Honesty via Choice-Matching

Jakša Cvitanić,* Dražen Prelec,† Blake Riley‡and Benjamin Tereick§

Abstract

We propose a class of mechanisms designed for eliciting honest responses to a multiple choice question (MCQ) when the truthfulness of these responses cannot be directly verified. Choice-matching mechanisms assign each respondent a score consisting of two terms: his score for predicting the answers of other respondents and the average prediction score of those respondents who give the same response to the MCQ. These mechanisms are truth-inducing when beliefs of respondents with the same truthful answers are sufficiently similar. We argue that our mechanisms are more suitable for practical implementation than existing alternatives. Going beyond predictions, choice-matching is a general method for making stated preferences incentive-compatible, by linking them to revealed ones.

Key words: Proper scoring rules, Bayesian Truth Serum, Peer Prediction, 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 10-08219.

†MIT, Sloan School of Management, Department of Economics, Department of Brain and Cognitive Sciences. E-mail: dprelec@mit.edu. Supported by Intelligence Advanced Research Projects Activity (IARPA) via the Department of Interior National Business Center contract number D11PC20058.

‡University of Illinois, Department of Economics. E-mail: blake.j.riley@gmail.com

§Erasmus School of Economics, Rotterdam. E-mail: tereick@ese.eur.nl. Supported by ERC Starting Grant 638408.

1 Introduction

A firm that wants to learn about how its customers value a new product, a government agency asking an external expert to evaluate the long-term impact of a policy, and an economist working with self-reported panel-data, all need to rely on claims made by individuals whose honesty may be in doubt.

In this paper,[1] we propose a new class of incentive-compatible "choice-matching" mechanisms for situations like these. The idea is to link explicit opinions and judgments, or "type-declarations," with an auxilliary game that reveals types, but only implicitly via a separating equilibrium with respect to some other action. In our canonical example, type-declarations are made by answering a multiple choice question (MCQ), and the auxilliary game asks these same respondents to predict how often each answer was taken by all other respondents. Predictions are then scored for accuracy using a standard proper scoring rule, and answers to the MCQ are credited with the average prediction accuracy score of all respondents who endorsed that particular answer. We prove that this mechanism is truth-inducing under standard assumptions, namely, that respondents are risk-neutral Bayesian maximizers who have a common prior and update their beliefs in an impersonal fashion. However, the guiding idea behind choice-matching is more general, and the auxilliary game can involve other types of questions or actions.

Incentives for non-verifiable MCQs based on respondents' predictions of the answer distribution were first introduced by Prelec (2004), with the Bayesian Truth Serum (BTS) mechanism. Under BTS scoring, every strict Bayes-Nash equilibrium is type separating (Cvitanic et al., 2016) for samples greater than a number N, but N itself is dependent on the common prior, which the planner may not know. This large sample requirement along with the non-transparent nature of the scoring formula has stimulated search for alternatives, especially in the computer science community (e.g. Witkowski and Parkes, 2012a, Radanovic and Faltings, 2014, Zhang and Chen, 2014).[2] Like these recent approaches, our choice-matching works with small samples. Its major practical advantages are the simplicity of the scoring principle and robustness with respect to underlying assumptions.

In the next section, we illustrate choice-matching using the example of consumer evaluations of a trial product. In section 3, we present the model and the main results and contrast our method to alternatives. In section 4, we explain how the choice-matching payment rules can be used without asking for predictions of other respondents' answers. For instance, we can ask for predictions of a verifiable random variable (such as GDP or future sales) if it is plausible that these predictions should be correlated with answers to the MCQ. Going beyond prediction, a generalized version of choice-matching can deal with situations in which true answers are entirely determined by preferences and not by beliefs. In the appendix, we discuss how a budget-balanced version of our mechanisms can be used to discourage collusion among respondents.

[1] This paper integrates results from three independent, previously unpublished documents: Cvitanić and Prelec (2014), Riley (2014) and Tereick (2016).

[2] An advantage of the BTS, however, is that its equilibrium payments rank respondents according to their expertise (Prelec 2004, Cvitanic et al. 2016), a property exploited by Prelec et al. (2017) to improve upon majority voting.

2 Example

Suppose that a firm hopes to launch a new product and gathers a sample of respondents to evaluate a trial version. After a testing phase, respondents are asked to provide a rating on a scale from 1 to 5. The firm would like to know the true percentage of each of the five ratings, but may be worried that respondents might not provide honest responses, for lack of effort, or because they feel obliged to endorse the product. The firm wants to design a mechanism that assigns a score to each respondent such that he maximizes his (expected) score by-responding truthfully, given that everyone else tells the truth. It is simple to show that rules based only on responses to the MCQ cannot achieve this aim (Radanovic and Faltings, 2013; Cvitanić, et al. 2016). Therefore, the MCQ must be supplemented with an auxiliary question. In the canonical version of choice-matching, this auxiliary question asks each respondent to predict the share of each of the five ratings submitted by other respondents in the sample. Respondents are then matched to those respondents who choose the same answer to the original question.

For concreteness, let $x^r = (x_1^r, ..., x_5^r)$ denote the rating of respondent r, where $x_k^r = 1$ if r's answer to the question is "k" and $x_k^r = 0$ else. For instance, if r rates the product with only one star, that response is coded as $x^r = (1,0,0,0,0)$. The probability vector $y^r = (y_1^r, .., y_5^r)$ is the prediction of respondent r, where y_k^r is the predicted share of answer k.

The "prediction score" that respondent r receives is some function $S(y^r, \bar{x}^{-r})$ of his prediction vector and the actual ratings distribution $\bar{x}^{-r}$, excluding his own rating from the average. Using strictly proper scoring rules, the firm can choose the function S such that $ES(y,X)$ is maximized for $y = E[X]$ (Savage, 1971, Gneiting and Raftery, 2007). We then write $\bar{S}(x^r, x^{-r}, y^{-r})$ for the average prediction score achieved by respondents other than r who report the same rating as respondent r. If no respondent has selected the same rating as r, then $\bar{S}$ is undefined and respondent r receives a score of zero. Importantly, respondent r also receives zero if any other rating is not chosen by at least one respondent. This "trigger device" cannot be manipulated, as we explain in section 3.

