00}(z_1,z_2) \end{bmatrix} = \begin{bmatrix} g(z_1) + (1-z_2) - k \mid z_1 - z_2 \mid & g(1-z_1) \\ g(z_1) & g(1-z_1) + z_2 - k \mid z_1 - z_2 \mid \end{bmatrix}.$$

(ii) With $k > 1$,

$$\begin{bmatrix} f_{11}(z_1,z_2) & f_{10}(z_1,z_2) \\ f_{01}(z_1,z_2) & f_{00}(z_1,z_2) \end{bmatrix} = \begin{bmatrix} g(z_1) + (1-z_1) - k \mid z_1 - z_2 \mid & g(1-z_1) \\ g(z_1) & g(1-z_1) + z_1 - k \mid z_1 - z_2 \mid \end{bmatrix}.$$

(iii) With $k > 0$,

$$\begin{bmatrix} f_{11}(z_1,z_2) & f_{10}(z_1,z_2) \\ f_{01}(z_1,z_2) & f_{00}(z_1,z_2) \end{bmatrix} = \begin{bmatrix} g(z_1) - k \mid z_1 - z_2 \mid & g(1-z_1) \\ g(z_1) & g(1-z_1) - k \mid z_1 - z_2 \mid \end{bmatrix}.$$

Remark 5.28

(i) If the players' belief is such that, when trying to pool, the probability that the metapredictions will be different is strictly positive, then these mechanisms are robust to pooling, if we choose k large enough. However, similarly to previous examples, the players might want to pool and declare the same metaprediction $z_1 = z_2 = z^*$, where z^* maximizes $g(z)$ if they declare type 1, and maximizes $g(1-z)$ if they declare type 0. This will dominate the honest equilibrium, if they are sure they will pool successfully.

(ii) The approach in the previous proposition can be used to create many examples. More precisely, if $\tilde{f}$ is a mechanism from the proposition and f is any ($\mathcal{X}^+$, IC) mechanism, then $\tilde{f} + f$ is ($\mathcal{X}^+$, IC).

Example 5.29 Frugal, continuous and bounded ($\mathcal{X}^+$, IC) mechanism.

Consider a mechanism (ii) in the above proposition. Let $C > 0$ be the minimal expected value the players would accept in the truthful equilibrium. To minimize the planner's cost, we need to have

$$C = x[g(x) + (1 - x)] + (1 - x)g(1 - x)$$

$$C = yg(y) + (1 - y)[g(1 - y) + y].$$

The two equalities are equivalent. As an example, let's take the quadratic strictly proper scoring rule $g(x) = x(2 - x) + D$. We get

$$C = x^2(2 - x) + x - x^2 + (1 - x)^2(1 + x) + Dx + D(1 - x).$$

This is satisfied for all x if and only if

$$D = C - 1.$$

6 Discussion and Conclusions

In this paper, we have addressed the problem of designing incentive-compatible mechanisms for eliciting truthful responses in the context of binary choice questions involving two respondents, referred to as the 2×2 case. We note that the case of more than two respondents can be dealt with using the idea, familiar in the literature, of rewarding respondent r based only on her reports and reports of one other respondent called the "peer agent", or peer respondent. More precisely, in a "peer-based mechanism" for each respondent we randomly select (independently of everything else) the peer respondent, and apply a mechanism that is IC for two respondents. [5]

Our study presents several contributions and findings. First, we characterize expected payoffs that can be implemented by IC mechanisms. We then construct a variety of IC rules that not only encourage truthful reporting, but also adhere to desirable properties such as robustness, simplicity, boundedness, continuity, and frugality. Second, our analysis extends to the models which are not necessarily IID

[5] To see that a peer-based mechanism is IC also for R respondents, denote by $F(u^r, z^r, u^s, z^s)$ the expected payoff to respondent r conditional on his peer being player s. Suppose everyone other than r is declaring truthfully. The (unconditional) expected payoff for r is $\frac{1}{R-1}\sum_{s=1}^{R-1} F(u^r, z^r, u^s, z^s)$. Since, given truthful u^s, z^s, $F(u^r, z^r, u^s, z^s)$ is maximized at the honest values of u^r, z^r for every s, so is the expected payoff.

conditionally on a state of nature. While conditionally IID models have been extensively studied, our work demonstrates that 2 × 2 IC rules can also be constructed in more general models. This finding fills a gap in the existing literature and broadens the applicability of IC mechanisms.

Given that we have identified many IC mechanisms and provided a number of examples, it might be helpful to provide recommendations which ones to use in practice. This will depend on the application at hand and may be somewhat subjective. For one thing, it may be preferable in most cases to use a ($\mathcal{X}$, IC) mechanism, since it is difficult to guarantee in practice that the respondents' beliefs correspond to $y < x$. Here are some choices we find appealing:

– Example 5.20, if it is not a repeated game and we don't fear collusion (i.e., if we believe revealing honest types is the most likely equilibrium). It is robust to metapredictions and can be extended to $\mathcal{X}$.

– Examples 5.3 and 5.4, if misreporting of metapredictions is not a concern. They are robust (somewhat) to pooling of types and are defined on $\mathcal{X}$.

