Honesty via Choice-Matching

Jakă Cvitanić,[*] Draěn Prelec,[†] Blake Riley[‡]and Benjamin Tereick[§]

Abstract

We introduce choice-matching, a class of mechanisms for eliciting honest responses to a multiple choice question (MCQ), as might appear in a market research study, opinion poll or economics experiment. Under choice-matching, respondents are compensated through an auxiliary task, e.g., a personal consumption choice or a forecast. Their compensation depends both on their performance on the auxiliary task, and on the performance of those respondents who matched their response to the MCQ. We give conditions for such mechanisms to be strictly truth-inducing, focusing on a special case in which the auxiliary task is to predict the answers of other respondents.

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-1810807.

[†] 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.

Disclaimer: The views and conclusions expressed herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/NBC, or the U.S. Government.

1 Introduction

More than other social sciences, economics favors the hard data revealed by agents' decisions, as opposed to surveys of their opinions or tastes. Yet subjective reports are often an irreplaceable source of information. Firms that want to know how consumers would value a hypothetical new product, government agencies asking external experts to assess the long-term impact of a policy, and economists working with self-reported panel-data all rely on answers provided by individuals whose effort or honesty may be in doubt. We propose a new class of incentive-compatible mechanisms for this problem that we call **choice-matching**.[1] The idea is to link explicit opinions and judgments, with an auxiliary task that reveals the respondents' "types", but only implicitly. In our canonical version, respondents answer a multiple choice question (MCQ), and the auxiliary task asks respondents to predict how often each answer was chosen by all other respondents.

A respondent's score is then a weighted sum of a prediction accuracy score and the average prediction accuracy score of all the respondents who endorse the same answer to the MCQ. We state conditions under which this mechanism is truth-inducing, the major one being that there is a one-to-one correspondence between predictions and answers to the MCQ. However, the guiding idea behind choice-matching is more general, and the prediction question can be replaced with other kinds of tasks or games.

We illustrate choice-matching with an example of consumer evaluations of a trial product in section 2. Section 3 presents the model and our results in the canonical version of our mechanism. Section 4 explains how the choice-matching payment rules can be used without asking for predictions of other respondents' answers, and discusses possible applications in marketing and experimental economics. Section 5 compares our method to previous proposals. Section 6 concludes.

2 Example

Suppose that a firm intends to launch a new product and asks respondents to evaluate a trial version on a scale from 1 to 5. A rating that best represents a respondent's true opinion is their **type**. The firm would like to know the actual percentage of each type, but may be worried that some respondents may

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

not answer honestly, for lack of effort or because they feel obliged to endorse the product. The firm wants to provide monetary incentives for honest answers. For these incentives to be strict, there must be some additional input to the mechanism beyond the answers themselves (Radanovic and Faltings, 2013; Cvitanić et al., 2017). In our canonical version of choice-matching, this input is a respondent's prediction of the relative frequency of each of the five ratings submitted by other respondents in the sample.

Let $x^r = (x_1^r, ..., x_5^r)$ denote the reported rating of respondent r, where $x_k^r = 1$ if r's answer is "k" and $x_k^r = 0$ otherwise. $y^r = (y_1^r, .., y_5^r)$, is then a vector of predictions, where y_k^r is the prediction by respondent r of the frequency of others choosing k as their MCQ answer.

The "prediction score" that respondent r receives is a function $S(\bar{x}^{-r}, y^r)$ of his prediction vector y^r and the distribution $\bar{x}^{-r}$ of the reported ratings' frequencies, excluding his own reported rating. The firm chooses S to be a proper scoring rule (see Savage, 1971, Gneiting and Raftery, 2007, and Definition 2 below). It is then in the interest of each respondent to report their expected value of the frequencies, which we call their **best prediction**. We denote by $\bar{S}^{-r}$ the average prediction score of respondents other than r who report the same rating as r, if any. Respondent r receives a score of zero if among the others' ratings there is any missing ratings option, not reported by at least one respondent other than r. If there is no missing option, he receives

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

with $\lambda \in (0, 1)$ an arbitrary weight chosen by the survey planner.

Why might we expect this payment formula to be truth-inducing? Observe, first, that respondent r can safely assume that the payment formula is in effect, because the presence or absence of a missing option does not hinge on his rating, and if the formula is not in effect he receives zero no matter what he does. This triggering mechanism was first employed by Baillon (2017), for the binary case.

The term λS provides incentives for giving best predictions because S is a strictly proper scoring rule. If a respondent believes that those who share his type also have (roughly) the same expectations of type frequencies, he will anticipate that they will also give (roughly) the same predictions. Then, if all respondents other than him provide honest ratings, the average of the prediction scores of all those with the same rating is (roughly) equal to the score corresponding to their best prediction. By providing his honest rating, a respondent is scored according to a prediction (roughly) identical to the one he made himself, ensuring that the second term provides the incentives to answer the MCQ honestly. These incentives apply even if some respondents do not provide a personal prediction. Thus, in a large survey most respondents could be asked just for their ratings (letting $\lambda = 0$ in their scoring formula).

For the above arguments, it is not crucial that respondents use what we call best predictions. What is important is that respondents of the same type believe that approximately the same predictions are optimal, and respondents of different types believe that different predictions are optimal, as we discuss below. Previous mechanisms have also invoked this assumption in some form, but with additional requirements or limitations (Prelec, 2004; Witkowski and Parkes, 2012; Radanovic and Faltings, 2013, 2014; Zhang and Cheng, 2014; Baillon, 2017). We will compare choice-matching to these proposals at the end of the paper.

3 Model and Results

3.1 Setup

Generalizing the previous example, let $A = \{1, \ldots, M\}$ denote the set of possible answers to an MCQ, and let N be the number of respondents. The respondent's **type** is denoted by random vector T^r, where $T_i^r = 1$ if $i \in A$ would be respondent r's honest answer to the question, and $T_i^r = 0$ otherwise.[2] $X^r = (X_1^r, ..., X_M^r)$ denotes the chosen answer or **reported type** of respondent r, where $X_k^r = 1$ if r's answer is "k" and $X_k^r = 0$ otherwise. The answer is **honest** iff $X^r = T^r$. For $i \in A$, $\bar{T}_i^{-r} = \frac{1}{N-1}\sum_{s' \neq r} T_i^{s'}$ is the frequency of types i in the sample, after excluding respondent r, and similarly, $\bar{X}_i^{-r} = \frac{1}{N-1}\sum_{s \neq r} X_i^s$ is the frequency of reported types i. Y_i^r is respondent r's prediction of $\bar{X}_i^{-r}$. The realizations of X_i^r and Y_i^r are denoted x_i^r, y_i^r, the realization of T_i^r is t_i^r, and so on.