In the alternative state, if all five possible evaluations are given by at least one respondent other than r, respondent r receives

$$\lambda S(y^r, \bar{x}^{-r}) + (1-\lambda)\,\bar{S}(x^r, x^{-r}, y^{-r})$$

with $\lambda \in (0,1)$ an arbitrary weight assigned to r's prediction score by the survey planner.

Why might we expect this payment formula to be truth-inducing? To fix ideas, image two states of nature, one in which the tested product has high quality and one in which it has low quality. In the high-quality state, we should expect a larger share of higher ratings and the difference between the two states should be especially pronounced at the extremes. In this case, respondents who evaluate the product differently should have different beliefs about the distribution of ratings. A respondent who rates the product at a 5, should evaluate the-high-quality world as more likely than a respondent who rates the product at a 4 and much more likely than a respondent who rates the product at a 1.

If one's rating is the only piece of evidence to distinguish between the two states, the beliefs of those respondents with the same rating should be (roughly) identical. When facing the above incentives, each respondent should therefore expect that those respondents with the same rating will – on average – score higher. This provides the desired truth-telling incentives: Since we use strictly proper scoring rules to construct S, every respondent has incentives to provide

their true prediction. Furthermore, since each respondent should expect that those respondents who provide the same rating get the highest prediction scores, they have an incentive to be honest with regard to the MCQ as well.

3 The Model and the Results

3.1 The setup

To generalize our example, let $A = \{1, \ldots, M\}$ denote the set of possible answers to a MCQ. Random variables X_i^r take value zero for all choices $i \in A$ not declared by respondent r, and equal 1 if i is the answer declared by r. We refer to vector X^r as the type-declaring variable. Each respondent r also provides a $y-$response denoted by random variables Y_i^r. We refer to Y^r as the type-separating variable. In this section, the variables Y_i^r represent a point prediction of the average answer of other respondents, $\bar{X}_i^{-r} = \frac{1}{N-1}\sum_{s \neq r} X_i^s$ for each $i \in A$. The respondent's **type** is the pair (T^r, P^r), consisting of his true answer T^r to the MCQ as well as a probability measure P^r on the distribution of true answers in the population (not conditioning on his own true answer). $E^r[.]$ is the expectation operator corresponding to P^r. We sometimes refer to the true answer t^r as the ***x*-type** of respondent r and, when there is no ambiguity, simply as his type. The realizations of X_i^r and Y_i^r are denoted x_i^r, y_i^r, the realization of T^r is t^r and so on.

A respondent's score will depend on the type-declaration and predictions of all respondents, and the following choice-matching trigger:

Definition 1. The **matching trigger** $\mathcal{M}^r$ of respondent r is the event in which each answer option is chosen by at least one respondent other than r. In other words, $\mathcal{M}^r$ occurs if and only if there is no $i \in A$ with $\bar{x}_i^{-r} = 0$. Furthermore, $\mathcal{E}^r$ is r's **matching trigger under honesty**, that is $\mathcal{E}^r$ denotes that each $i \in A$ is the honest answer of at least one respondent other than r.

Assumptions A1-A4 below are sufficient, but not necessary, for choice-matching to be honesty-inducing. Subsection 3.3 discusses ways in which they can be relaxed.

Assumption 1. ***Common prior.*** *There exists a common prior on the distribution of x-types in the population. That is, it is common knowledge among individuals that for all respondents r, we have $P^r(.) = P$ and $E^r[.] = E[.]$ for some probability measure $P(.)$ and connected expectation operator $E[.]$.*

Importantly, this assumption does not imply that the survey planner knows the prior P.

Assumption 2. ***Non-degeneracy.*** *For any respondent r and any realization t^r: $P^r(\mathcal{E}^r \mid T^r = t^r) > 0$.*

Each r believes that with positive probability each answer to the MCQ is the x-type of at least one other respondent. While non-degeneracy is a technical assumption, it cannot be relaxed without changing our method substantially. Our method thus requires a minimal amount of care in the survey design: The planner should not include MCQ answer options for which respondents are sure that they will not be endorsed by anyone.

To simplify notation, we write $y^{r,k}$ for the conditional expectation $E^r\left[\frac{1}{N}\sum_{s' \neq r} T^{s'} \mid t_k^r = 1, \mathcal{E}^r\right]$. This is r's expectation of the average x-type – conditioning on $\mathcal{E}^r$ – when r's true answer to the MCQ is $k \in A$. We then assume:

Assumption 3. ***Stochastic relevance.*** *For any two respondents r, s and any answer options, $k, \ell \in A$:*

$$y^{r,k} \neq y^{s,\ell} \text{ if } k \neq \ell$$

Stochastic relevance is a mild requirement (Miller et al., 2005). Essentially, it states that there is something to learn about the responses of others from your own response. For our main result, we also assume the converse of stochastic relevance:

Assumption 4. ***Impersonal updating.*** *For any two respondents r, s and any answer option, $k \in A$:*

$$y^{r,k} = y^{s,k}$$

Impersonal updating is a more demanding assumption, stating that all respondents who share the same x-type have identical posteriors. Under a common prior and Bayesian updating, it is implied when individual x-types follow a multinomial distribution with unknown probability parameters. In such a model, we can imagine that there is a general population with unknown frequencies $p = (p_1, p_2, \ldots, p_M)$ of each type and respondents think of themselves as a random draw from this overall population.[3]

3.2 Inducing Honesty via Choice-Matching

In the current subsection, we will assume that assumptions A1-A4 hold. Under these assumptions, we can model the strategic setting induced by our payment rule as a Bayesian game in which a respondents' true answer completely determines his type. In this game, a pure strategy for respondent r is a function $\sigma(t^r) = (\sigma_x(t^r), \sigma_y(t^r))$ that maps his type to a response (x^r, y^r). The profile of all respondents' pure strategies is denoted $\sigma(t)$, with entries $\sigma^r(t^r)$, and the profile excluding player r is denoted $\sigma^{-r}(t^{-r})$. We consider only pure strategies and suppose that each respondent maximizes the expected value of his score, conditional on his type.

