Eliciting and estimating valid subjective probabilities: An experimental investigation of the Exchangeability Method

Simone Cerroni[a,*], Sandra Notaro[a], W. Douglass Shaw[b,c]

[a] Department of Economics, University of Trento, Via Inama 5, 38122, Trento, Italy

[b] Department of Agricultural Economics, Texas A&M University, College Station, TX 77843-2124, United States

[c] Hazards Reduction and Recovery Center, Texas A&M University, College Station, TX 77843-3137, United States

[*]Corresponding Author. E-mail address: simone.cerroni@unitn.it. Phone: +39 02 52036963. Fax: +39 02 52036946

Abstract

Using a laboratory experiment, we investigate whether incentive compatibility affects subjective probabilities elicited via the Exchangeability Method, an elicitation technique consisting of several chained questions. We hypothesize that subjects who are aware of the chaining strategically behave and provide invalid subjective probabilities, while subjects who are not aware of the chaining state their real beliefs and provide valid subjective probabilities. The validity of subjective probabilities is investigated using de Finetti's notion of coherence, under which probability estimates are valid if and only if they obey all axioms of probability theory.

Four experimental treatments are designed and implemented. Subjects are divided in two initial treatment groups: in the first, they are provided with real monetary incentives, and in the second, subjects are not. Each group is further sub-divided in two treatment groups, in the first, the chained structure of the experimental design is made clear to the subjects, while, in the second, the chained structure is hidden by randomizing the elicitation questions.

Our results suggest that subjects provided with monetary incentives and randomized questions provide valid subjective probabilities because they are not aware of the chaining which undermines the incentive compatibility of the Exchangeability Method.

Keywords: Subjective probabilities; Exchangeability; Validity; Pesticide residue; Apples

JEL: C44; D81; I10

Highlights

> Incentive compatibility determines the validity of elicited beliefs.

> Randomized questions make the chaining less transparent.

> Randomized questions and monetary incentives increase validity rates.

Eliciting and estimating valid subjective probabilities: An experimental investigation of the Exchangeability Method

1. Introduction

During the last two decades, many social scientists have become more interested in investigating and eliciting subjective probabilities of everyday events. The main reason to pursue this line of inquiry is because many choices in the real world involve future outcomes and take place under uncertainty, hence, people often behave and make decisions according to their beliefs and expectations. Manski (2004) demonstrates the importance of subjective probabilities in several branches of applied economics, ranging from the influence of households' probabilistic income expectations on their consumption and saving decisions, to the impact of students' probabilistic expectations of the returns (again, in income terms) to education on schooling choices.

Expectations on risky and uncertain outcomes, which lie outside of the financial domain, are potentially complex, but also important to deal with. These have been neglected in economics until quite recently, perhaps, because they pertain to issues which are more difficult to address than financial risk and uncertainty, such as stock market activity. Early work on subjective probability pertained to another issue that is relatively simple to understand and for which outcomes are readily observable with short delays: the weather, specifically, temperature and precipitation forecasts (e.g., Brier, 1950; Baillon, 2008).

A domain where subjective probabilities have been recognized to be crucial in understanding and predict people's choice behavior is food safety, but little in this area has been done to explore subjective probability elicitation. Despite many studies have shown how consumers' probabilistic expectations of food safety might affect purchases (e.g., Buzby et al., 1998; Williams and Hammit, 2001), they often use very simple and

rough methods for eliciting subjective probabilities, which often consist in directly asking subjects a guess of the probability that given outcomes will occur in the future. The key problem, with issues such as food safety, is that the nature of the uncertainty is less accessible to laypeople, and the primary outcome, the health effect, may be unobservable for quite some time to come[1]. However, a recent study suggests that uncertainty in food safety decisions may be quite important (Kivi and Shogren, 2010).

In this paper, we investigate and elicit consumers' perceptions of the probability that given levels of pesticide residues will be present in apples produced in the future in the Province of Trento (Italy). Pesticide residues pose health risks to people who eat apples, and, thus, people's perceptions of their presence can affect their preferences for agricultural policies that local authorities are planning to incentivize the production of healthy apples. The investigation of this topic might be very important to this region as apple production is a key sector of its economy (P.A.T., 2010). Generally, the presence of pesticides in food is quite important, as we all must eat; several studies shown that human exposures to chemicals are associated with risks to human health, they may even produce very severe illnesses as cancer (Alavanja et al., 2004).