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

Definition 1. The **choice-matching trigger** $\mathcal{M}^r$ of respondent r is the event such that each answer $i \in A$ is reported at least once by respondents other than r. That is, $\mathcal{M}^r$ occurs if and only if there is

[2] We will say in this case that r has type i.

no $i \in A$ with $\bar{x}_i^{-r} = 0$. The **type-matching trigger** $\mathcal{E}^r$ of respondent r is the event such that each type $i \in A$ is represented at least once among respondents other than r.

Note that when all respondents other than r respond honestly, the choice-matching trigger becomes identical to the type-matching trigger. Assumptions A1-A4 below are sufficient, but not necessary, for choice-matching to be honesty-inducing. In subsection 3.2 we explain how they can be relaxed.

Assumption 1. ***Common prior.*** *There exists a common prior on the distribution of* $T^1,...,T^N$.

Importantly, we do not assume that the survey planner knows the prior, only that it is known to the respondents.

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

Each respondent considers it possible that all distinct types are represented in the rest of the sample. Although a technical assumption, non-degeneracy cannot be relaxed without changing our approach substantially. The method thus puts some constraints on the survey design: the number of answer options must be less than the number of respondents (continuous scales are not allowed), and all options must have a chance of receiving an honest endorsement.

We now write r's expectation of others' type frequencies, conditioning on own type k and $\mathcal{E}^r$, as

$$p^{r,k} = E\left[\bar{T}^{-r} \mid T_k^r = 1, \mathcal{E}^r\right]$$

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

$$p^{r,k} \neq p^{s,i} \, \text{if } k \neq i$$

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

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

Our version of stochastic relevance differs slightly from the standard one (Miller et al., 2005), because expectations are also conditioned on the type-matching trigger $\mathcal{E}^r$ (this further implies $N \geq M+2$, since $N = M+1$ and $\mathcal{E}^r$ yield $p^{r,k} = (1/M, .., 1/M)$, for all r, k).

Stochastic relevance is a mild requirement, with substantial support in the experimental psychology literature (Marks and Miller, 1987). The finding there that respondents' answers about their personal characteristics are strongly correlated with their predictions about the distribution of answers in the sample was labeled the "false consensus effect" (Ross et al., 1977), although, as noted by Dawes (1989), it can be consistent with Bayesian updating.

Stochastic relevance may be violated if the distribution of a particular characteristic – for instance gender – is common knowledge. However, such a characteristic can be made stochastically relevant if combined with characteristics whose distribution is not common knowledge. For example, a respondent could be asked to indicate both gender and beverage preference, and to predict their joint distribution in the sample. If the correlation between beverage preference and gender is not known a priori, then one's own gender will be informative of the joint gender-preference distribution.

Impersonal updating is a more demanding assumption, stating that all respondents of the same type have identical expectations about type frequencies. It will not hold exactly in practice, as respondents with identical answers to an MCQ are likely to report differing predictions. As we discuss below, the main result of this section continues to hold when impersonal updating is only approximately true.

Assumptions A1-A4 generically hold in a setting in which respondents' types are independently and identically distributed conditional on a certain state of the world.[3] In our example in section 2 for instance, we could imagine that different states of the world correspond to the true quality of the trial product and that the better the quality, the higher the probability for each respondent to have a high honest rating.[4]

3.2 Inducing Honesty via Choice-Matching

We now assume A1-A4 and model the strategic setting induced by our payment rule as a Bayesian 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 $\sigma^r(t^r)$,

[3] An exact statement on the relation of the conditional i.i.d. assumption and our assumptions is given in the online appendix. One sufficient condition for our assumptions to hold then is that $N > M + 1$ and that the state of the world has a continuous distribution.

[4] Furthermore, our assumptions can hold even if the exact distribution of types is common knowledge. For example, with $N = 4$, $M = 2$ and common knowledge that there are two respondents of each type, each type knows that among the remaining three respondents there is one that matches his type and two who do not. A pair of respondents will therefore have the same expectations if and only if the pair are of the same type, satisfying assumptions A3 and A4.

and the profile excluding player r is denoted $\sigma^{-r}(t^{-r})$.

Given a real-valued payment rule $R(\sigma^r(t^r), \sigma^{-r}(t^{-r}))$, a set of response strategies is a (Bayesian) **Nash equilibrium**, if, for any response $(x,y) \neq (\sigma^r(t^r))$, 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** Nash equilibrium. A Nash equilibrium 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 $y \neq \sigma_y^r(t^r)$.

A strategy profile is **honest** if every respondent r reports $x^r = t^r$. A payment rule which has a Nash equilibrium that is strict in x and y and honest is called **strictly incentive compatible**. To make choice-matching incentive compatible, we use strictly proper scoring rules:

Definition 2. Let Z be a random vector of positive frequencies with dimension $M > 1$ which satisfies $\sum_{k=1}^{M} Z_k = 1$. Let y be a prediction of Z. We say that the scoring rule S is **strictly proper** if $E\left[S\left(y, Z\right)\right]$ is uniquely maximized for $y = E\left[Z\right]$.

Well-known strictly proper scoring rules are the quadratic $S\left(y, Z\right) = -\sum_{k=1}^{M}\left(y_k - Z_k\right)^2$, and the logarithmic $S\left(y, Z\right) = \sum_{k=1}^{M} Z_k \log\left(y_k\right)$.

Definition 3. Consider a respondent r who reports k as his type, and let S be strictly proper. We say that a collection of payment rules $R_{S,\lambda}^r$ induces **choice-matching** if

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

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

where $\lambda \in (0,1)$ and $\bar{S}^{-r}(x^r, x^{-r}, y^{-r})$ is the average prediction score achieved by the respondents other than r who submit $x^s = x^r$:

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

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

In words, if all M possible answers are chosen by at least one respondent other than r, choice-matching assigns him a score that is a weighted average of his own prediction score (that is, $S(y^r, \bar{x}^{-r})$) and the prediction score of those respondents who report the same type. Otherwise, he receives zero.