Given a real-valued payment rule $R(\sigma^r(t^r), \sigma^{-r}(t^{-r}))$, we call a set of response strategies a (Bayesian) **Nash Equilibrium**, **NE**, if, for any-responses $(x, y) \neq (\sigma^r(t^r))$, we have

$$E\left[R\left(\sigma_x^r(t^r), \sigma_y^r(t^r); \sigma^{-r}(t^{-r})\right) - R\left(x, y; \sigma^{-r}(t^{-r})\right) \mid T^r = t^r\right] \geq 0 \tag{3.1}$$

That is, by deviating in responses (x, y), player r would be worse off (in expectation) than by not deviating. If the inequality is strict we speak of a **strict NE**. An NE is strict in x if the inequality is strict whenever $x \neq \sigma_x^r(t^r)$ and, analogously, it is strict in y if the inequality is strict whenever $y \neq \sigma_y^r(t^r)$.

We further call a strategy profile **honest** if every respondent r reports $x^r = t^r$ and $y^r = y^{r,k}$ if and only if k is r's honest answer. This definition of honesty differs from the previous literature in which honest y-responses are not conditioning on the event $\mathcal{E}^r$ (with the exception of Baillon, 2017). For many reasonable priors there is only a slight

[3]For a more extensive discussion of the empirical meaning of stochastic relevance and impersonal updating as well as for supporting evidence from psychology see Prelec (2004) and Baillon (2017).

difference between these two definitions. A payment rule which has a NE that is strict and honest is called **strictly incentive compatible**. It is called strictly incentive compatible in x if the NE is honest and strict in x (and analogously for y).

Definition 2. Given an integer $I > 1$, we say that functions $f(p;j)$, $j = 1,...,I$, form a **strictly proper scoring rule** (SPSR) if, for any probability vectors $p = (p_1,\ldots,p_I)$, $q = (q_1,\ldots,q_I)$, $q \neq p$, we have:

$$\sum_{j=1}^{I} p_j f(p;j) > \sum_{j=1}^{I} p_j f(q;j) \tag{3.2}$$

There are infinitely many functions $f(.)$ which satisfy inequality 3.2, most notably the quadratic scoring rule $f(p;j) = 2p_j - \sum_{i=1}^{M} p_i^2$ and the logarithmic scoring rule $f(p;j) = \log p_j$. In the latter case, 3.2 is known as the Gibbs inequality. Given a series of categorical variables, we can use scoring rules to elicit the expected value of the series' average. For variables x_j^1, ...,x_j^N, $j = 1,...,I$ where x_j^s takes value 1 for some j and value 0 otherwise, we let:

$$\frac{1}{N}\sum_{s=1}^{N}\sum_{j=1}^{I} x_j^s f(p;j) = \sum_{j=1}^{I} \bar{x}_j^s f(p;j) := S(p,\bar{x}) \tag{3.3}$$

for a given SPSR f. With this construction, we make choice-matching incentive compatible in y.

Definition 3. Consider a respondent r who declares k to be his true answer. Let f be a strictly proper scoring rule and S as defined in 3.3. Given $\lambda \in (0,1)$, a payment rule $R_{S,\lambda}$ induces **choice-matching** if

(a) In the event $\mathcal{M}^r$:

$$R_{S,\lambda}(x^r,y^r) = \lambda S(y^r) + (1-\lambda)\bar{S}^{-r}(x^r)$$

where $\bar{S}^{-r}(x^r)$ is the average prediction score achieved by those respondents other than r who submit $x^s = x^r$:

$$\bar{S}^{-r}(x^r) = \frac{\sum_{s\neq r} x^r \cdot x^s \, S(y^s)}{\sum_{s\neq r} x^r \cdot x^s}$$

(b) and $R_{S,\lambda}(x^r,y^r) = 0$ otherwise.

In words, if all M possible answers are declared by at least one respondent other than r, choice-matching assigns him a score that is a weighted average of his own prediction score and the prediction score of those respondents who declare the same x-choice. Otherwise, he receives zero.

We now come to the main result of the paper.

Proposition 1. *Assume A1-A4 and at least two more respondents than the number of possible answers in the multiple-choice question. Then any payment rule $R_{S,\lambda}$ that induces choice-matching is strictly incentive compatible for every $\lambda \in (0,1)$.*

Proof. By the construction of S, choice-matching is incentive compatible in y. To see that choice-matching is also incentive compatible in x, fix a respondent r with honest answer k. We only need to consider $\bar{S}^{-r}$. Suppose that all

the players other than r play the NE strategies. The difference in expected information score for player r between non-deviation and deviating from t^r to some other response x^r with $x^r_i = 1$ is:

$$Pr(\mathcal{E}^r \mid T^r) \times (1-\lambda) E\big[S^{-r}(t^r_x) - \bar{S}^{-r}(x^r) \mid T^r = t^r, \mathcal{E}^r\big]$$

From non-degeneracy, $Pr(\mathcal{E}^r \mid T^r) \times (1-\lambda) > 0$. Furthermore, due to impersonal updating and via construction of $R_{S,\lambda}$:

$$E\big[\bar{S}^{-r}(t^r_x) - \bar{S}^{-r}(x^r) \mid T^r = t^r, \mathcal{E}^r\big] = E\Big[S(y^{r,k}) - S(y^{r,i}) \mid T^r = t^r, \mathcal{E}^r\Big]$$

By construction of S from a SPSR, we must have $E\big[S(y^{r,k}) - S(y^{r,i}) \mid T^r = t^r, \mathcal{E}^r\big] > 0$ for any $i \neq k$. Thus, $R_{S,\lambda}$ is strictly incentive compatible in x as well. ■

The matching trigger prevents a respondent from influencing whether or not he is matched. From a respondent's perspective, the distribution of other answers either does or does not have a vacancy, i.e., an answer not declared by anyone. If there is such a vacant anwer (or more than one), then either the respondent declares this answer and is not matched (as he is the only one declaring it), or he declares a different answer, in which case no one chooses the vacant answer, and again he is not matched. If there are no vacant answers, then he is matched no matter what he does.