– Example 5.29, if we are confident that $y < x$ holds in the population. It is bounded, continuous and implements constant expected payoffs and is therefore frugal.

In future studies, it would be of interest to analyze further (for example, in experiments) which ones among the various rules we identify would be optimal to use in a given application, and to analyze whether some of our findings can be extended to the case of more than two multiple choices.

7 Appendix

The statement that Ω -i.i.d model leads to $x > y$, in our notation, is folklore to the experts. We provide a quick proof for reader's convenience. Let Ω denote the state of nature with two states, i.e. Ω is a random variable with values in probability distributions on (Yes, No), denoted by $(T, 1-T)$ and $(F, 1-F)$, where $0 < T, F < 1$ and $T \neq F$. Let us denote $p := Prob(\Omega = (T, 1-T))$. We assume here that $(T^r : r \in R)$ are Ω - i.i.d. . Hence

$$T = P(T^r = Yes \mid \Omega = (T, 1-T)), F = P(T^r = Yes \mid \Omega = (1-F, F)).$$

The system is then determined via the distribution of the vector (T^r, Ω), i.e., it is given via the Q-matrix of the form

$$Q = \left(\begin{array}{cc} Tp & F(1-p) \\ (1-T)p & (1-F)(1-p) \end{array} \right). \tag{7.1}$$

Using direct calculation, we then obtain

$$x = \frac{T^2 p + F^2(1-p)}{Tp + F(1-p)} \qquad y = \frac{T(1-T)p + F(1-F)(1-p))}{(1-T)p + (1-F)(1-p)}$$

which leads to

$$x - y = \frac{p(1-p)}{P(N)P(Y)}(F-T)^2 > 0,$$

where $P(N) = P(T^r = No), P(Y) = P(T^r = Yes)$.

Proof of Lemma 3.10. The proof is rather obvious.

Proof of Lemma 3.11. We show only the $\mathcal{X}^+$ case, since $\mathcal{X}^-$ goes in an analogous way. Similarly, the proof is analogous in the case "f_{00} is a constant". Hence, assume that there exists $C \in \mathbb{R}$ such that $f_{11} \equiv C$. Then (3.8) would lead to the property that for every $x, y, z \in (0,1)$, such that $0 < y < x < 1$ and $x \neq z$,

$$[f_{10}(x,y) - f_{10}(z,y)](1-x) > 0.$$

Since $(1-x) > 0$, we obtain $f_{10}(x,y) - f_{10}(z,y) > 0$. This would imply that for a fixed $y < min(x,z)$, we would have $[f_{10}(x,y) - f_{10}(z,y)] > 0$ and $[f_{10}(z,y) - f_{10}(x,y)] > 0$, which is a contradiction.

Proof of Lemma 3.12.

Assume that $f_{10} \equiv 0 \equiv f_{01}$. Then (3.6) with $z = y$ and (3.7) with $w = x$ leads to

$$f_{11}(x,x)\frac{x}{1-x} > f_{00}(y,y) > f_{11}(x,x)\frac{y}{1-y}$$

whenever $x, y \in (0,1), x \neq y$. Fix $x \in (0,1)$ and consider $y \in (0,1) \setminus \{x\}$. We obtain

$$0 < f_{11}(x,x)[\frac{x}{1-x} - \frac{y}{1-y}] = \frac{f_{11}(x,x)}{(1-x)(1-y)}(x-y).$$

It follows that $f_{11}(x,x) \neq 0$. If $f_{11}(x,x) > 0$, then $x > y$. This is a contradiction since we allow $y > x$, as well. Similarly for $f_{11}(x,x) < 0$.

Proof of Proposition 3.13.

This proof is rather straightforward from (3.6) - (3.9).

Proof of Remark 3.14.

(ii) Using (3.6) with $x = z$ and $f(z_1, z_2) = z_1$ gives

$$[f_{11}(z) - f_{01}(z)]z + [f_{10}(z) - f_{00}(z)](1 - z) > 0,$$

for every $0 < z < 1$. Similarly, using (3.7) with $y = w$ gives

$$[f_{01}(w) - f_{11}(w)]w + [f_{00}(w) - f_{10}(w)](1 - w) > 0,$$

for every $0 < w < 1$, which is a contradiction.

(iii) Using (3.8) with $f(z_1, z_2) = z_2$ gives

$$0 < [f_{11}(x) - f_{11}(x)]x + [f_{10}(y) - f_{10}(y)](1 - x) = 0;$$

clearly a contradiction.

Proof of Example 3.15.

Since $f_{10} \equiv 0 \equiv f_{01}$, Lemma 3.12. implies that f is not an ($\mathcal{X}$, IC) mechanism. For the $\mathcal{X}^+$ case, (3.6) becomes

$$[f_{11}(x, x) - f_{01}(z, x)]x + [f_{10}(x, y) - f_{00}(z, y)](1 - x) =$$

$$(1 - x)x + (1 - x)(\mid z - y \mid -y) = (1 - x)[x - y + \mid z - y \mid].$$

Hence, we need to show that, for every $x, y, z \in (0, 1)$, such that $y < x$, we have $x - y + \mid z - y \mid > 0$, which is obvious. Similarly for (3.7).