We now come to the main result of this section.

Proposition 1. *Under Assumptions A1-A4 any collection of payment rules $R^r_{S,\lambda}$ that induces choice-matching is strictly incentive compatible.*

Proof. We show that providing honest responses and best predictions is a Nash equilibrium that is honest and strict in x and y, hence strictly incentive compatible. Fix a respondent r with type k and suppose that all respondents other than r provide honest responses and best predictions. Because respondent r cannot influence the choice-matching trigger and receives zero if he is not matched, he will condition his payoffs on being choice-matched. Given that other respondents answer honestly, this means conditioning on the type-matching event $\mathcal{E}^r$. With his prediction y^r, r thus maximizes

$$E\left[S(y^r, \bar{T}^{-r}) \,|\, \mathcal{E}^r, T^r_k = 1\right]$$

Since S is strictly proper, r maximizes his expected payoff with $y^r = p^{r,k}$, i.e., what we call r's "best prediction" is the optimal prediction from r's perspective.[5] To see that choice-matching is strictly incentive compatible in x, consider the difference in the expected score for respondent r between reporting t^r honestly and deviating by making some dishonest report x^r with $x^r_i = 1$:

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

From non-degeneracy, $Pr(\mathcal{E}^r \mid T^r) \times (1-\lambda) > 0$. Furthermore, by impersonal updating all respondents of the same type have the same best prediction, so that the average $\bar{S}^{-r}$ is equal to the corresponding score S, and we have

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

[5] Due to non-degeneracy, $\tilde{E}[\cdot] := E\left[\cdot \,|\, \mathcal{E}^r, T^r_k = 1\right]$ is also an expectation operator, so that when S is strictly proper, for any random vector of frequencies Z, $\tilde{E}[S(r, Z)]$ is maximized by $r = \tilde{E}[Z]$.

By definition of $p^{r,k}$ and since S is strictly proper, we have $E\left[S(p^{r,k},\bar{x}^{-r}) - S(p^{r,i},\bar{x}^{-r}) \mid T^r = t^r, \mathcal{E}^r\right] > 0$ for any $i \neq k$. Thus, $R_{S,\lambda}$ is strictly incentive compatible. ■

3.3 Robustness of the Result

In the following, we discuss extensions of Proposition 1. First, there may be equilibria other than the honest one. However, if respondents cannot communicate with each other, it is plausible that respondents of the same type will adopt the same strategies (perhaps mixed). Under this **symmetry** assumption, we can show that all strict equilibria are either honest, or honest up to permutation of responses, that is, all respondents of type k report type i, all respondents of type i report type j, and so on. Second, we discuss in which way assumptions A1-A4 can be relaxed, while still allowing for a strict Nash equilibrium in honest strategies. Third, we point out that it is not necessary that all respondents report both an answer to the MCQ and a prediction, and that proper scoring can be replaced with other belief elicitation mechanisms.

A strategy profile σ is **symmetric** if respondents of the same type have identical strategies, i.e., if for all respondents r and s and all types t^r, t^s we have $\sigma^r(t^r) = \sigma^s(t^s)$ whenever $t^s = t^r$. The profile $\tilde{\sigma}$ is a **permutation profile** of σ if for any respondent r and any $k \in A$, we have $\tilde{\sigma}_{\pi(k)}\left(t^t\right) = \sigma_k(t^r)$, where π is a permutation of the answers to the MCQ.

Corollary 1. *Any permutation profile of honest responses is a strict Nash equilibrium. Moreover, when σ is a symmetric strict Nash equilibrium, then σ is a permutation profile of honest responses.*

Proof. The first part of Corollary 1 follows directly from the proof of Proposition 1, replacing k by $\pi(k)$. The second part rules out pooling equilibria as follows. Since σ is symmetric, respondents of the same type always provide the same answers. First, if respondents of different types report different answers, the reported types $\sigma_x(t)$ are a permutation of the (actual) type profile. As in the proof of Proposition 1, any respondent r of type k then maximizes his expected prediction score by setting $y^r_\ell = p^{r,k}_{\sigma_x^{-1}(\ell)}$ for all $\ell \in A$. Taken together, $\sigma(t)$ is a permutation profile of honest responses. Second, if respondents of two different types report the same type, there must be a missing answer option and thus all respondents receive zero. Since respondents could also receive zero by endorsing the missing answer option, the equilibrium cannot be strict. ■

Permutation profiles are a relabeling of the answers to the MCQ, and provide no benefit in score relative to the focal honest equilibrium. It is hard to see how such relabeling might be individually or collectively advantageous. Pooling equilibria may be attractive to respondents who would like to conceal their type. However, as the above proof shows, they are either not strict or not symmetric (or both).

Proposition 1 and Corollary 1 still hold when we weaken the impersonal updating assumption A4. Note first that A4 does not affect incentives for the predictions. If all respondents other than r are honest, r should report his best prediction $p^{r,k}$ when his type is k. It is intuitive that r will have strict incentives to truthfully report his type if the best prediction of all individuals of his type is "closer" to his best prediction than the best prediction of individuals of some other type. The precise definition of "closer" in the proof of such a result will depend on the scoring rule S. For instance, with the logarithmic scoring rule, assumption A4 can be replaced by the assumption that the relative entropy between the best predictions of two individuals with the same type is smaller than the relative entropy between the best predictions of two individuals with different types. As long as the expectations of respondents are clustered in this sense, truth-telling will remain strictly incentive-compatible.

Our method further allows that only a subsample of the respondents answer both the MCQ and the prediction question, while the remaining respondents answer the MCQ only. To accommodate this option theoretically, we redefine the matching trigger such that for each answer choice $i \in A$ there is a respondent $s \neq r$ who answers i *and* submits a prediction. Respondents who only answer the MCQ will receive the score $\bar{S}$, provided that the (redefined) event $\mathcal{M}^r$ occurs. This simplifies the mechanism for those respondents without losing incentive compatibility.

Finally, by using other types of scoring rules one can adjust choice-matching to respondents who are not risk-neutral or do not even maximize expected utility (Offerman et al, 2009). Whichever method is chosen, choice-matching adds minimal difficulty to these procedures, since it relies on the same scoring principle for rewarding answers to the MCQ as for rewarding predictions. A respondent who understands the elicitation method will thus easily understand choice-matching incentives as well.