This idea was first employed by Baillon (2017), for the binary case. As pointed out by him, the sample must have at least two more respondents than the number of possible answers to the MCQ, or $N > M+1$. If $N = M+1$, then the only way to prevent a vacancy after excluding one respondent is if other respondents each choose a different answer. Therefore, non-zero scores are possible only if there is a uniform distribution over answers. As this is common knowledge, all predictions will be uniform irrespective of respondent type, eliminating choice-matching incentives. Formally, stochastic relevance fails if $N = M+1$, so the requirement that $N > M+1$ in the proposition is in fact redundant.

3.3 Robustness: No common prior

Our main result holds in a model without a common prior P. Since a Bayesian game requires a common prior, we need to adjust our definition of incentive compatibility which we do in a similar manner as Witkowski and Parkes (2012b), Radanovic and Faltings (2014) and Baillon (2017). We call a rule R incentive compatible if:

$$E^r\left[R_{S,\lambda}\left(t^r, y^{r,k}; T^{-r}\right) - R_{S,\lambda}\left(x, y; T^{-r}\right) \mid t^r_k = 1\right] > 0$$

for any responses $(x,y) \neq \left(t^r_x, y^{r,k}\right)$ and each respondent r. We continue to assume non-degeneracy and stochastic relevance. In contrast to subsection 3.1 however, it is now possible that $y^{r,k} \neq y^{s,k}$ for two respondents r, s (violating impersonal updating). In order to measure the divergence between these posterior expectations, we use a **divergence function** $d(x,y) : \Delta^m \times \Delta^m \to \mathbb{R}_+ \cup \{\infty\}$ mapping two probability vectors into the extended positive real line. We require that $d(x,y) = 0$ if and only if $x = y$, and define:

Definition 4. Posterior expectations $y^{r,k}$ of respondents satisfy **closeness with respect to divergence** d if:

$$d\left(y^{r,k}, y^{s,k}\right) < d\left(y^{r,k}, y^{s',k'}\right)$$

for any respondents r, s and s', and for all $k, k' \in A$ with $k \neq k'$.

That is, under closeness the posteriors of two respondents with the same x-type can differ due to different prior assumptions, but they still agree more (as measured by d) than two individuals with different x-types. Put differently, respondents may bring different information about the distribution of x-types to the survey, but the information about their own honest response dominates differences with regard to the remaining information. We can exploit such a structure via a well-known relationship between scoring rules and divergence measures:

Definition 5. An SPSR $f(q;j)$ is **effective** with respect to divergence function d if for probability vectors p^1, p^2, q:

$$d\left(p^1, q\right) \leq d\left(p^2, q\right) \Longleftrightarrow \sum_i q_i \, f(p^1; i) \geq \sum_i q_i \, f(p^2; i)$$

That is, when p^1 is "closer" according to d to the true probability q than p^2 is, then the expectation of scoring rule $f(p, i)$ using true probability q is higher for $p = p^1$ than for $p = p^2$.

Proposition 2. *Suppose that closeness holds with respect to divergence d and that SPSR $f(q, j)$ is effective with respect to d. Then, choice-matching using $S(y^r)$ as defined in 3.3 is strictly incentive compatible.*

Proof. To simplify the exposition of this proof, we write $\tilde{E}^r = E^r\left[. \mid \mathcal{E}^r, T^r\right]$ for player r's expectation operator conditional on his x-type and conditional on r's matching trigger under honesty. Compared to proposition 1, incentive compatibility for y-responses is unchanged. For x-responses, consider an individual respondent r with honest answer k and suppose that all other respondents respond truthfully. Then, the difference of the expected payoff between reporting t^r and deviating to an x-response x^r with $x^r_{k'} = 1$ for some $k' \neq k$ is:

$$\tilde{E}^r\left[\bar{S}^{-r}(t^r_x) - \bar{S}^{-r}(x^r)\right] \geq \min_{s,\, s' \neq r} \tilde{E}^r\left[\sum_{i=1}^{M} \bar{x}_j^{-r}\left[f\left(y_i^{s,k}, i\right) - f\left(y_i^{s',k'}, i\right)\right]\right]$$

$$= \min_{s,\, s' \neq r} \sum_{i=1}^{M} \tilde{E}^r\left[\bar{x}_j^{-r}\right]\left(f\left(y_i^{s,k}, i\right) - f\left(y_i^{s',k'}, i\right)\right)$$

$$= \min_{s,\, s' \neq r}\left[\sum_{i=1}^{M} y_i^{r,k} f\left(y_i^{s,k}, j\right) - \sum_{j=1}^{M} y_j^{r,k} f\left(y_j^{s',k'}, j\right)\right] > 0$$

where the last inequality follows since by definition of closeness and the effectiveness relation between d and f, the bracketed term must be strictly positive for all respondents r, s, s'. ∎

Friedman (1983) and Nau (1985) characterize scoring rules which are effective with respect to a *metric* d, including the well-known quadratic scoring and spherical scoring rules. In addition to their contribution, it is easy to verify that

the logarithmic scoring rule is effective with respect to d when d is defined as relative entropy (which is not symmetric and hence not a metric). Thus, for two plausible divergence measures of posteriors, relative entropy and quadratic distance, the two most commonly used scoring rules can be used to allow a strictly separating NE with choice-matching.