There are many different ways to elicit subjective probabilities and several are briefly discussed below. We use an innovative technique for eliciting probabilities, known as the Exchangeability Method (EM), recently used by Baillon (2008). He elicited subjective probabilities for future daily temperature in Paris, the euro/dollar exchange rate, and the daily variation of the French stock index CAC 40. His experimental subjects were asked to estimate these for a given day about four weeks after the experiment was conducted. The same technique was further developed by

[1] Short-term food sickness is perhaps observable after a short delay, but ethics in experiments preclude subjecting subjects to this.

Abdellaoui et al. (2011) to elicit subjective probabilities and investigate ambiguity attitudes related to similar topics[2].

The EM consists of a set of binary questions where subjects are asked to bet a certain amount of money on a given outcome rather than on an alternative outcome. In each question, the outcomes which are presented to the subject result from a bisection procedure of the whole state space of the random variable under study. When subjects become indifferent between the two outcomes, they are assumed to perceive both as equally likely and subjective probabilities can be estimated. The sequential splitting process behind the EM makes this elicitation procedure chained, in the sense that the outcomes presented in each question depends on the outcome that has been chosen in the previous one.

The incentive compatibility of the EM might be questioned because previous experimental studies have shown that chained elicitation mechanisms are not necessarily incentive compatible. In fact, the provision of monetary incentives to subjects, based on their choice behavior during the experiment, might induce them to not state their real beliefs, but, instead, to strategically behave to be better rewarded upon completion of the tasks for the experiment (e.g., Harrison, 1986).

In this paper, we investigate whether the lack of incentive compatibility of the EM due to both the presence of chained questions and no provision of real monetary incentives, affects the validity of subjective probabilities elicited by such technique. We determine and measure the validity of subjective probabilities elicited via the EM implementing a method based on de Finetti's notion of *coherence* (1937). By using this

[2] They elicited subjective probabilities related to the daily variation of the French stock index CAC 40, temperature in Paris and also in a randomly drawn remote country for a given day about 3 months after the experiment.

approach we essentially aim to identify the best way for eliciting subjective probabilities via the EM, in terms of validity[3].

The remainder of the paper is laid out as follows. We first highlight the main strengths and limitations of the EM by comparing it to other techniques for eliciting beliefs. Next, describe our testable hypotheses and the methodology used to measure validity of subjective probabilities. Finally, we offer some conclusions based on the experimental results we have obtained.

2. Methods for eliciting subjective probabilities

The simplest way to elicit subjective probabilities consists of asking people to directly state the chance that a specific magnitude of the outcome will happen in the future (Spetzler and Stael Von Holstein, 1975). Asking simple, direct questions is common in a host of previous health-risk studies, such as those involving smoking cigarettes (e.g., Viscusi, 1990; Gerking and Khaddaria, 2011), drinking contaminated water (e.g., Jakus et al. 2009; Shaw et al., 2012), or eating unhealthy food (e.g., Buzby et al., 1998; Williams and Hammit, 2001).

However, unless subjects are asked to state a chance for each of all possible specific magnitudes of outcomes, the information gathered from such an easy question is very limited. Using a direct approach like this, we might learn about only one point, or about a very narrow range, in the individual's subjective probability distribution.

The reliability of subjective probabilities elicited via this family of techniques, called *direct methods*, have also been often questioned, particularly by psychologists, on the grounds that laypeople may be neither familiar with the notion of probability per se,

[3] Since this experiment is conducted in the lab, with a controlled environment and real monetary incentives, we only refer to the internal validity of elicited risk estimates. Hence, we cannot analyze the external validity of our results, being aware that elicited estimates in the lab might be different from those elicited in the field, where it is impossible to control for many confounding factors (for instance, background risk) (Harrison et al., 2007).

nor willing to put efforts into thinking in probabilistic terms (Mansky, 2004)[4]. Some have gone as far to suggest that individuals do better in understanding risks with verbal, rather than numerical percentage or probability scales (see discussion in Weinstein and Diefenbach, 1997). Several economic studies provide supporting evidence that people have problems with open-ended questions about probability estimates (e.g., Jakus et al., 2009; Riddel and Shaw, 2006)

Other approaches, called *indirect methods*, may overcome some of the limitations that direct methods have. Here, probability measures are indirectly estimated at the points for which subjects show their indifference between choices involving lotteries or gambles. Indirect techniques have often been used for eliciting probabilities related to financial outcomes (e.g., Andersen et al., 2010; Offerman et al., 2009) because actual monetary payments for played-out bets make the elicitation mechanism incentive compatible and appear to be relatively easy for subjects to understand. Quite recently, a few scholars have used indirect methods to estimate subjective probabilities related to health and environmental outcomes (e.g., Fiore et al., 2009; Cerroni and Shaw, 2012). As noted in the introduction, the limited use of these *indirect methods*, for eliciting probabilities related to health and environmental outcomes, is due to the fact that very long term health and environmental outcomes cannot be played out at the end of experiments in the lab setting, thus again making incentive compatibility a potential issue. Fiore et al. (2009) and Cerroni and Shaw (2012) both rely on hypothetical portrayals of adverse forest impacts, and, in the former study, the authors explore the use of virtual forest fires in the experimental setting.