4 Choice-Matching Generalized

Making predictions about other respondents' answers is an attractive default for our auxiliary task. However, there are situations in which its use cannot be recommended. 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, and the planner does not want to include another question in the survey to reinstate stochastic relevance. Second, some respondents may have trouble understanding the payments made according to proper scoring rules. Finally, in some situations respondents could expect that predictions of some other respondent type might be more accurate. For example, if a survey asks about occupation, a respondent could have an incentive to claim to be someone that has specialized knowledge of the empirical distribution, for instance a labor economist.

In this section, we show that there is a general principle behind choice-matching which can be employed by methods that do not rely on predictions.

4.1 General Mechanism

To formalize the underlying principle of choice-matching, 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 reported type x^r. Since respondents do not necessarily report a prediction, we refer to y^r in the following as r's y**-response**, which may take values in some general response set Ω. In the prediction-based model of Section 3, the utility functions correspond to the (realized) prediction score:

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

Note that in section 3 we implicitly assumed that the planner knows the respondents are risk-neutral. However, in general, choice-matching does not require such knowledge. What is needed is that the reward system used by the planner induces a game in which different strategies are optimal for different types.

Definition 4. Let G be a (Bayesian) game given by the collection of N respondents, a set of types 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) $\sigma_y^r(t^r) = y^{*k}$ for some $y^{*k} \in \Omega$ 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 σ in which respondents declare their types honestly (condition (i)), and in which respondents of identical types give the same y-response and respondents of different types give different y-responses (condition (ii)). Furthermore, this profile is a Nash 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, so that the game G alone is not sufficient to reveal types. For instance, in the setting discussed in section 3, it would not be sufficient to only score respondents on their own predictions, since then they would have no incentive to report their honest answer to the MCQ.

In section 3, conditions $(i)-(iii)$ correspond to separation in the prediction task, which – under assumptions A1-A4 – can be achieved by choosing a strictly proper scoring rule. The idea of the next proposition is that prediction scoring can be replaced by any game that satisfies conditions $(i)-(iii)$.

Proposition 2. *Let $G = \left\langle N, A, \Omega, \{u_k\}_{k\in A}, P\right\rangle$ be a type-separating game. Under assumptions A1-A2, any payment rule is strictly incentive compatible if it induces a game $\left\langle N, A, \Omega, \{V_k\}_{k\in A}, P\right\rangle$ in which on event $\mathcal{M}^r$:*

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

where $\lambda \in (0,1)$ and $\bar{u}$ is the average utility value achieved by the respondents other than r who submit $x_s = x_r$, i.e.,

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

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

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

Proof. The proof is given in the online appendix. It is almost identical to the special case in section 3. ■

4.2 Example and Applications

To see how choice-matching can be applied when the second question is not a prediction, consider again the example from section 2 in which each respondent gives a rating for a trial product. As an alternative to scoring respondents on predictions, the company could use the following procedure: first, the company presents a list of existing products to each respondent and each respondent chooses the product on the list he likes most. Again, we use the matching trigger: if there is a product that is not chosen by anyone, then r receives nothing, or some pre-defined, fixed participation fee. Otherwise, r participates in a "product lottery": he receives the product he chose with probability λ, and else, he receives the product chosen by a respondent randomly selected among those who give the same rating.

Intuitively, these incentives work if respondents expect that those who have the same assessment of the trial product will have (approximately) the same preference with regard to the products on the list, which is also what the type-separation assumption in proposition 2 entails.[6]

Experimental economics is another promising application domain for generalized choice-matching. Suppose that an experimenter is interested in the willingness-to-pay for a good that cannot be offered for sale, perhaps because its price exceeds the research budget or because it is hypothetical in nature. As the auxiliary task, the experimenter can then ask for the willingness-to-pay for a good which can be incentivized, and link the hypothetical willingness-to-pay to the stated willingness-to-pay demonstrated in the lab. Choice-matching also makes it possible to incentivize survey responses about behavior outside the lab, by linking the respondent's reward to the one obtained by respondents who give the same answers to the survey questions. For instance, in an experiment on decision-making under risk, respondents often choose among a number of gambles, with one of the gambles randomly chosen to be played out. Using choice-matching, the experimenter could ask a survey question related to risk attitude, e.g., which insurance a respondent has bought. She could then provide incentives for this survey question by paying respondents with some probablility according to the gambles chosen by respondents with the same survey answer.

[6] The type-separation assumption can be relaxed in a similar fashion as the impersonal updating assumption in subsection 3.1.

5 Comparison with Existing Methods

Incentives for non-verifiable MCQs based on predictions of the answer distribution were introduced by Prelec (2004), through the Bayesian Truth Serum (BTS) mechanism. Under BTS, a reported answer receives a high score if its actual frequency exceeds predicted frequency (and predictions are scored with the logarithmic version of S). The intuition here is that a respondent of, say, type k expects that other types will underestimate the frequency of k-types in the population, making an honest k-report optimal ex-ante. The exact BTS scoring formula is quite opaque and its theoretical guarantee holds only in the large sample limit (Prelec 2004, Cvitanić et al. 2017). However, the scores deliver a bonus property, in that they also reflect respondent expertise. Precisely, if types are i.i.d conditional on an underlying state of the world, then equilibrium BTS scores rank types according to their posteriors on the actual world state. In principle, one can back out the true world state by focusing on the highest-scoring respondents, and thus improve on majority voting and other crowd-wisdom algorithms (Prelec et al., 2017).

Finite sample incentive-compatibility is obtained by several recent mechanisms, typically under weaker assumptions than needed for BTS. For binary questions, these include the Robust Bayesian Truth Serum (RBTS) (Witkowski and Parkes, 2012) and Baillon's (2017) "Bayesian Market." The latter translates reported types and predictions into the buying and selling of securities whose value is defined by the distribution of reported answers. Some respondents may find the competitive market trading setting more engaging than scoring rules. For general MCQs, Radanovic and Faltings (2013) is mathematically simpler than both BTS and RBTS, but requires that the highest expected frequency of each type is held by the type itself, i.e, $p_k^{r,k} > p_k^{r,\ell}$, for each $k, \ell \in A$ with $k \neq \ell$.[7] The mechanism of Zhang and Chen (2014) involves a two-stage game where respondents revise their predictions after receiving the answer of another respondent. Survey implementation would therefore require contacting respondents more than once.