3.4 Alternative Methods

As mentioned at the start, there is now a growing literature on mechanisms for eliciting non-verifiable opinions.[4] Miller et al.'s (2005) "peer prediction" assumes that the planner knows the common prior and can compute everyone's prediction after receiving the answers to the MCQ. This idea has been adapted by Zhang and Chen (2014) to a setting in which the prior is unknown to the survey planner. Their "Generalized peer prediction" lets respondents first report an answer and after receiving the answer of another respondent make a prediction. They then receive a score for their own prediction and for the prediction of the respondent who received their answer. This two-stage structure is a practical disadvantage for large-scale surveys in which respondents are typically contacted once.[5]

A series of papers have tried to eliminate the prediction question (Radanovic and Faltings, 2015, Radanovic et al., 2016, Shnayder et al., 2016, Agarwal et al., 2017, Liu and Chen, 2017). These methods either need to make distributional assumptions on the prior, or extract the prior using machine learning and require large amounts of data.

More in the spirit of the current approach is the "Robust bayesian truth serum" (RBTS) of Witkowski and Parkes (2012a). The RBTS can work in a setting with only 3 respondents. Its payment structure cleverly exploits a symmetry property of the quadratic scoring rule. It is however difficult to extend this idea to non-binary settings. Furthermore, the payment rule of the RBTS is quite complicated, so that it cannot be considered an improvement over the BTS in this regard. Baillon's (2017) "Bayesian market" constitutes a step forward with respect to accessibility. In the Bayesian market, answers to the MCQ and predictions by respondents are translated to buying and selling decisions, which may be more natural to respondents than engaging with scoring rules. As in the case of the RBTS, this idea can however not be easily extended to situations with more than two possible answer options. The method by Radanovic and Faltings (2013) works for any number of possible answers and is mathematically simpler than both the BTS and the RBTS, but it requires an additional assumption on the information structure: For each $k, \ell \in A$ with $k \neq \ell$, it needs to hold that $y_k^{r,k} > y_k^{r,\ell}$, the so-called self-predicting assumption.

Closest to choice-matching is the "Divergence-based bayesian truth serum" (DIV) of Radanovic and Faltings (2014) and the "minimum truth serum" suggested by Riley (2014). Expressed in our notation, the latter is given by the following payment formula:

$$R_S^{MTS}(x^r, y^r) = \begin{cases} \min\left\{S(y^r), \bar{S}^{-r}(x^r)\right\} & \text{in the event } \mathcal{M}^r \\ S(y^r) & \text{otherwise.} \end{cases}$$

[4] A minor difference between methods such as the BTS and choice-matching and the methods proposed in the computer scence literature discussed in this subsection is that in the latter, respondents are often asked to predict the answer of a single agent. However, these mechanisms are usually isomorphic to one which asks about an average answer.

[5] In principle, this problem could be solved by using a "strategic implementation" where participants are contacted once and make a hypothetical second prediction conditional on all answers they might receive from another respondent. However, this adjustment would complicate their method substantially, especially if M is large.

where S is constructed from a strictly proper scoring rule f, as in choice-matching. In contrast to choice-matching, the minimum truth serum is not always strictly incentive compatible without further assumptions on the updating process.[6] The DIV of Radanovic and Faltings (2014), in its "non-parametric" version, assigns a respondent r to two peer agents and penalizes r if one of the peers gives the same answer to the MCQ, while the other disagrees, and yet the prediction of the latter is closer to r's than the one of the former. DIV is incentive compatible under very similar conditions like the ones we have stated in subsection 3.3.

Looking ahead to implementation, choice-matching has a major advantage over non-parametric DIV in that the score for the type-declaration does not depend on the respondent's prediction (the mechanism is decomposable, according to the definition of Radanovic and Faltings, 2013). Therefore, it is possible to let (some) respondents decide whether or not to make a prediction, as proposed by Riley (2014). We can redefine the matching trigger such that for each answer option $i \in A$, there is a respondent $s \neq r$ who answers i *and* submits a prediction. Respondents who decide to submit a type declaration only, will receive the score $\lambda \bar{S}(x^r)$, provided that the (redefined) event $\mathcal{M}^r$ occurs. To ensure that respondents still maximize their expected score by submitting a prediction, it is sufficient to choose a scoring rule with a positive lower bound to construct S. This construction allows respondents to skip the prediction question if they prefer to do so, while leaving the truth-telling incentives of the remaining respondents intact – as long as enough respondents submit a prediction such that $\mathcal{M}^r$ occurs with positive probability.

A further implication is that a planner may decide to not even ask some respondents for a prediction, once enough predictions have been collected by previous respondents. This way, the planner can reduce the difficulty of the task of respondents who enter the survey at a later stage. These later responses can thus be collected in a faster manner and, consequently, at lower cost.

4 Choice-Matching Generalized

Making predictions about other respondents' answers is an attractive default for our auxiliary question. It gives rise to a plausible structure of beliefs which are sufficient for a truthful equilibrium. Furthermore, respondents can be paid as soon as all responses have been selected.

However, on some occasions using predictions may not be ideal. First, it may be that the stochastic relevance of individual answer types is weak since the distribution of types in a population is well known, for instance if a survey asks respondents for their gender. Second, some respondents may have trouble understanding the payments made according to proper scoring rules. Finally, in some situations respondents could expect that the predictions of respondents with different answers to the MCQ will be more accurate. For instance, if a survey asks about field of study and highest attained degree, it may be optimal to report a PhD in quantitative social science. In this section, we show that there is a general principle behind choice-matching which can be employed by methods which do not rely on a prediction question.