The most popular of the indirect methods are called "*external reference events*", in which subjects are asked to choose between a lottery characterized by an uncertain

[4] Many studies investigated different approaches for communicating probabilities to laypeople and, then, eliciting their best estimate (e.g., Gigerenzer and Hoffrage, 1995; Hammit and Graham, 1999; Corso et al., 2001).

event (*U*), whose probability needs to be estimated, and a lottery characterized by an external reference event (*K*), whose probability is known and disclosed to subjects. The probability of the known event (K) is often visually presented through probability wheels, scroll bars, or other visual aids such as risk ladders, grids, or pie charts, all of which have been tested as probability communication devices (e.g., Morgan and Henrion, 1990). Once subjects become indifferent between the two lotteries, the uncertain outcome (*U*) is assumed to have the same probability of occurrence of the familiar outcome (*K*), so that $P(U) = P(K)$ (Spetzler and Stael Von Holstein, 1975).

Although these techniques are widely used, they may produce biased probability estimates because they ask subjects to process two sources of uncertainty at the same time: the first relates to the uncertain outcome (*U*), the second relates to the external reference event (K). Previous experimental studies have shown that individual choices depend on the source of uncertainty that subjects have been asked to consider[5] (e.g., Kilka and Weber, 2001; Abdellaoui et al., 2011), and, hence, elicitation mechanisms, which combine diverse sources of uncertainty, may become too complex and generate biased subjective probabilities (Baillon, 2008).

Source dependence does not appear to be an issue within another class of indirect methods, which use *internal events*. In these elicitation techniques, subjects deal with magnitudes of the outcomes, but not with their probabilities of occurrence. In fact, subjects are only asked to bet a certain amount of money on one of the several disjoint subspaces, in which the whole state space of the variable under study has been previously divided. When subjects become indifferent regarding betting on one disjoint subspace rather than on the others, subjects are assumed to perceive those subspaces as equally likely (Spetzler and Stael Von Holstein, 1975).

[5] Baillon (2008, p.77) defined a source of uncertainty as "…a set of events that are generated by a common mechanism of uncertainty".

The EM, which was first described by Raiffa (1968) and more recently implemented by Baillon (2008) and Abdellaoui et al. (2011), belongs to this latter class of probability elicitation techniques. In the specific case of the EM, each question gives subjects the chance to bet on one of two disjoint subspaces, as the whole state space of the random variable under study is sequentially divided using a bisection process. The subdividing procedure of the event space makes each binary question of the EM chained to the previous one. In fact, the sub-events that subjects face in each question depend on the sub-event that has been chosen in the precedent question.

As noted in the introduction, chained techniques for eliciting preferences or beliefs are perhaps not incentive compatible. Strategic behaviors might have strong impacts on elicited subjective probabilities (Harrison, 1986) and chained questions may propagate subjects' strategic choices made during the choice-tasks (e.g., Spetzler and Stael Von Holstein, 1975; Wakker and Deneffe, 1996). Previous investigations that rely on chained games and real monetary incentives have validated their results by using subjects' statements of unawareness about the presence of chaining in the games (Van de Kuilen et al., 2006; Abdellaoui et al., 2011).

Baillon (2008) dealt with this problem by randomizing the order of questions. The questions are not sequentially presented and, thus, the chaining is less transparent to subjects because they are no longer aware of the relationship between the disjoint subspaces they face in one question and the subspace they have chosen in the previous one. Developing this experimental design with randomized questions, one hopes that telling the truth becomes the simplest and most efficient strategy that subjects can use when they play the EM (Baillon, 2008).