The (non-parametric) Divergence-based Bayesian Truth Serum (DIV) of Radanovic and Faltings (2014) is closest to (prediction-based) choice-matching. The basic idea is to penalize disagreement in predictions among respondents reporting the same type. A theoretical disadvantage of DIV relative to choice-matching, is that it allows non-honest symmetric strict equilibria that also payoff dominate

[7] This assumptions for instance rules out the common knowledge case of stochastic relevance mentioned in footnote 4.

honest ones. A practical disadvantage is that with DIV each respondent must provide a prediction.

Several mechanisms get by without asking for predictions. The “peer prediction” method of Miller et al. (2005) assumes that the planner knows the common prior, and is in effect able to compute the posteriors that ideal respondents would supply. More recently, data-intensive methods have attempted to estimate the prior from distributional assumptions or by machine learning (Radanovic and Faltings, 2015, Radanovic et al., 2016, Shnayder et al., 2016, Agarwal et al., 2017, Liu and Chen, 2017). These methods are suitable for settings where large quantities of data are collected on similar questions (e.g., different product ratings).

6 Conclusion

We have proposed a new way to elicit honest answers to a multiple choice question when honesty cannot be verified, and the planner functions as an outsider, agnostic about the distribution of answers in the population. The method asks respondents to engage in an incentivized auxiliary task, which in our canonical version is to predict the distribution of answers in the sample. By answering the MCQ, respondents pool with other respondents endorsing that same answer, and receive the same reward for the auxiliary task. It is not necessary for everyone to participate in the auxiliary task; the minimal number is just two plus the number of possible answers in the MCQ.

Compared to alternatives in the literature, our mechanism may be easier to explain. Previous empirical tests of the BTS (John et al., 2012 and Weaver and Prelec, 2013) used the so-called “intimidation method” (Frank et al. 2017), where respondents are told that they do not need to understand how the payment formula works, only that it is in their interest to be truthful. Although such black-box instructions do affect the aggregate answer distribution, it is not clear what fraction of respondents find the claim credible.

The specific prediction-based method may not be effective when there is strong public information about the distribution of types, or when respondents of the same type, as defined by the MCQ, have very different information about this distribution. The choice-matching principle can still be applied provided there is some task that induces separation among the honest answers. The planner only needs to find a task for which a separating equilibrium exists, she does not need to know the strategies that constitute this equilibrium.

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 there are no consequences for responding carelessly or dishonestly. These unverifiable variables may often be the key variables of social science interest. By linking stated and revealed preferences our method erases, in principle, the methodological boundary between those two types of data.

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 ï¿œikić, H. (2017) Incentive Compatible Surveys via Posterior Probabilities. Submitted.

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

[5] Dawes, R.M. (1989) Statistical criteria for establishing a truly false consensus effect. Journal of Experimental Social Psychology, 25(1), 1-17.

[6] 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.

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

[10] Marks, G., & Miller, N. (1987) Ten years of research on the false-consensus effect: An empirical and theoretical review. Psychological Bulletin, 102(1), 72.

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

[12] Offerman, T., Sonnemans, J., Van De Kuilen, G. and Wakker, P. (2009) A truth serum for non-bayesians: Correcting proper scoring rules for risk attitudes. The Review of Economic Studies 76(4), 1461-1489.

[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] Ross, L., Greene, D., and House, P. (1977) The "false consensus effect": An egocentric bias in social perception and attribution processes. Journal of experimental social psychology, 13(3), 279-301.

[21] Savage, L. J. (1971) Elicitation of personal probabilities and expectations. Journal of the American Statistical Association, 66(336), 783-801.

[22] 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.

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

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

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

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

[27] 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.

Appendix

A Stochastic Relevance Under Conditional Independence and Identical Distribution

Here we show that our stochastic relevance assumption A3 is satisfied in the common setup in which there is an unknown state of the world conditioning on which respondent types are independently and identically distributed.[8]

We first derive conditions which guarantee stochastic relevance without a matching trigger. These conditions are always satisfied when the state of the world has a continuous distribution, and is generically satisfied when it has a discrete distribution. In the second step, we show that our matching trigger adds only minor additional requirements.

A.1 Stochastic relevance with respect to the state of the world

As before, let $M > 1$ be an integer and $A = \{1, ..., M\}$. We further write Δ^M for the M-dimensional simplex. We start by considering a single respondent, whose type we denote T. Suppose that Ω is a random vector taking values in Δ^M, representing the state of the world, and write

$$P\left(T = k \,|\, \Omega = \omega\right) = \omega_k \tag{A.1}$$

for all $k \in A$. In the following, we assume that Ω is a continuous random variable or a discrete random variable, and we denote by $P(\omega)$ its probability density/mass function. In the appendix, we slightly abuse notation by writing $T = k$ instead of $T_k^r = 1$ to mean that respondent r's type is k.

Definition 5. Stochastic relevance w.r.t. to Ω holds if for all different $k, j \in A$:

$$E\left(\Omega | T = k\right) \neq E\left(\Omega | T = j\right) \tag{A.2}$$

Our first result states:

[8] Note that assumption A4 is then also satisfied.

Proposition 3. *Stochastic relevance w.r.t. to* Ω *holds if and only if for all different* $k, j \in A$ *and any constant* $\lambda \in \mathbb{R}$*:*

$$P(\Omega_k = \lambda\Omega_j) < 1 \tag{A.3}$$

In words, stochastic relevance w.r.t. Ω holds if the ratio of any two state components Ω_k, Ω_j is not known with certainty ax ante. This condition is satisfied if Ω follows a continuous distribution on Δ^M because the subset of Δ^M on which $\Omega_k = \lambda\Omega_j$ has zero measure. If the distribution of Ω is discrete, then A.3 is violated only if for all ω s.t. $P(\omega) > 0$, $\omega_k = \lambda\omega_j$.