[6] For example, let $\frac{1}{N} > \varepsilon > 0$ and suppose that $y^{r,1} = \left(\frac{N-1}{N} + \varepsilon, \frac{1}{N} - \varepsilon\right)$ and $y^{r,2} = \left(\frac{N-1}{N}, \frac{1}{N}\right)$, where we slightly abuse notation to let $y^{r,k}$ be the posterior expectation without conditioning on $\mathcal{E}^r$. In this case, $S\left(y^{r,1}\right) < S\left(y^{r,2}\right)$ whenever $\mathcal{M}^r$ occurs. Respondents whose honest answer is 1 then do not have strict truth-telling incentives since $\min\left\{S(y^r), \bar{S}^{-r}(x^r)\right\} = S(y^r)$ on $\mathcal{M}^r$, regardless of their x-response.

To formalize the general principle, we first introduce real-valued utility-functions $u_k(y^r, x^{-r}, y^{-r})$ for $k \in A$ that depend on all variables except a respondent's type-declaration, x^r. Since y^r is not necessarily a prediction, we let the y-responses be taken from some general response set Ω.

Definition 6. Let G be a (Bayesian) game given by the collection of the respondent set N, a set of potential type declarations A, a set of potential y-responses Ω, a prior P and utilities $\{u_k\}_{k\in A}$. The game G is **type-separating** if there is a profile σ such that for every respondent r and every $k \in A$:

(i) $\sigma_x^r(t^r) = t^r$

(ii)$\sigma_y^r(t^r) = y^{*k}$ if and only if $t_k^r = 1$ and

(iii) $E\left[u_k\left(y^{*k}, t^{-r}, \sigma_y^{-r}\left(t^{-r}\right)\right) \mid t_k^r = 1, \mathcal{E}^r\right] > E\left[u_k\left(y, t^{-r}, \sigma_y^{-r}\left(t^{-r}\right)\right) \mid t_k^r = 1, \mathcal{E}^r\right]$ for any $k \in A$, $y \in \Omega$, $y \neq y^{*k}$.

In words, in a type-separating game there is a profile $\sigma^*(.)$ in which respondents declare their types truthfully (condition (i)) and in which respondents with identical answers give the same y-response and respondents with different answers give different y-responses (condition (ii)). Furthermore, this profile is an equilibrium (condition (iii)). Importantly, this equilibrium is not strict in x^r, since $u\left(y^r, x^{-r}, y^{-r}\right)$ does not depend on x^r at all.

As explained below in more detail, in our model from section 3 the type-separating game is the game induced by the prediction score alone.

Proposition 3. *Let $G = \langle N, , A, \Omega, \{u_k\}_{k\in A}, P\rangle$ be a type-separating game. Under assumptions A1-A2, any payment rule which induces a game $\langle N, , A, \Omega, \{V_k\}_{k\in A}, P\rangle$ in which*

$$V_k(x^r, y^r, x^{-r}, y^{-r}) = \lambda u_k\left(y^r, x^{-r}, y^{-r}\right) + (1-\lambda)\,\bar{u}_k\left(x^r, x^{-r}, y^{-r}\right)$$

in case of $\mathcal{M}^r$, where $\lambda \in (0,1)$ and

$$\bar{u}_k\left(x^r, x^{-r}, y^{-r}\right) = \frac{\sum_{s\neq r} x^r \cdot x^s\, u_k\left(y^s, x^{-s}, y^{-s}\right)}{\sum_{s\neq r} x^r \cdot x^s}$$

and on the complement of $\mathcal{M}^r$:

$$V_k(x^r, y^r, x^{-r}, y^{-r}) = 0$$

is strictly incentive compatible.

Proof. The proof is almost identical to the special case in section 3. The difference in expected utility for respondent r between non-deviation and deviating from t^r to some other response x^r with $x_i^r = 1$ is:

$$P(\mathcal{E}^r \mid t_k^r = 1) \times (1-\lambda)\, E\left[\bar{u}_k\left(t^r, x^{-r}, y^{-r}\right) - \bar{u}_k\left(x^r, x^{-r}, y^{-r}\right) \mid t_k^r = 1, \mathcal{E}^r\right]$$

From non-degeneracy, $P(\mathcal{E}^r \mid t_k^r = 1) \times (1-\lambda) > 0$. Furthermore, due to construction of $\bar{u}\left(x^r, x^{-r}, y^{-r}\right)$:

$$\bar{u}_k\left(t^r, x^{-r}, y^{-r}\right) - \bar{u}_k\left(x^r, x^{-r}, y^{-r}\right)$$

$$= u_k\left(y^{*k}, x^{-r}, y^{-r}\right) - u_k\left(y^{*i}, x^{-r}, y^{-r}\right)$$

for $i \neq k$. Since the game G is type-separating, we have $E\left[u_k\left(y^{*k}, x^{-r}, y^{-r}\right) - u_k\left(y^{*i}, x^{-r}, y^{-r}\right) \mid t_k^r = 1, \mathcal{E}^r\right] > 0$ which gives the required result. ■

To make the proposition concrete, we first explain how it covers our model in section 3. In this model, the type-separating game is based on the prediction score:

$$u_k\left(y^r, x^{-r}, y^{-r}\right) = \begin{cases} S\left(y^r, x^{-r}, y^{-r}\right) & \text{in the case of } \mathcal{M}^r \\ 0 & \text{otherwise.} \end{cases}$$

Given that all respondents report their answers honestly, it is a strict best response to set $y^r = y^{r,k}$, so that $y^{*k} = y^{r,k}$ for all r and all $k \in A$, which makes the game induced by the prediction score a type-separating game. Under this definition, the function $V(x^r, y^r, x^{-r}, y^{-r})$ from proposition 3 equals the payment rule of choice matching R_λ^ρ as defined in definition 3.