For (3.8) we obtain

$$[f_{11}(x, x) - f_{11}(z, x)]x + [f_{10}(x, y) - f_{10}(z, y)](1 - x) =$$

$$= [(1 - x) - ((1 - x) - \mid z - x \mid)]x = \mid z - x \mid x > 0,$$

since $x > 0$ and $x \neq z$. The argument for (3.9) is similar.

Proof of Lemma 4.5.

We consider all four inequalities needed for the IC property.

$$(3.6)\quad [f_{11}(x, x) - f_{01}(z, x)]x + [f_{10}(x, y) - f_{00}(z, y)](1 - x) =$$

$$= -\frac{V_0(z, x)}{z}x + [\frac{V_1(x, y)}{1 - x} + [\frac{K(z, y)}{(1 - z)(1 - y)zy}](1 - x) =$$

$$= V_1(x, y) - \frac{V_0(z, x)}{z}x + \frac{1 - x}{(1 - z)(1 - y)zy}K(x, y) =$$

$$= V_1(x,y) - V_1(x,z) + V_1(x,z) - \frac{V_0(z,x)}{z}x + \frac{1-x}{(1-z)(1-y)zy}K(x,y).$$

If $z = x$, the inequality holds because of (i). If $z \neq x$, then $V_1(x,z) - \frac{V_0(z,x)}{z}x > 0$ follows by (ii). Using (iii), there is such $K(x,y)$ so that $\frac{V_1(x,y)-V_1(x,z)}{1-x} + K(z,y) \geq 0$.

$$(3.7)\quad \forall x,y,w,x \neq y, [f_{01}(y,x) - f_{11}(w,x)]y + [f_{00}(y,y) - f_{10}(w,y)](1-y) =$$
$$= \frac{V_0(y,x)}{y}y - \frac{V_1(w,y)}{1-w}(1-y) + \frac{yK(w,x)}{(1-x)(1-w)wx} =$$
$$= V_0(y,x) - V_0(y,w) + V_0(y,w) - \frac{V_1(w,y)}{1-w}(1-y) + \frac{yK(w,x)}{(1-x)(1-w)wx} > 0.$$

Again, using (i) the inequality is true when $y = w$. When $y \neq w$, then again we use (ii) for the middle term, and because of (iv) we can select K so that we have

$$\frac{V_0(y,x) - V_0(y,w)}{y} + K(w,x) \geq 0.$$

$$(3.8)\quad z \neq x,\ [f_{11}(x,x) - f_{11}(z,x)]x + [f_{10}(x,y) - f_{10}(z,y)](1-x) =$$
$$= \frac{x}{(1-z)(1-x)zx}K(z,x) + [\frac{V_1(x,y)}{1-x} - \frac{V_1(z,y)}{1-z}](1-x) =$$
$$= \frac{1}{(1-z)(1-x)z}\Big[\frac{K(z,x)}{z} + [V_1(x,y)(1-z) - (1-x)V_1(z,y)](1-x)\Big].$$

It suffices to show that we can select K so that $\mid V_1(x,y)(1-z) - (1-x)V_1(z,y) \mid \leq K(z,x)$, which is true by (i).

$$(3.9)\quad w \neq y, x \neq y,\ [f_{01}(y,x) - f_{01}(w,x)]y + [f_{00}(y,y) - f_{00}(w,y)](1-y) =$$
$$= [\frac{V_0(y,x)}{y} - \frac{V_0(w,x)}{w}]y + \frac{K(w,y)}{(1-w)(1-y)wy}(1-y) > 0;$$

since, by (i), we can always choose K such that the following inequality holds

$$[\frac{V_0(y,x)}{y} - \frac{V_0(w,x)}{w}]y + \frac{K(w,y)}{(1-w)wy} > 0.$$

This completes the proof of the lemma.

Proof of Theorem 4.7.

We want to prove that Lemma 4.5. holds. Observe that property (i) holds by the assumptions of the theorem. We want to prove (ii). We want to extend our functions to the set $0 < y < x < 1$ so that property (4.4) holds.

We define V_0, V_1 on diagonal such that $V_0(x,x) = 0 = V_1(x,x)$ and on $0 < y < x < 1$ by $V_1(x,y) := (y-x)(1-x)x$, $V_0(y,x) := (y-x)x(1-x)$. Consider the following three equalities:

$$\frac{y}{x}V_1(x,y) = (y-x)y(1-x)$$

$$V_0(y,x) = (y-x)x(1-x)$$

$$\frac{1-y}{1-x}V_1(x,y) = (y-x)x(1-y).$$

The three equalities show that property (4.4) holds on the set $0 < y < x < 1$. By the assumption of the theorem, this proves that (ii) is fulfilled.