Proof. We first rewrite (A.2). W.l.o.g. take $k = 1$, $j = 2$. For any $i \in A$:

$$E(\Omega_i \,|\, T = 1) = \int\limits_{\Delta^M} P(\omega \,|\, T = 1)\,\omega_i\, d\omega = \int\limits_{\Delta^M} \omega_i \frac{P(T = 1 \,|\, \omega)}{P(T = 1)} P(\omega)\, d\omega \tag{A.4}$$

$$= \int\limits_{\Delta^M} \frac{\omega_i\omega_1}{\int_{\Delta^M} P(\omega')\,\omega'_1 d\omega'} P(\omega)\, d\omega = \frac{E(\Omega_i\Omega_1)}{E(\Omega_1)} \tag{A.5}$$

And analogously, $E(\Omega_i \,|\, T = 2) = \frac{E(\Omega_i\Omega_2)}{E(\Omega_2)}$. Thus, stochastic relevance w.r.t. Ω holds if and only if for any $i \in A$:

$$\frac{E(\Omega_i\Omega_1)}{E(\Omega_1)} \neq \frac{E(\Omega_i\Omega_2)}{E(\Omega_2)} \tag{A.6}$$

We next show that (A.3) implies (A.6) by contradiction. Suppose that (A.3) holds but that stochastic relevance w.r.t. Ω is violated, i.e. for any $i \in A$:

$$\frac{E[\Omega_i\Omega_1]}{E[\Omega_1]} \quad = \quad \frac{E[\Omega_i\Omega_2]}{E[\Omega_2]} \tag{A.7}$$

In particular, setting $i = 1$ and $i = 2$, respectively, (A.7) implies:

$$E\left[(\Omega_1)^2\right] = E[\Omega_1\Omega_2]\frac{E[\Omega_1]}{E[\Omega_2]}, \qquad E\left[(\Omega_2)^2\right] = E[\Omega_1\Omega_2]\frac{E[\Omega_2]}{E[\Omega_1]} \tag{A.8}$$

which yields

$$E\left[(\Omega_1)^2\right]E\left[(\Omega_2)^2\right] = \left(E\left[\Omega_1\Omega_2\right]\right)^2 \tag{A.9}$$

Note that (A.9) is an instance of the Cauchy-Schwarz inequality which holds with equality if and only if there is $\lambda \in \mathbb{R}$ such that

$$P(\Omega_1 = \lambda \Omega_2) = 1 \tag{A.10}$$

Thus, we have a contradiction and (A.3) indeed implies (A.6).

Finally, it is easy to see that (A.6) implies (A.3). Note that $P(\Omega_k = \lambda \Omega_j) = 1$ implies $\lambda = \frac{E(\Omega_2)}{E(\Omega_1)}$. This immediately yields equation (A.7), in violation to (A.6). ■

Corollary 2. *Stochastic relevance w.r.t. to Ω holds if and only if for all different $k, j \in A$ there is a set $S \subset \Delta^M$ of possible values of Ω with*

$$P(\Omega \in S \,|\, T = k) \neq P(\Omega \in S \,|\, T = j) \tag{A.11}$$

In words, stochastic relevance w.r.t. to Ω holds if and only if every type disagrees with all other types about the probabilities of some states of the world.

Proof. Consider the sets

$$S_1 = \left\{\omega : \frac{\omega_k}{E[\omega_k]} > \frac{\omega_j}{E[\omega_j]}\right\}, \qquad S_2 = \left\{\omega : \frac{\omega_k}{E[\omega_k]} < \frac{\omega_j}{E[\omega_j]}\right\} \tag{A.12}$$

and recall that $P(\omega \,|\, T = k) = \frac{\omega_k}{E[\omega_k]} P(\omega)$. Thus, for $i = 1, 2$ we have

$$P(\Omega \in S_i \,|\, T = k) = \int\limits_{S_i} \frac{\omega_k}{E[\omega_k]} P(\omega)\, d\omega$$

and hence $P(\Omega \in S_i \,|\, T = k) = P(\Omega \in S_i \,|\, T = j)$ only if $P(\Omega \in S_i) = 0$. From Proposition 3, we know that stochastic relevance holds if and only if for all different $k, j \in A$ and any constant $\lambda \in \mathbb{R}$, (A.3) holds. It is straightforward to verify that (A.3) is equivalent to either $P(\Omega \in S_1) > 0$ or $P(\Omega \in S_2) > 0$ (or both). It follows that stochastic relevance holds if and only if $P(\Omega \in S_1 \,|\, T = k) \neq P(\Omega \in S_1 \,|\, T = j)$ or $P(\Omega \in S_2 \,|\, T = k) \neq P(\Omega \in S_2 \,|\, T = j)$. ■

A.2 Stochastic Relevance in a Finite Sample Distribution

Consider a finite sample of N respondents with types $T^1,...T^N$ which satisfy conditional independence w.r.t. Ω, that is, for all $s \neq r$ and $i,k \in A$:

$$P\left(T^s = i\,|\,\Omega = \omega, T^r = k\right) = P\left(T^s = i\,|\,\Omega = \omega\right) \tag{A.13}$$

and are identically distributed, so that for all $s,r \leq N$:[9]

$$P\left(T^s = i\,|\,\Omega = \omega\right) = P\left(T^r = i\,|\,\Omega = \omega\right) = \omega_i \tag{A.14}$$

Next, recall that we defined $\bar{T}_i^{-r}$ as the frequency of respondents whose type is i in the sample excluding r.

Proposition 4. *Under conditional i.i.d, stochastic relevance w.r.t.* Ω *implies stochastic relevance w.r.t.* $\bar{T}^{-r}$.

Proof. W.l.o.g let $r = 1$. Then, for all $i,k \in A$:

$$E\left[\bar{T}_i^{-1}\,|\,T^1 = k\right] = \frac{1}{N-1}\sum_{s=2}^{N} Pr\left[T^s = i\,|\,T^1 = k\right] = Pr\left[T^2 = i\,|\,T^1 = k\right] \tag{A.15}$$

$$= \int\limits_{\Delta^M} P\left(T^2 = i\,|\,T^1 = k, \Omega = \omega\right) P\left(\Omega = \omega\,|\,T^1 = k\right) d\omega = E\left[\Omega_i\,|\,T^1 = k\right] \tag{A.16}$$

Thus, for given $j,k \in A$, we get that $E\left[\bar{T}^{-1}\,|\,T^1 = k\right] \neq E\left[\bar{T}^{-1}\,|\,T^1 = j\right]$ if $E\left[\Omega\,|\,T^1 = k\right] \neq E\left[\Omega\,|\,T^1 = j\right]$. ■

A.3 Sufficient Conditions For Assumption A3

We will next state a result which adapts propositions 3 and 4 to our matching trigger. As before, we write:

$$p^{r,k} := E\left[\bar{T}^{-r}\,|\,\mathcal{E}^r, T^r = k\right] \tag{A.17}$$

[9] Note that it is not restrictive that we identify the state ω with the probability ω_i. Suppose for instance that in the trial product example from section 2, honest ratings are distributed conditionally i.i.d on the quality of the product, represented by Ω. We can then simply define a new random vector $\tilde{\omega}$, s.t. $P\left(T^r = i\,|\,\Omega = \omega\right) = \tilde{\omega}_i$.