As an example of how choice-matching can be applied when the second question is not a prediction at all, consider again the example from section 2. As a reward for their participation in the product trial, each respondent could be allowed to participate in a "product lottery". First, they choose from a list of products the firm has already launched. If each star rating is chosen by one respondent other than r, respondent r receives the product they choose with probability λ and otherwise receive the product chosen by a respondent randomly selected among those giving the same star rating.

Formally, we can represent the list as a set $L = \{1, ..., ..., \ell\}$, when y^r is a selection from L and $u_k\left(y^r, x^{-r}, y^{-r}\right)$ simply indicates how much a respondent of type k values the product (so that it does not in fact depend on x^{-r} and y^{-r}). If respondents then receive their own selection with probability λ and otherwise receive the selection of another respondent who gave the same rating, their expected utility equals the expression $V(x^r, y^r, x^{-r}, y^{-r})$ from proposition 3. Thus, when the game induced by asking for a selection from the list is type separating,[7] the proposition tells us that the product lottery makes truth-telling a strict Bayes-Nash equilibrium.

Using individual selections from a list as a type-separating game could find further applications in the study of decision-making under risk. These experiments usually let respondents choose among a variety of risky gambles, one of which is used to determine payments. Oftentimes, there is a final survey which asks about behavior outside the laboratory. For instance, the survey may ask which types of insurance a respondent possesses. While responses to such questions previously had to be taken at face value, choice-matching makes it possible to incentivize them by paying a respondent according to the gambles chosen by respondents who give the same answer to the survey question. This can make those answers more credible and allow more reliable inferences about the connection of behavior in- and outside

[7] For this, we need that respondents with different honest star ratings of the trial product select different products from L and those with the same true rating select the same. As in section 3, it is not necessary that this requirement holds strictly. It can be relaxed in a similar fashion as we relaxed the impersonal updating assumption in subsection 3.3.

the laboratory.

Proposition 3 also makes apparent that we could choose y^r to be a prediction not about other respondents' type declarations, but about verifiable random variables. For example, suppose that the original MCQ asks respondent to rate the statement "the fiscal stimulus applied by the Obama administration since 2009 accelerated the recovery of the US economy after the subprime mortgage crisis" on a scale from 1 to 5, where 1 means strong disagreement and 5 means strong agreement. An auxiliary question could be a prediction about macroeconomic indicators, such as GDP, interest rate or unemployment rate. Since the assessment of the fiscal stimulus should correspond to a specific macroeconomic view, respondents should plausibly expect that those respondents who evaluate the stimulus in the same manner, adhere to the "correct" macroeconomic conception and should therefore also be better forecasters of macroeconomic variables.

Relating to our discussion at the end of section 3, we can further see that we can reduce the burden on respondents by requesting predictions not about *all* possible answer options, but only over convenient subsets. In the example from section 2, the company could ask respondents to predict the shares of ratings higher and lower than 3 stars. This would not affect the choice-matching incentives.

5 Conclusion

In this paper, we have proposed a simple way to elicit honest responses from many agents to a multiple choice question, even if these answers cannot be verified and the planner has no prior knowledge about the distribution of honest answers in the population. Compared to alternative methods in the literature, our method is easy to explain to respondents. This is relevant from a practical standpoint: In empirical tests of the BTS, John et al. (2012) and Weaver and Prelec (2013), respondents were not informed about the payment structure but were only told that it was in their interest to be truthful. This has also been called the "intimidation method" (Frank et al., 2017). While the black-box presentation mode in these studies did change the answer distribution, it is reasonable to assume that transparent methods will be even more effective. A further practical advantage of choice-matching is that part of the respondent pool (potentially the majority) only needs to answer one question. This is particularly helpful in the design of large-scale online incentive syste-ms which currently rely on sophisticated machine learning techniques and non-transparent payment rules.

The practical features of choice-matching do not come at the cost of stronger assumptions on the underlying setting. To the contrary, we have shown that our method is honesty-inducing under fairly general assumptions. It is only needed that honest answers are informative and that posteriors of respondents with the same honest answer to the MCQ have beliefs which are more similar to each other than the beliefs of respondents whose honest answer to the MCQ differs.

Using predictions may not be ideal when there is strong public information about the distribution of types, when respondents do not understand proper scoring rules well or when respondents may differ with regard to prediction accuracy for other reasons than the information contained in their answer type. In this case, the general principle behind choice-matching can still be applied, whenever there is a task which induces separation among honest answers. We have illustrated potential applications in customer research or in the study of risk. The general insight is that to

design an incentive compatible reward scheme, the planner does not need to know *which* options respondents choose in a separating game, it is enough that she knows that separating strategies exist in order to align the game with the type declaration question.

While surveys have played a major role as a research tool in other social sciences, economists have traditionally been suspicious of stated preferences and beliefs since it is of no consequence for respondents if they do not answer truthfully. However, it is often that these unverifiable variables, which can tap the expressive range of ordinary langauge, that are precisely the variables of interest. By linking stated to revealed preferences and beliefs our method erases, in principle, the methodological boundary between these two types of data.

Appendix

Budget-Balancing

Throughout our paper, we have assumed that individuals maximize their score individually and have no means of colluding with each other. This may seem problematic, because our method is susceptible to an attack, in which respondents agree on a response and then each make the same prediction. It can be verified straightforwardly that such collusion is even a strict Bayesian equilibrium.

It should be noted that our trigger device, conditioning payment on the matching trigger $\mathcal{M}^r$, constitutes an impediment to such attacks since participants need to coordinate such that each answer is taken by at least two respondents. Against methods without the device, respondents could each report the exact same answer, making coordination much simpler.