The effect of real monetary incentives on the elicitation of subjective probabilities has been investigated in another recent application of the EM, by Abdellaoui et al. (2011). After having tested that subjects were unaware of the chained structure of the

EM, they next compare subjective probabilities provided by two groups of subjects, one provided with monetary incentives and the other not. They did not find any substantial difference between subjective probabilities elicited from the two groups, but do not provide a logical explanation as to why subjects provided with money incentives should have greater or lower beliefs than others.[6]

In contrast, we argue that monetary incentive may affect the validity of subjective probabilities elicited via the EM depending on whether subjects are aware of the chaining or not. In particular, we believe that monetary incentives and the ordering of questions may affect the incentive compatibility of the Exchangeability Method, and therefore, the validity of subjective probabilities elicited by using this technique. Here, we don't want to confuse truth with validity, in fact, as reported below, incentive compatibility and validity are separate and distinct concepts. An elicitation mechanism is incentive compatible if subjects have an incentive to state their real beliefs (Vossler and Evans, 2009), while subjective probabilities are valid if and only if they obey all axioms and theorems of probability theory (de Finetti, 1937).

In this paper, we hypothesize that subjective probabilities elicited via incentive incompatible mechanisms, which induce subjects to not truly state their beliefs, are likely to be invalid, in the sense that they do not obey to axioms and theory of probability theory.

To test our predictions, we create a validation method based on the de Finetti's notion of coherent probability measures (1937; 1974a; 1974b), under which, subjective probabilities are coherent if and only if they obey all axioms and theorems of probability theory. The choice of using the de Finetti's notion of coherence to define

[6] They found that probability distribution functions of median temperature in Paris in a given day for both groups are quite well calibrated with historical distribution of temperature in that particular day. In contrast, they found that probability distribution functions of median daily variation of the French stock index CAC 40 in a given day for both groups differ from historical distribution of CAC 40 daily variation in that particular day.

valid subjective probabilities relies on the fact that the EM is based on the assumption of exchangeability-based probabilistic sophistication (Chew and Sagi, 2006), that, in turn, is based on the idea of equal likelihoods of exchangeable events (de Finetti, 1937)

3. Specific Objectives

Previous applications of the Exchangeability Method have not directly investigated the effect of chaining on subject's choice-behaviors (see Baillon, 2008; Abdellaoui et al., 2011), but they have simply tried to avoid the use of identifiable chained questions in their experimental designs. As noted before, this is due to the fact that previous experimental studies have shown how the provision of chained questions along with real monetary incentives make the elicitation mechanism incentive incompatible (Harrison, 1986).

In line with the above discussion, we hypothesize that subjective probabilities elicited via an incentive incompatible mechanism more likely turn out to be invalid. In particular, we hypothesize that subjective probabilities elicited via the EM, using sequential questions along with real monetary incentives, are invalid because, when the chaining is clear to subjects, monetary incentives will encourage them to strategically behave. In contrast, when random questions are provided in the EM along with monetary incentives, subjective probabilities are valid because when the chaining is less transparent to subjects, monetary incentives induce them to state their real beliefs, or at least, to invest more cognitive effort into the elicitation process[7].

We also hypothesize that subjects provided with random questions will perform better than those provided with sequential questions in terms of validity, even in the absence of actual monetary rewards at the end of the experiment. We expect those who are not aware of the chaining will provide invalid subjective probabilities, but less so

[7] We thank an anonymous reviewer for suggesting this possibility.

than those who are aware of the chaining structure. Our prediction is supported by the fact that questions related to the elicitation of the third quartile ask subjects to choose between two prospects that they have already ruled out in previous questions. The issue is evident to subjects when questions are sequentially ordered, while it is less transparent when they are randomly ordered. This may affect the validity of probabilities elicited using sequential questions as subjects may perceive questions to be meaningless and may invest less cognitive effort in playing the game.

To test our hypotheses, we first need to understand whether elicited subjective probabilities are valid or not. The empirical way we tested the validity of subjective probability elicited via the EM is described below.

4. The Experimental Design

4.1. The empirical application

Our specific application consists of investigating uncertain outcomes related to fire blight, a bacterial disease that has threatened apple orchards in the Province of Trento, at least since 2003 (EMF, 2006). This phytopathology damages and kills apple plants resulting in substantial losses in the production of apples. The best available science predicts a future spread of the disease in apple orchards of the Province of Trento, since suitable climatic conditions for the biology of the bacterium *Erwinia amylovora* are likely to occur in the future (unpublished results by Edmund Mach Foundation).

Although Italian farmers currently control the fire blight by using pesticides, chemicals might be not efficient enough to prevent the future spread of this apple disease. Nevertheless, the future production of apples in the Province of Trento (around 420.000 tons at the present time) might not decrease if farmers start implementing new

adaptation strategies against fire blight. However, the only strategy that is currently available to farmers is the introduction of new active principles, such as the antibiotic streptomycin, that is currently forbidden by the Italian legislation, but that has been already used in U.S., Germany, Belgium and Netherlands for controlling the fire blight (Németh, 2004).