Recall that our assumption A3 states that if $j \neq k$, then:

$$p^{r,k} \neq p^{r,j}. \tag{A.18}$$

It will further prove useful to use the following identities, where $\Omega > 0$ means that $\Omega_k > 0$ for all $k \in A$.

Lemma 1. *If the types* $T^1, ..., T^N$ *are conditionally i.i.d w.r.t.* Ω *and* $P(\Omega > 0) > 0$, then for $s \neq r$ and any $i \in A$:

$$(i)\, P(T^s = i | \Omega = \omega, T^r = k, \mathcal{E}^r) = P(T^s = i | \Omega = \omega, \mathcal{E}^r) = \frac{1}{N-1} + \frac{N-1-M}{N-1}\omega_i \tag{A.19}$$

$$(ii)\, E(\bar{T}_i^{-r} | T^r = k, \mathcal{E}^r) = \frac{1}{N-1} + \frac{N-1-M}{N-1} E(\Omega_i | T^r = k, \mathcal{E}^r) \tag{A.20}$$

The first identity allows us to conclude that if $T^1, ..., T^N$ are conditionally i.i.d w.r.t. Ω, then $T^1, ..., T^N$ are also conditionally i.i.d w.r.t. to both Ω and the matching trigger. The second identity allows us to relate stochastic relevance w.r.t. Ω to stochastic relevance w.r.t. $\bar{T}^{-r}$.

Proof. In the following, w.l.o.g fix $r = k = 1$. We first prove identity *(i)*. For $s \neq 1$:

$$P(T^s = i | \Omega = \omega, T^1 = 1, \mathcal{E}^1) = \frac{P(T^s = i, \mathcal{E}^1 | \Omega = \omega, T^1 = 1)}{P(\mathcal{E}^1 | \Omega = \omega, T^1 = 1)} \tag{A.21}$$

Let θ^{-1} be the set of all possible $t^{-1} = (t^2, ..., t^N)$ and let $\theta_{\mathcal{E}}^{-1}$ be the set of all t^{-1} such that 1's matching trigger is in effect. We then get for the numerator:

$$P(T^s = i, \mathcal{E}^1 | \Omega = \omega, T^1 = 1) = \sum_{t^{-1} \in \theta_{\mathcal{E}}^{-1}} P(T^s = i, T^{-1} = t^{-1} | \Omega = \omega, T^1 = 1) \tag{A.22}$$

$$= \sum_{\substack{t^{-1} \in \theta_{\mathcal{E}}^{-1}, \\ s.t. t^s = i}} P(\, T^{-1} = t^{-1} | \Omega = \omega, T^1 = k) = \sum_{\substack{t^{-1} \in \theta_{\mathcal{E}}^{-1}, \\ s.t. t^s = i}} P(T^{-1} = t^{-1} | \Omega = \omega) \tag{A.23}$$

where the first step uses that each part of the sum is zero when $t^s \neq k$ and the second step uses conditional independence. Similarly, we get $P\left(\mathcal{E}^1|\Omega=\omega, T^1=1\right)=P\left(\mathcal{E}^1|\Omega=\omega\right)$ for the denominator. Thus:

$$P(T^s=i|\Omega=\omega, T^1=1, \mathcal{E}^1)=\frac{P(T^s=k, \mathcal{E}^1|\Omega=\omega)}{P\left(\mathcal{E}^1|\Omega=\omega\right)}=P(T^s=i|\Omega=\omega, \mathcal{E}^1) \tag{A.24}$$

which is the first part of identity (i). To get the second part, note that due to conditional i.i.d. w.r.t. Ω we can rewrite

$$P(T^{-1}=t|\Omega=\omega)=\prod_{s=2}^{N}\omega_{t^s} \tag{A.25}$$

so that (A.23) yields the same expression for all $s\neq 1$. Thus, we are allowed to write

$$P(T^s=i|\Omega=\omega, \mathcal{E}^1)=\frac{1}{N-1}\sum_{s'=2}^{N}P(T^{s'}=i|\Omega=\omega, \mathcal{E}^1) \tag{A.26}$$

$$=E(\bar{T}_i^{-1}|\Omega=\omega, \mathcal{E}^1)=\frac{1}{N-1}E((N-1)\,\bar{T}_i^{-1}|\Omega=\omega, \mathcal{E}^1) \tag{A.27}$$

Let $Z=(N-1)\,\bar{T}^{-1}-1_M$, where 1_M denotes an M-dimensional vector of 1s. Note that conditioning on $\Omega=\omega$ and $\mathcal{E}^1$, Z is distributed according to a multinomial distribution, with $N-M-1$ draws and parameters $\omega_1, ..., \omega_M$, such that:

$$(N-1)\;E(\bar{T}_i^{-1}|\Omega=\omega, \mathcal{E}^1)=1+E(Z_i\,|\Omega=\omega,\, Z\geq 0\,) \tag{A.28}$$

$$=1+(N-M-1)\,\omega_i \tag{A.29}$$

which also shows the second part of identity (i).

To get to identity (ii), observe that for all $i\in A$

$$E(\bar{T}_i^{-1}|T^1=k, \mathcal{E}^1)=\int_{\Delta^M}P(\,\omega|T^1=k, \mathcal{E}^1)\,E(\bar{T}_i^{-1}|T^1=k,\,\Omega=\omega,\,\mathcal{E}^1\,)d\omega \tag{A.30}$$

$$= \int_{\Delta^M} P(\,\omega | T^1 = k, \mathcal{E}^1)\, \frac{1}{N-1} \sum_{s=2}^{N} P(T^s = i\, | T^1 = k, \Omega = \omega, \mathcal{E}^1\,) d\omega \tag{A.31}$$

$$= \int_{\Delta^M} P(\,\omega | T^1 = k, \mathcal{E}^1)\, \left[\sum_{s=2}^{N} \frac{1}{(N-1)^2} + \frac{N-M-1}{(N-1)^2} \omega_i \right] d\omega \tag{A.32}$$

$$= \frac{1}{N-1} + \frac{N-M-1}{N-1} E(\omega_i | T^1 = k, \mathcal{E}^1) \tag{A.33}$$

where in (A.32) we used identity (*i*). This proves identity (*ii*) as well. ■

We will now come to the key result:

Proposition 5. *Let $N-1 > M$ and suppose that*

(i) The types $T^1, ..., T^N$ are conditionally i.i.d w.r.t. Ω,

(ii) $P(\Omega > 0) > 0$

(iii) and for all different $k, j \in A$ and any constant $\lambda \in \mathbb{R}$:

$$P(\Omega_k = \lambda \Omega_j \,|\, \Omega > 0) < 1 \tag{A.34}$$

Then, assumption A3 holds.