Beside the coordination problem that colluding respondents need to solve, there is a way to make choice-matching robust to collusion among *all respondents* by using a **budget-balanced** version of it. We say that a scoring rule is budget-balanced if for any configuration of responses, the total score of respondents is zero. Given a choice-matching payment rule R, we define.

$$R^0(x^r, y^r) = R(x^r, y^r) - \frac{1}{N}\sum_{s=1}^{N} R(x^s, y^s)$$

Obviously, we have $\sum_{r=1}^{N} R^0(x^r, y^r) = 0$ for all N. When the underlying scoring rule $f(.)$ is bounded,[8] we further have $\lim_{n\to\infty} R^0(x^r, y^r) - R^0(\tilde{x^r}, y^r) = R(x^r, y^r) - R(\tilde{x^r}, y^r)$, since the subtracted term $\bar{R}$ does not depend on the responses of an individual respondent in the limit. Then, R^0 is strictly incentive compatible in x and strictly separating in y when R is.

Alternatively, if there are $N \geq M+3$ respondents, we can build N subsamples, each excluding one of the N respondents. We then calculate R for the responses of each subset. Each respondent receives the score from each of the $N-1$

[8] If we use *regular* scoring rules (Gneiting and Raftery, 2007) which only (potentially) assign infinite scores in case of a forecast of 0, we can restrict the respondents answers by allowing no predictions smaller than $\frac{1}{N}$. This does not affect incentives in a truthful equilibrium, because conditioning on $\mathcal{E}^r$, each respondent knows with certainty that $\bar{x}_k \geq \frac{1}{N}$ for all k.

subsamples of which he is part, deducted by the total score of the subsample of which he is not part. Formally:

$$R^1(x^r, y^r, x^{-r}, y^{-r}) = \frac{1}{N-1} \sum_{s \neq r} \left[R(x^r, y^r, x^{-r,s}, y^{-r,s}) - R(x^s, y^s, x^{-r,s}, y^{-r,s}) \right]$$

where $x^{-r,s}$ is the vector of x-responses, excluding the answers of respondents r and s. This formula is a generic device to achieve budget-balancing in an incentive compatible manner. In our opinion, it is advisable to use R^1 as a budget-balancing device only in small samples, when the limit result does not have force. If the number of respondents is large however, we recommend to use R^0 since this rule is easier to understand.

References

[1] Agarwal, A., Mandal, D., Parkes, D. and Shah, Nisarg. (2017) Peer Prediction with Heterogeneous Users. In Proceedings of the 18th ACM Conference on Economics and Computation.

[2] Baillon, A. (2017) Bayesian markets to elicit private information. Proceedings of the National Academy of Sciences, 114:30, 7958–7962.

[3] Cvitanić, J., Prelec, D., Radas, S. and Šikić, H. (2017) Incentive Compatible Surveys via Posterior Probabilities. Submitted.

[4] Cvitanić, J. and Prelec, D. (2014) Honesty Via Type-Matching. Working Paper.

[5] Frank, M. R., Cebrian, M., Pickard, G., and Rahwan, I. (2017) Validating Bayesian truth serum in large-scale online human experiments. PloS one, 12(5), e0177385.

[6] Friedman, D. (1983) Effective scoring rules for probabilistic forecasts. Management Science, 29, 447-454.

[7] Gneiting, T. and Raftery, A.E. (2007) Strictly Proper Scoring Rules, Prediction, and Estimation. Journal of the American Statistical Association, 102, 359–378.

[8] John, L.K., Loewenstein, G. and Prelec, D. (2012) Measuring the prevalence of questionable research practices with incentives for truth telling. Psychological Science, 23, 524-532.

[9] Kong, Y. and Schoenebeck, G. (2016) A framework for designing information elicitation mechanisms that reward truth-telling. arXiv preprint arXiv:1605.01021.

[10] Liu, Y. and Chen, Y. (2017) Machine-Learning Aided Peer Prediction. Proceedings of the 2017 ACM Conference on Economics and Computation.

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

[12] Nau, R.F. (1985) Should scoring rules be "effective"?. Management Science, 31, 527-535.

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

[14] Prelec, D., Seung, H. S., and McCoy, J. (2017) A solution to the single-question crowd wisdom problem. Nature, 541(7638), 532-535.

[15] 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).

[16] Radanovic, G. and Faltings, B. (2014) Incentives for truthful information elicitation of continuous signals. In Proceedings of the 28th AAAI Conference on Artificial Intelligence (AAAI 14).

[17] Radanovic, G. and Faltings, B. (2015) Incentive Schemes for Participatory Sensing. In Proceedings of the 2015 International Conference on Autonomous Agents and Multiagent Systems (AAMAS).

[18] Radanovic, G., Faltings, B. and Jurca, R. (2016) Incentives for Effort in Crowd-sourcing using the Peer Truth Serum. ACM Transactions on Intelligent Systems and Technology (TIST), 7.4.

[19] Riley, B. (2014) Minimum truth serums with optional predictions. In Proceedings of the 4th Workshop on Social Computing and User Generated Content (SC14).

[20] Shnayder, V., Agarwal, A., Frongillo, R. and Parkes, D. (2016) Informed Truthfulness in Multi-Task Peer Prediction. In Proceedings of the 2016 ACM Conference on Economics and Computation, 179–196.

[21] Tereick, B. (2016) Credible Truth-Telling Mechanisms For Subjective Truths (unpublished master's thesis). Tinbergen Institute, Netherlands.

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

[23] Weaver, R. and Prelec, D. (2013) Creating truth-telling incentives with the Bayesian truth serum. Journal of Marketing Research, 50, 289-302.

[24] Witkowski, J. and Parkes, D.C. (2012a) A Robust Bayesian Truth Serum for Small Populations. In Proceedings of the 26th AAAI Conference on Artificial Intelligence (AAAI 13).

[25] Witkowski, J., and Parkes, D.C. (2012b) Peer Prediction Without a Common prior. In Proceedings of the 13th ACM Conference on Electronic Commerce (EC12).

[26] Zhang, P. and Chen, Y. (2014) Elicitability and knowledge-free elicitation with peer prediction. In Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems, 245-252.