In the context presented here, we focus on three diverse random variables: the percentage (or number) of days in which the infestation will occur during the blossoming period in 2030 (g)[8], the number of apples containing at least one residue in a sample of 100 apples in 2030 (a)[9], and the number of apples containing more than 1 residue in a sample of 100 apples in 2030 (r)[10]. These variables have been selected among many other possible measures of pest infestation, or apple contamination, after having interviewed approximately 20 focus group subjects.

4.2. The Exchangeability Method and the related game

Let a random variable under study in the EM game (EG) be g. The EG uses a series of binary questions to reveal an individual's underlying cumulative distribution function (CDF) over an event x that is drawn from an event space, $S_G = G_1^1$. The first step of the EG establishes the lower and upper bounds of the event space, defined as g_0 and g_1. Each subject is asked the bounds for outcomes outside of which they are essentially certain the outcome cannot happen at all — i.e., the bounds that pertain to a non-zero probability of an outcome.

The second step of the EG involves asking a series of questions that establish the value of $g_{1/2} \in S_G$ that corresponds with the 50th percentile of the subjective CDF, in

[8] The blossoming period usually occurs in April in Trentino.
[9] This is the number of apples at least one residue beyond the level of 0 mg/kg.
[10] This is the number of apples containing at least two residues beyond the level of 0 mg/kg.

other words, the median estimate. This series of questions asks the subject to choose between binary prospects. In the first binary question, S_G is divided at a point g_a into two prospects, say $G_a=\{g_0<x<g_a\}$ and $G_a'=\{g_a\leq x<g_1\}$, where $g_a=\{g_0 + [(g_1-g_0)/2]\}$. If G_a was chosen by the individual, the implication is that the individual believes the probability of occurrence of the sub-event G_a is equal to that of the sub-event G_a', so that $P(G_a)\geq P(G_a')$ and $g_a\geq g_{1/2}$. A follow-up binary question is then asked of this same individual, using a new value g_b and two new prospects G_b and G_b'. If G_a was chosen in the first question, then $g_a<g_b$. However, if G_a' was chosen in the first question, then $g_a>g_b$. This process is repeated until the individual reaches a value g_z such that she is indifferent between G_z and G_z'. When this point is reached, it follows that $g_z=g_{1/2}$, $G_z=G_2^1$, $G_z'=G_2^2$, and $P(G_z)=P(G_z')$. This process describes the "chaining" or interdependence of these binary outcome questions.

A similar process can be followed to determine other points for the individual's subjective CDF; in theory as many as the researcher wants to identify. However, there is a limit to how many separate points can be elicited because of potential exhaustion of the subject. For example, to determine the value of $g_{1/4}\in S_G$ that corresponds with the 25th percentile, a gamble is proposed that is contingent on a value of x that is lower than $g_{1/2}$, obtained in the previous step. Once again, a sequence of values, g_a, g_b, …, g_z is used, but in this next case (the quartile) the initial upper bound is $g_{1/2}$. In the first new binary question, subjects choose between the following binary prospects, $G_a=\{g_0<x<g_a\}$ and $G_a'=\{k_1\leq x<g_{1/2}\}$. As above, this process is repeated until the individual is indifferent between G_z and G_z', so that $g_z=g_{1/4}$ $G_z=G_4^1$, $G_z'=G_4^2$, and $P(G_z)=P(G_z')$ (see Figure 1 and Appendix A). At the end of the EG, the second binary question that subjects have already answered is presented again to them in order to test the consistency of their choice behaviors.

4.3. Other games

The Repeated Exchangeability Game (REG) consists in eliciting a new measure of the median value of individual CDFs, say $g_{1/2}$', through a second round of Exchangeability Game. This round differs from the first one because the lower and upper bounds of the event space are now not defined by g_0 and g_1, but instead by the subjective estimates of the quartiles $g_{1/4}$ and $g_{3/4}$ elicited via the Exchangeability Game (see Example 2 in Appendix A).