From the following it follows easily that (iii) and (iv) are fulfilled as well.

$$\sup_{x\in(0,1)} \frac{V_1(x,z)}{1-x} \leq \sup_{0<x<z} \frac{V_1(x,z)}{1-x} + \sup_{z\leq x<1} \frac{V_1(x,z)}{1-x} \leq \frac{K}{1-z} + 1$$

$$\sup_{x\in(0,1)} \frac{V_0(y,x)}{y} \leq \sup_{0<y<x} \frac{V_0(y,x)}{y} + \sup_{x\leq y<1} \frac{V_0(y,x)}{y} \leq \frac{K}{x} + (y-x)x(1-x).$$

Using Lemma 4.5. it is now straightforward to check that the theorem holds.

Proof of Theorem 4.8.

Again we use Lemma 4.5. Consider $g(x) := \sup_{0<y<x<1} \frac{V_1(x,y)}{y}$.

$$\sup_{0<x<1} g(x) = \sup_{0<x<1} \sup_{0<y<x} \frac{V_1(x,y)}{y} \leq sup_{(y,x)} \frac{V_1(x,y)}{yx(1-x)} < +\infty;$$

by the assumption of the theorem. This implies that $g(x)$ is bounded.

For every $0 < y < x \Longrightarrow V_1(x,y) \leq g(x)y \leq g(x)x$. For $0 < y < x < 1$ we define $\tilde{V}_1(x,y) := V_1(x,y) - g(x)x < 0$. It is clear that $\tilde{V}_1(x,y)$ is bounded.

On the set $0 < x \leq y < 1$ we define $\tilde{V}_1(x,y) := (y-x)x(1-x)y$, and $\tilde{V}_0(y,x) := (y-x)x(1-x)y$.

We want $\tilde{V}_0(y,x)$ defined on $0 < y < x$ such that $\frac{y}{x}\tilde{V}_1(x,y) > \tilde{V}_0(y,x) > \frac{1-y}{1-x}\tilde{V}_1(x,y)$. Obviously, we can select such $\tilde{V}_0(y,x)$. We then have that

$$\tilde{V}_0(y,x) < \frac{y}{x}(V_1(x,y) - g(x)x) = \frac{y}{x}V_1(x,y) - yg(x) < V_1(x,y) - yg(x).$$

Since this last expression is bounded, $\tilde{V}_0$ is bounded. This proves properties (*i*) and (*ii*) from Lemma 4.5. We need to check (*iii*) and (*iv*). Regarding (*iii*), we derive the following inequalities

$$\sup_{x\in(0,1)} \frac{\tilde{V}_1(x,y)}{1-x} \leq \sup_{x\leq y<1} \frac{\tilde{V}_1(x,y)}{1-x} + \sup_{y<x<1} \frac{\tilde{V}_1(x,y)}{1-x} \leq 1 + \sup_{y<x<1} [\frac{V_1(x,y)}{1-x} - g(x)\frac{x}{1-x}] \leq 1 + \sup_{y<x<1} \frac{V_1(x,y)}{1-x} < +\infty.$$

To derive the above inequalities we used that $\sup_{x\leq y<1} \frac{\tilde{V}_1(x,y)}{1-x} \leq 1$, that $g(x)\frac{x}{1-x} > 0$, and that $\sup_{y<x<1} \frac{V_1(x,y)}{1-x}$ is bounded; the last being true by the assumption of the theorem. Therefore the property (*iii*) holds.

Similarly, to prove that (*iv*) holds, we get

$$\sup_{y\in(0,1)} \frac{\tilde{V}_0(y,x)}{y} \leq \sup_{0<y<x<1} \frac{\tilde{V}_0(y,x)}{y} + \sup_{0<x\leq y<1} \frac{\tilde{V}_0(y,x)}{y} < \sup_{0<y\leq x<1} \frac{\tilde{V}_0(y,x)}{y} + 1 \leq \sup_{0<y<x<1} \frac{\tilde{V}_1(x,y)}{x} + 1 < +\infty.$$

The above shows that (*iv*) holds.

Using Lemma 4.5. there is $\tilde{f}$ which is $(\mathcal{X}^+, IC)$, such that $E[true \mid T^r = 1] = \tilde{V}_1(x,y)$ and $E[true \mid T^r = 0] = \tilde{V}_0(x,y)$. Now we define f, using Proposition 3.13 (*b*), in the following way:

$$f_{11}(z_1,z_2) = \tilde{f}_{11}(z_1,z_2) + g(z_2), \quad f_{01}(z_1,z_2) = \tilde{f}_{01}(z_1,z_2) + g(z_2).$$

Observe that we leave f_{00} and f_{10} intact. In this way we get f which is $(\mathcal{X}^+, IC)$, and

$$E[true \mid T^r = 1] = f_{11}(x,x)x + f_{10}(x,y)(1-x) = g(x)x + \tilde{f}_{10}(x,y)(1-x) = V_1(x,y).$$

This completes the proof.

Proof of Example 5.3.