Proof. Again fix $r = 1$. Invoking Lemma 1, we have

$$E(\bar{T}_i^{-1} | T^1 = k, \mathcal{E}^1) = \frac{1}{N-1} + \frac{N-1-M}{N-1} E(\Omega_i | T^1 = k, \mathcal{E}^1) \tag{A.35}$$

Thus, under assumptions *(i)* and *(ii)*, A3 holds if and only if $E(\Omega | T^1 = k, \mathcal{E}^1) \neq E(\Omega | T^1 = j, \mathcal{E}^1)$ for different j, k.

Note that $P(\Omega > 0) > 0$ implies $P(\Omega > 0 \,|\, T^r = i) > 0$ for all $i \in A$. Under conditional i.i.d., this implies that for any $i \in A$:

$$P\left(\mathcal{E}^1 \,|\, T^1 = i\right) > P\left(T^2 = 1, ..., T^{M+1} = M \,|\, T^1 = i\right) \tag{A.36}$$

$$= \int\limits_{\Delta^M} P\left(\omega \,|\, T^1 = i\right) \prod_{k=1}^{M} \omega_k > 0 \tag{A.37}$$

Thus, we can define $\tilde{P}(\cdot) := P\left(\cdot\,|\,\mathcal{E}^1\right)$ and $\tilde{E}(\cdot) := E\left(\cdot\,|\,\mathcal{E}^1\right)$. Lemma 1 ensures that if $T^1,...,T^N$ are independent conditional on Ω, they are independent conditional on Ω when replacing P and E by $\tilde{P}$ and $\tilde{E}$. We can thus apply proposition 3 to $\tilde{P}(\cdot)$ and $\tilde{E}(\cdot)$, which yields that assumption A3 holds if and only if for all $\lambda \in \mathbb{R}$

$$\tilde{P}(\Omega_k = \lambda\Omega_j) = P\left(\Omega_k = \lambda\Omega_j\,|\,\mathcal{E}^1\right) < 1 \tag{A.38}$$

Finally, note that

$$P\left(\Omega_k = \lambda\Omega_j\,|\,\mathcal{E}^1\right) = 1 - \int\limits_{\Delta^M} P\left(\omega\,|\,\mathcal{E}^1\right) I_{\left\{\omega_k \neq \lambda\omega_j\right\}} d\omega \tag{A.39}$$

$$P(\Omega_k = \lambda\Omega_j\,|\,\Omega > 0) = 1 - \int\limits_{\Delta^M} P(\omega\,|\,\Omega > 0)\, I_{\left\{\omega_k \neq \lambda\omega_j\right\}} d\omega \tag{A.40}$$

and that

$$P\left(\omega\,|\,\mathcal{E}^1\right) = P\left(\omega\,|\,\mathcal{E}^1, \Omega > 0\right) = \frac{P\left(\mathcal{E}^1\,|\,\omega, \Omega > 0\right) P(\omega\,|\,\Omega > 0)}{P(\mathcal{E}^1\,|\,\Omega > 0)} \tag{A.41}$$

For $\omega > 0$, we have that $P\left(\mathcal{E}^1\,|\,\omega, \Omega > 0\right) > 0$, so that (A.41) is strictly larger than 0 if and only if $P(\omega\,|\,\Omega > 0) > 0$. Therefore, when (A.40) does not equal 1, (A.39) does not equal 1 either. Thus, assumption *(iii)* implies (A.38) and under *(i)* and *(ii)*, assumption *(iii)* is sufficient to guarantee A3.

■

Note that in addition to the requirements of propositions 3 and 4, proposition 5 adds only the additional requirements that the sample is sufficiently large ($N - 1 > M$) and that the linear independence needed for proposition 3 holds also when restricted to $\Omega > 0$. Again, given that $N - 1 > M$, it is sufficient that Ω has a continuous distribution on Δ^M. Thus, the only cases that are excluded by our matching trigger are knife edge cases in which Ω's distribution on Δ^M is discrete, and in which, given that all types have a strictly positive (conditional) probability, the ratio of at least two types is known. Since this peculiar situation is unlikely to occur in practice, the additional restriction implied by our trigger is a very minor one.

B Proof of Proposition 2

Recall the statement of Proposition 2:

Proposition. *Let $G = \left\langle N, A, \Omega, \{u_k\}_{k\in A}, P\right\rangle$ be a type-separating game. Under assumptions A1-A2, any payment rule is strictly incentive compatible if it induces a game $\left\langle N, A, \Omega, \{V_k\}_{k\in A}, P\right\rangle$ in which on event $\mathcal{M}^r$:*

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

where $\lambda \in (0,1)$ and $\bar{u}$ is the average utility value achieved by the respondents other than r who submit $x_s = x_r$, i.e.,

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

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

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

Proof. As in the proof of Proposition 1, observe that r cannot influence the choice-matching trigger and should thus condition his expected payoffs on being choice-matched. Since G is type separating, condition (*iii*) from definition 4 guarantees that the payment rule is strictly incentive compatible in y. Consider then the difference in the expected score for respondent r between reporting t^r honestly and deviating by making some dishonest report x^r with $x^r_i = 1$:

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

Due to the 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 $i \neq k$. Again, using condition (*iii*) from definition 4, we have that for r's conditional expectiation of this payoff difference:

$$E\left[u_k\left(y^{*k}, x^{-r}, y^{-r}\right) - u_k\left(y^{*i}, x^{-r}, y^{-r}\right) \mid T^r = k, \mathcal{E}^r\right] > 0$$

which gives the required result. ■