The Certainty Equivalent Game (CEG) is based on the notion of certainty equivalents (CE), defined as the sure amount of money that makes subjects indifferent to gamble. For the CEG, the subjects are presented with two choice tasks, say CT1 and CT2, both containing six binary questions. In each question of the first choice task (CT1), the subject is asked to choose between a lottery (Lottery 1), in which he or she wins a monetary outcome *x* if the real outcome G_j^i will happen in the future (or a null monetary outcome otherwise), and a sure payment *z,* varying from 0 to 100€. In the same way, in the CT2, subjects are asked to choose between a lottery (Lottery 2), in which they win a monetary outcome *x* if the real outcome G_j^k will happen in the future (or a null monetary outcome otherwise), and a sure payment *z* varying from 0 to 100€. Hence, each subject is presented with two choice tasks characterized by six binary matching question where he or she has to choose between options A (bet *x* € on the occurrence of G_j^i in CT1 or G_j^k in CT2) and B (take the amount of money z = 0, 25, 49, 51, 75, and 100€) (see Example 3 in Appendix A). The certainty equivalent for the lottery described in option A is determined by looking at the first question of the choice task in which the subject switches from choosing option A to choose option B. Recall that G_j^i and G_j^k are the couple of sub-spaces that have been already judged to be

equally likely by the subjects themselves, during the earlier Exchangeability Game. Each subject in our study was presented with this game three times for each variable of interest in the study. In the first, the two lotteries involved in the game are denoted as G_2^1 and G_2^2, in the second, they are G_4^1 and G_4^2, and in the third, they are G_4^3 and G_4^4 [11].

4.4. The sample

The sample of laboratory subjects consists of 80 individuals who were randomly recruited outside the main supermarkets of Trento and asked to come in the experimental lab of the University of Trento for a compensation of 25€ (show-up fee). Given the fact that we recruit non-students and, then, we bring them in the lab, we can define our study as an artefactual field experiment (Harrison and List, 2004). Our sample consists of people between 18 and 70 years age who live in the Province of Trento and the sample is balanced regarding the gender. They are not strictly speaking, a simple random sample of the population, because they were recruited outside food markets, but as most people visit such markets to obtain food, they probably are quite representative of people leaving in this Province. Moreover, the random nature of the sample may be biased by subjects' motivation to participate in the experiment. For example, subjects may participate because they were interested in the topic or because they were in need of the show-up fee.

4.5. Treatments

Selected participants were randomly assigned to four subsamples or treatment groups, where each treatment is characterized by a different experimental design: "real

[11] Both games have been already used to test exchangeability in other experimental applications (e.g., Baillon, 2008; Abdellaoui et al., 2011).

incentives-random questions" (22 subjects)[12], "real incentives-sequential questions" (23 subjects), "hypothetical incentives-random questions" (19 subjects), and "hypothetical incentives-sequential questions" (16 subjects). For the "hypothetical incentives" treatments, subjects are only given a show-up fee, while in the "real incentives" treatments, subjects are told that one randomly selected individual from each group has the chance to win additional 100€ based on her/his choices during the experiment. Specifically, one subject is to be randomly selected at the end of the experiment and one of the questions she/he answers during the experiment is also randomly selected to be played out. The lucky subject is selected through the draw of a numbered chip from a bingo cage (Cage 1). The total number of chips is equal to the total number of participants in each session, so that each subject has an equal chance of being selected. The question with the potential pay-out is also selected through the draw of a numbered chip from another bingo cage (Cage 2), that contains as many numbered chips as the number of questions that the subject answered during the experiment. The drawn participant wins the additional 100€ if and only if the event she/he had chosen in the drawn question contains the value of the random variable under consideration that the best science currently predicts. This prediction is based on the research conducted by the Edmund Mach Foundation (EMF). This procedure for the determination of a "win" in the lottery situation is similar to that used by Fiore et al. (2009) in their virtual experiment on the risk of wild fires. Despite some participants already being aware of the existence of the EMF, all subjects are provided with general information about the EMF's research that provides science-based estimate of probabilities. Note that even when all subjects receive the same information, it is a common finding that they may

[12] The original "real incentives-random questions" treatment had 23 subjects, however we deleted observations gathered from one particular subject who declared that she has made a mistake during the tasks. Given that subjects did not have the chance to correct their errors during the experiment and chained experimental designs propagate mistakes, our subjects were asked to declare if they unintentionally made errors answering experimental questions.

not form the same subjective estimates (e.g. Riddel and Shaw, 2006; Shaw et al., 2012). In all treatments subjects were provided with precise information about the values that the random variables under study had in the last ten years (from 2000 to 2010) and, then, they were asked to play the games.