We need to check (3.6) to (3.9). Since (3.7) goes in analogous way to (3.6), and (3.9) to (3.8), we will check (3.6) and (3.8). For (3.6) consider $x, y, z \in (0,1)$, such that $x \neq y$. We obtain

$$[f_{11}(x,x) - f_{01}(z,x)]x + [f_{10}(x,y) - f_{00}(z,y)](1-x) = ,$$

$$= -(z-x)(1-z)x + (y-x)x(1-x) + 2\frac{\mid z-y \mid}{1-y}(1-x) =$$

$$= x[(y-z)(1-x) + (x-z)^2] + 2 \mid z-y \mid \frac{1-x}{1-y}.$$

If $y = z$, then $x \neq z$, and we obtain $x(x-z)^2 > 0$. If $y \neq z$, then our expression is bigger or equal than

$$(1-x) \mid z-y \mid [\frac{2}{1-y} \pm x] > (1-x) \mid z-y \mid (2 \pm x) > 0.$$

For (3.8) consider $x, y, z \in (0,1)$ such that $y \neq x \neq z$. We have

$$[f_{11}(x,x) - f_{11}(z,x)]x + [f_{10}(x,y) - f_{10}(z,y)](1-x) =$$

$$= 2\frac{\mid z-x \mid}{x}x + [(y-x)x - (y-z)z](1-x) =$$

$$= 2 \mid z-x \mid + y(1-x)(x-z) + (1-x)(-x^2+z^2) =$$

$$= \mid z-x \mid [2 \pm y(1-x) \mp (1-x)(z+x)] =$$

$$= \mid z-x \mid [2 \pm (1-x)(y-(z+x)] > 0,$$

since $z \neq x$ and $\mid (1-x)(y-(z+x)) \mid < \mid y-(z+x) \mid < 2$.

Proof of Example 5.4.

Consider $x, y, z, w \in (0,1)$ such that $x \neq y$. Then (3.6) gives

$$[f_{10}(x,y) - f_{00}(z,y)](1-x) + [f_{11}(x,x) - f_{01}(z,x)]x =$$

$$= f_{10}(x,y)(1-x) - f_{01}(z,x)x - f_{00}(z,y)(1-x) =$$

$$= (y^2-x^2)[(1-y^2)+(1-x^2)]y(1-x) - x(z^2-x^2)[(1-z^2)+(1-x^2)](1-x) - f_{00}(z,y)(1-x).$$

Now $(1-x) > 0$ factors out and we need to check

$$(y^2-x^2)[(1-y^2)+(1-x^2)]y - (z^2-x^2)[(1-z^2)+(1-x^2)]x - f_{00}(z,y) =$$

$$= (y^2-x^2)[(1-y^2)+(1-x^2)](y-x) + (y^2-x^2)[(1-y^2)$$

$$+(1-x^2)]x - (z^2-x^2)[(1-z^2)+(1-x^2)]x - f_{00}(z,y) =$$

$$(y-x)^2(y+x)[(1-y^2)+(1-x^2)] + x(1-x^2)[(y^2-x^2)-(z^2-x^2)]$$

$$+x[(y^2-x^2)(1-y^2) - (z^2-x^2)(1-z^2)] - f_{00}(z,y)$$

Since the first term is always positive, this is strictly greater than

$$x(1-x^2)(y^2-z^2) + x[(y^2-z^2)+(z^4-y^4)+x^2(y^2-z^2)] - f_{00}(z,y) >$$

$$> 2x(1-x^2)(y^2-z^2) + x(z^4-y^4) - f_{00}(z,y) > -2 \mid y^2-z^2 \mid - \mid y^4-z^4 \mid - f_{00}(z,y) \geq 0.$$

Considering (3.7) we obtain

$$[f_{01}(y,x) - f_{11}(w,x)]y + [f_{00}(y,y) - f_{10}(w,y)](1-y) =$$

$$= f_{01}(y,x)y - f_{10}(w,y)(1-y) - f_{11}(w,x)y =$$

$$= (y^2-x^2)[(1-y^2)+(1-x^2)](1-x)y - (y^2-w^2)[(1-y^2)+[(1-w^2)]y(1-y) - f_{11}(w,x)y.$$

Now $y > 0$ factors out and we need to check

$$(y^2-x^2)[(1-y^2)+(1-x^2)](1-x) - (y^2-w^2)[(1-y^2)+(1-w^2)](1-y) - f_{11}(w,x) =$$

$$= (y^2-x^2)[(1-y^2)+(1-x^2)](y-x)$$

$$+(y^2-x^2)[(1-y^2)+(1-x^2)](1-y) - (y^2-w^2)[(1-y^2)+(1-w^2)](1-y) - f_{11}(w,x)$$

Since the first term is always positive, the above is strictly greater than

$$(1-y)[(y^2-x^2)[(1-y^2)+(1-x^2)] - (y^2-w^2)[(1-y^2)+(1-w^2)]] - f_{11}(w,x) =$$

$$= (1-y)[-y^2x^2 - x^2(1-y^2) - x^2(1-x^2) + y^2w^2 + w^2(1-y^2) + w^2(1-w^2)] - f_{11}(w,x) =$$

$$= (1-y)[2(w^2-x^2)+(x^4-w^4)] - f_{11}(w,x) > -2 \mid w^2-x^2 \mid - \mid w^4-x^4 \mid - f_{11}(w,x) \geq 0.$$