In the "sequential questions" treatments subjects are asked to answer questions that allow us to elicit the percentiles of their CDFs in the following order: $g_{1/2}$, $g_{1/4}$, $g_{3/4}$, $a_{1/2}$, $a_{1/4}$, $a_{3/4}$, $r_{1/2}$, $r_{1/4}$, and $r_{3/4}$. In the "random questions" treatments, this chained structure of the game is hidden through a mixed up order of questions determined once and for all. In fact, we elicit the percentiles of subjects' CDFs in the following order: $g_{1/2}$, $a_{1/2}$, $r_{1/2}$, $g_{1/4}$, $a_{1/4}$, $r_{1/4}$, $g_{3/4}$, $a_{3/4}$, and $r_{3/4}$.

It follows that each subject, regardless of the treatment group to which she/he is randomly assigned, plays the Exchangeability and the other games three times, one for each random variable under study.

5. Hypotheses about the validity of subjective probabilities

To investigate the effect of sequential (or random questions) and real (or hypothetical) monetary incentives on the validity of subjective probabilities elicited via the Exchangeability Game, we first need to understand whether gathered estimates are valid or not. Given the theoretical background of the EG, we argue that subjective probabilities elicited via this technique are valid if the only if the exchangeability assumption is satisfied, otherwise they are invalid. In fact, under the exchangeability assumption, subjective probabilities elicited via the EG satisfy all definitions, axioms and theorems of probability theory. Considering two disjoint sub-events, G_j^i and G_j^k, the exchangeability assumption is satisfied when the two sub-events are exchangeable, in the sense that the probability related to the occurrence of one must be equal to the

probability of occurrence of the other (see Appendix B). When the assumption holds we fail to reject the following null hypothesis (H_0) and we consider elicited subjective probabilities valid:

H_0: $P\left(G_j^i\right) = P\left(G_j^k\right), \forall k \neq i, k \leq n$

H_1: $P\left(G_j^i\right) \neq P\left(G_j^k\right), k \neq i, k \leq n$

We test this hypothesis, and, thus, the validity of subjective probabilities elicited via the EM by investigating whether subjects' choice behaviors are consistent across the Exchangeability Game, the Repeated Exchangeability Game, and Certainty Equivalent Game. In particular, we test the following two hypotheses:

Hypothesis 1. We test whether the exchangeability assumption is satisfied or not by comparing the estimates of $g_{1/2}$ obtained from the Exchangeability Game and the estimates of $g_{1/2}$' obtained from Repeated Exchangeability Game. The exchangeability assumption is satisfied, and, thus, the subjective probability of the event $g_{1/2}$ is valid, if and only if we fail to reject the following null hypothesis (H_0):

H_0: $g_{1/2} = g_{1/2}$'

H_1: $g_{1/2} \neq g_{1/2}$'

Hypothesis 2. We test whether the exchangeability assumption is satisfied or not by comparing the certainty equivalents that subjects are willing to accept to give up the possibility to play the lotteries presented in the matched pairs of choice tasks, $\left[L\left(x : G_j^i\right)\right]$ in CT1 and $\left[L\left(x : G_j^k\right)\right]$ in CT2 (Certainty Equivalent Game). The exchangeability assumption is satisfied, and, thus, the subjective probability of the event presented in

both CT1 and CT2 is valid, if and only if we fail to reject the following null hypothesis (H_0):

$$H_0: CE\left[L\left(x : G_j^i\right)\right] = CE\left[L\left(x : G_j^k\right)\right], \text{ with } k \neq i,\ k \leq j$$

$$H_1: CE\left[L\left(x : G_j^i\right)\right] \neq CE\left[L\left(x : G_j^k\right)\right]$$

6. Testing hypotheses

Before testing the hypotheses above, we first check the consistency of subjects' choice behaviors by examining their answers to the repeated binary questions presented at the end of the Exchangeability Game. In the 66.51% of cases, subjects' choices are the same in the original and repeated questions. This result is quite encouraging, given that Baillon (2008) found a consistency rate of 70.51% applying the same procedure to evaluate consistency, but investigating random variables more familiar to subjects than the ones we have examined here. Further, the McNemar test shows that subjects' choices are consistent even across treatments (Table 1).

Next, testing our hypotheses at sample level, we determine whether subjects, belonging to diverse experimental treatments, provide valid subjective probabilities or not. This allows us to test predictions presented above, in particular, the fact that subjects provided with real monetary incentives and random questions state valid subjective probabilities, while the others do not. Recall that subjects are assumed to provide valid subjective probabilities if the exchangeability assumption holds, and thus, if and only if we fail to reject the null hypotheses presented in *Hypotheses 1* and *2*.

We test *Hypotheses 1* and *2* by using non-parametric tests such as the Wilcoxon Matched-Pairs Signed-Ranks test (WMP) and the Sign Test of Matched Pairs (SMP).