Similarly, considering (3.8) we get

$$[f_{11}(x,x) - f_{11}(z,x)]x + [f_{10}(x,y) - f_{10}(z,y)](1-x) =$$

$$= (1-x)[(y^2-x^2)[(1-y^2)+(1-x^2)]y - (y^2-z^2)[(1-y^2)+(1-z^2)]y] - f_{11}(z,x)x =$$

$$= (1-x)[y^2(z^2-x^2)+(2-y^2)(z^2-x^2)+(x^4-z^4)]y - f_{11}(z,x)x =$$

$$= (1-x)y(z^2-x^2)[(1-x^2)+(1-z^2)] - f_{11}(z,x)]x = \Delta.$$

If $z > x$, then Δ is obviously positive. If $z < x$, then Δ equals

$$\Delta = (1-x)y(z-x)(z+x)[(1-x^2)+(1-z^2)] - f_{11}(z,x)x$$

Since $z + x < 2x$ and $z - x < 0$, the above is greater than

$$(1-x)y(z-x)2x[(1-x^2)+(1-z^2)] - f_{11}(z,x)x =$$

$$= x[2(1-x)y[(1-x^2)+(1-z^2)](z-x) - f_{11}(z,x)] > x[-4 \mid z-x \mid - f_{11}(z,x)] > 0.$$

Analogously, (3.9) gives

$$[f_{01}(y,x) - f_{01}(w,x)]y + [f_{00}(y,y) - f_{00}(w,y)](1-y) =$$

$$= y(1-x)[(y^2-x^2)[(1-y^2)+(1-x^2)]-(w^2-x^2)[(1-w^2)+(1-x^2)]]-f_{00}(w,y)(1-y) =$$

$$= y(1-x)[x^2y^2-x^2w^2+y^2(1-y^2)+y^2(1-x^2)-w^2(1-w^2)-w^2(1-x^2)]-f_{00}(w,y)(1-y).$$

Using $(w^4-y^4)=(y^2-w^2)(-y^2-w^2))$ this becomes

$$y(1-x)[(y^2-w^2)[(1-y^2)+(1-w^2)]]-f_{00}(w,y)(1-y)=\Theta$$

If $y>w$, then Θ is obviously positive. If $y<w$, then $y<w<0$ and $1-w<1-y$, i.e.

$$\Theta = y(1-x)(y^2-w^2)[(1-y)(1+y)+(1-w)(1+w)]-f_{00}(w,y)(1-y) >$$

$$> y(1-x)(y^2-w^2)(1-y)[(1+y)+(1+w)]-f_{00}(w,y)(1-y) >$$

$$> (1-y)[-4 \mid y^2-w^2 \mid -f_{00}(w,y)] > 0.$$

Proof of Theorem 5.6.

Given $f_{10}\equiv 0\equiv f_{01}$, (3.8) and (3.9) become

$$[f_{11}(x,x)-f_{11}(z,x)]x>0 \ for \ x\neq z$$

$$[f_{00}(y,y)-f_{00}(w,y)](1-y)>0 \ for \ y\neq w.$$

Similarly (3.6) and (3.7) become, for every $0<y<x<1$ and for every $z,w\in(0,1)$

$$f_{11}(x,x)x>(1-x)f_{00}(z,y)$$

$$f_{00}(y,y)(1-y)>yf_{11}(w,x);$$

Having $f_{00}(z,y)<f_{00}(y,y)$ for $z\neq y$, the first inequality is equivalent to $f_{11}(x,x)x>(1-x)f_{00}(y,y)$. Similarly the second, i.e., we have condition (4.4) on $\mathcal{X}^+$.

Proof of Theorem 5.8.

Given $f_{10}\equiv 0\equiv f_{01}$, condition (4.4) on $\mathcal{X}^+$ is equivalent to

$$f_{00}(y,y)\frac{1-y}{y}>f_{11}(x,x)>f_{00}(y,y)\frac{1-x}{x}$$

for every $0<y<x<1$. Since in this case $\frac{1-y}{y}>\frac{1-x}{x}$, it leads to both $f_{11}(x,x)$ and $f_{00}(y,y)$ being strictly positive. By letting $y\nearrow x$ and using left-continuity assumption, we obtain that, for every $0<x<1$,

$$f_{11}(x,x)=\frac{1-x}{x}f_{00}(x,x).$$

Using the first starting inequality, we obtain that, for every $0 < y < x < 1$,

$$f_{11}(y, y) > f_{11}(x, x),$$

i.e., $x \longrightarrow f_{11}(x, x)$ is strictly positive and strictly decreasing. By the left continuity assumption, it follows that $x \longrightarrow f_{11}(x, x)$ is also left-continuous. Using the second starting inequality, we obtain that, for every $0 < y < x < 1$,

$$f_{00}(x, x) = \frac{x}{1-x} f_{11}(x, x) > f_{00}(y, y),$$

i.e., $x \longrightarrow f_{00}(x, x)$ is strictly positive, strictly increasing and left-continuous. Monotonicity of both functions implies the existence of the following limits

$$\lim_{x \to y} [f_{00}(y, y) \frac{1-y}{y}] \geq \lim_{x \to y} f_{11}(x, x) \geq \lim_{x \to y} [f_{00}(y, y) \frac{1-x}{x}].$$

Hence,

$$f_{11}(y, y) = f_{00}(y, y) \frac{1-y}{y} \geq \lim_{x \to y} f_{11}(x, x) \geq f_{00}(y, y) \frac{1-y}{y} = f_{11}(y, y),$$

which implies the continuity of $x \longrightarrow f_{11}(x, x)$, and therefore of $x \longrightarrow f_{00}(x, x)$. The remaining part of the proof is rather straightforward.

Proof of Remark 5.9.

The proofs of (i), (iii), (iv) are standard and fairly obvious. Consider (ii). For $a \geq 0$, $x \longrightarrow x^2 + a$ is obviously continuous, strictly positive and strictly increasing. For the function

$$f(x) = \frac{1-x}{x}(x^2 + a) = (1-x)x + a\frac{1-x}{x} = (1-x)[x + \frac{a}{x}];$$

we have $f'(x) = 1 - 2x - \frac{a}{x^2}$.

We want to ensure that $f'(x) < 0$, for every $0 < x < 1$. For example, this will be so if the equation $f'(x) = 0$ has no solutions in $(0, 1)$. This leads to the equation $h(x) = 0$, where $h(x) = x^2 - 2x^3 - a$. Since $h'(x) = 0$ has solutions $x = 0$ and $x = \frac{1}{3}$, we will obtain good enough a, if $h(\frac{1}{3}) < 0$. Hence, $\frac{1}{9} - \frac{2}{27} - a < 0$, i.e., $a > \frac{1}{27}$.

Proof of Proposition 5.14.

Using (5.12) and (5.13), apply (3.6) with $x = z$, to obtain that, for every $x, y \in (0, 1),\ x \neq y$,

$$[h_{11}(x) - h_{01}(x)]x + [h_{10}(y) - h_{00}(y)](1 - x) > 0.$$

Denote $A(x) := [h_{11}(x) - h_{01}(x)]$ and $B(y) := [h_{00}(y) - h_{10}(y)]$. Apply (5.12) and (5.13) on (3.7), with $y = w$. Hence, from (3.6) and (3.7), we obtain, for every $x, y \in (0,1),\ x \neq y$,

$$A(x)\frac{x}{1-x} > B(y) > A(x)\frac{y}{1-y};$$

which implies

$$\frac{A(x)}{(1-x)(1-y)}(x-y) > 0.$$

For $x > y$ we obtain $A(x) > 0$, while for $x < y$, we obtain $A(x) < 0$; clearly a contradiction.

Proof of Theorem 5.15.

Consider (3.6) with $x = z$; it gives

$$[h_{11}(x) - h_{01}(x)]x + [h_{10}(y) - h_{00}(y)](1-x) > 0.$$

Similarly for (3.7) and the second inequality in the Theorem.

Since in (3.8) we must have $x \neq z$, we obtain

$$(g_1(x) - g_1(z))x + (g_0(x) - g_0(z))(1-x) > 0.$$

Observe that (3.9.) leads to the same inequality. The rest of the proof is straightforward.

Proof of Remark 5.22.

Observe that both (3.6), with $z = y$, and (3.7) with $w = x$, are equal to (5.19). Hence, we have weak $(\mathcal{X}^+, IC)$. However, the case $x = z$ in (3.6) leads to

$$(f_{11}(x,x) - f_{01}(x,x))x + (f_{10}(x,y) - f_{00}(x,y))(1-x) = 0.$$

Therefore, we do not have $(\mathcal{X}^+, IC)$.

Proof of Theorem 5.23.

Consider (3.6) with $0 < y < x < 1$. In the case $z = y$ we obtain

$$(f_{11}(x,x) - f_{01}(y,x))x + (f_{10}(x,y) - f_{00}(y,y))(1-x) =$$

$$= -(1-x)h_1(x) + g_1(x)x - g_1(x)x - h_0(y)x + h_1(x)(1-x)$$

$$+g_0(y)(1-x) + y\frac{1-x}{1-y}h_0(y) - g_0(y)(1-x) =$$

$$= h_0(y)[y\frac{1-x}{1-y} - x] = \frac{h_0(y)}{1-y}(y-x) > 0.$$

In the case $z \neq y$, we obtain

$$(f_{11}(x,x) - f_{01}(z,x))x + (f_{10}(x,y) - f_{00}(z,y))(1-x) =$$

$$= -(1-x)h_1(x) + g_1(x)x - g_1(x)x - f_{01}(z,x)x + (1-x)h_1(x) + g_0(y)(1-x) =$$

$$= g_1(x)x + g_0(y)(1-x) - f_{01}(z,x)x.$$

For $z < x$, we obtain

$$g_1(x)x + g_0(y)(1-x) - g_1(x)x - h_0(z)x = g_0(y)(1-x) - h_0(z)x > 0.$$

For $z \geq x$, we obtain $g_1(x)x + g_0(y)(1-x) > 0$. The proof of (3.7) goes along the same lines.

Consider (3.8) with $0 < y < x < 1$ and $x \neq z$.

$$(f_{11}(x,x) - f_{11}(z,x))x + (f_{10}(x,y) - f_{10}(z,y))(1-x) =$$

$$= -(1-x)h_1(x) + g_1(x)x + (1-x)h_1(x) + (1-x)g_0(y) - f_{10}(z,y)(1-x) =$$

$$= g_1(x)x + g_0(y)(1-x) - f_{10}(z,y)(1-x).$$

For $z > y$, we obtain

$$g_1(x)x + g_0(y)(1-x) - (1-x)h_1(z) - (1-x)g_0(y) = g_1(x)x - (1-x)h_1(z) > 0$$

For $z \geq x$, we obtain $g_1(x)x + g_0(y)(1-x) > 0$. Similarly for (3.9).

Proof of Example 5.26.

Rather straightforward to check.

Proof of Proposition 5.27.

We prove (i) only, since (ii) and (iii) are completely analogous. Consider (3.6). For $0 < y < x < 1$ and $z \in (0,1)$ we obtain

$$[f_{11}(x,x) - f_{01}(z,x)]x + [f_{10}(x,y) - f_{00}(z,y)](1-x) =$$

$$= (g(x) + (1-x) - g(z))x + (g(1-x) - g(1-z) - y + k \mid z - y \mid)(1-x) =$$

$$= [(g(x) - g(z))x + (g(1-x) - g(1-z))(1-x)] + (1-x)[k \mid z - y \mid +(x-y)] > 0,$$

by (5.17) and the fact that $k > 0$ and $x > y$. The proof for (3.7) is analogous.

Consider (3.8). For $0 < y < x < 1$ and $z \in (0,1)$ such that $x \neq z$, we obtain

$$[f_{11}(x,x) - f_{11}(z,x)]x + [f_{10}(x,y) - f_{10}(z,y)](1-x) =$$

$$= (g(x) + (1-x) - g(z) - (1-x) + k \mid z - x \mid)x + (g(1-x) - g(1-z))(1-x) =$$

$$= [(g(x) - g(z))x + (g(1-x) - g(1-z))(1-x)] + k \mid z - x \mid x > 0;$$

by (5.17) and $k > 0$, $\mid z - x \mid > 0$, and $x > 0$.

References

[1] Baillon, A. (2017). Bayesian markets to elicit private information. *Proceedings of the National Academy of Sciences.*, 114 (30) 7958-7962. https://doi.org/10.1073/pnas.1703486114

[2] Cvitanić, J., Prelec, D., Radas, S. and , Šikić, H. (2020) Incentive Compatible Surveys via Posterior Probabilities. *Theory of Probability and its Applications*, 65, 368–408.

[3] Cvitanić, J., Prelec, D., Riley, B. and Tereick, B. (2019) Honesty via Choice-Matching. *American Economic Review: Insights*, 1, 179–192.

[4] Cvitanić, J., Prelec, D., Radas, S. and , Šikić, H. (2018) Game of duels: Information-theoretic axiomatization of scoring rules. *IEEE Transactions on Information Theory,*, 65(1), 530-537.

[5] Dasgupta, A., and Ghosh, A. (2013) Crowdsourced Judgement Elicitation with Endogenous Proficiency. In Proceedings of the 22nd ACM International World Wide Web Conference (WWW'13). 319–330.

[6] Frongillo, R. and Witkowski, J. (2017) A Geometric Perspective on Minimal Peer Prediction, ACM Transactions on Economics and Computation, Vol. V, No. N, Article A.

[7] Miller, N., Resnick, P. and Zeckhauser, R. (2005) Eliciting Informative Feedback: The Peer-Prediction Method. *Management Science* 51, 1359–1373.

[8] Prelec, D. (2004) A Bayesian Truth Serum for Subjective Data. *Science* 306, 462–466.

[9] Prelec, D. (2021) Bilateral Bayesian Truth Serum: The $n \times m$ Signals case. Available at SSRN 3908446.

[10] Radanovic, G. and Faltings, B. (2013) A Robust Bayesian Truth Serum for Non-binary Signals. In Proceedings of the 27th AAAI Conference on Artificial Intelligence (AAAI'13). 833–839.

[11] Radanovic, G. and Faltings, B. (2014) Incentives for truthful information elicitation of continuous signals. In Proceedings of the AAAI Conference on Artificial Intelligence (AAAI'14), 28(1), 770-776.

[12] Waggoner, B., and Chen, Y. (2013) Information Elicitation Sans Verification. In Proceedings of the 3rd Workshop on Social Computing and User Generated Content (SC13).

[13] Witkowski., J. (2014.) Robust Peer Prediction Mechanisms. Ph.D. Dissertation. Department of Computer Science, Albert-Ludwigs-Universitat Freiburg.

[14] Witkowski, J., and Parkes, D. (2013) Learning the Prior in Minimal Peer Prediction. In Proceedings of the 3rd Workshop on Social Computing and User Generated Content (SC'13).

[15] P. Zhang and Y. Chen, Elicitability and knowledge-free elicitation with peer prediction, in AAMAS'14: Proceedings of the 2014 International Conference on Autonomous Agents and Multiagent Systems (Paris, 2014), IFAAMAS, Richland, SC, 2014, pp. 245–252; also available online from http://yiling.seas.harvard.edu/publications/.