The SMP test is used because the assumptions behind the WMP test were not always satisfied in our sample. For example, the differences between the matched values provided by each subject were not always distributed symmetrically around the median point in our sub-samples (*symmetry assumption*).

While testing *Hypothesis 1*, we only investigates the validity of individual CDFs' medians ($g_{1/2}$, $a_{1/2}$, and $r_{1/2}$), as we rely on estimates elicited via the Exchangeability Game and Repeated Exchangeability Game, testing *Hypothesis 2*, we also examine the validity of individual CDFs' first and third quartiles ($g_{1/2}$, $a_{1/2}$, $r_{1/2}$, $g_{1/4}$, $a_{1/4}$, $r_{1/4}$, $g_{3/4}$, $a_{3/4}$, and $r_{3/4}$), as we rely on estimates elicited via the Exchangeability Game and Certainty Equivalent Game.

Further, we assess the *validity rate* (V) for each different experimental treatment, which is the percentage of valid subjective probabilities out of the total number of elicited estimates in each treatment. This rate allows us to quantitatively assess the validity of subjective probabilities for each treatment and test, once more, predictions presented in Paragraph 3. To compute validity rates, we first need to verify whether each observation ($g_{1/2}$, $a_{1/2}$, $r_{1/2}$, $g_{1/4}$, $a_{1/4}$, $r_{1/4}$, $g_{3/4}$, $a_{3/4}$, and $r_{3/4}$) provided by each subject (i = 1,...,80) is valid or not. For example, let's consider one specific experimental subject, who provides us with the estimate of $g_{1/2}$, we assume that this estimate is valid if and only if the certainty equivalents for Lottery 1 and 2, presented in the Certainty Equivalent Game, are equal, thus, $CE\left[L\left(x:G_2^1\right)\right]=CE\left[L\left(x:G_2^2\right)\right]$. This does not imply any statistical test, but just a simple check of the equality between $CE\left[L\left(x:G_2^1\right)\right]$ and $CE\left[L\left(x:G_2^2\right)\right]$.

In addition, by examining the dissimilarity between $CE\left[L\left(x:G_2^1\right)\right]$ and $CE\left[L\left(x:G_2^2\right)\right]$, we can also investigate how much elicited subjective probabilities are invalid. For each elicited probability, the dissimilarity is measured as the absolute value

of the difference between the certainty equivalents for Lottery 1 and Lottery 2, that is given by $\Delta(CE) = \left| CE\left[L\left(x: G_2^1\right)\right] - CE\left[L\left(x: G_2^2\right)\right] \right|$.

Based on these absolute values, we create an invalidity scale consisting of five categorical level of invalidity: very low invalidity when $\Delta(CE) \in \{x < 27\}$, low invalidity when $\Delta(CE) \in \{27 \leq x < 52\}$, medium invalidity when $\Delta(CE) \in \{52 \leq x < 77\}$, high invalidity when $\Delta(CE) \in \{77 \leq x < 101\}$, and, finally, very high invalidity when $\Delta(CE) \in \{x \geq 101\}$. These boundaries have been chosen as the absolute values, Δ(CE), are naturally grouped in five categories given the range of the sure amount of money x that subjects might accept instead of playing the lotteries presented in the Certainty Equivalent Game.

Using this classification, we calculate the percentage of invalid probability estimates that falls within each category of invalidity and, hence, we investigate how far off invalid probabilities are from being valid.

Finally, we hypothesize that, not only the features of the experimental setting may determine the validity of subjects' subjective probabilities, but also their socio-economic conditions. We econometrically test this hypothesis by estimating a model in which the discrete dependent variable captures the validity of each observation provided by each subject, while independent variables captures the characteristics of each experimental setting and other socio-economic variables which characterize subjects, allowing for some observable heterogeneity.

7. Results

7.1. Non-parametric tests

By testing *Hypothesis 1* for each experimental group of subjects, we identify the effect of our experimental designs on subjects' capability to provide valid estimates of

the median values. In the "real incentives-sequential questions" treatment we have 24 matched pairs of observations, in the "real incentives-random questions" 40, in the "hypothetical incentives-sequential questions" 22, and in the "hypothetical incentives-random questions" 26 (Table 2).

The validity of individual CDFs' medians ($g_{1/2}$, $a_{1/2}$, and $r_{1/2}$) is determined by testing *Hypothesis 1* via both the Wilcoxon Matched-Pairs Signed-Ranks (WMP) and the Sign Test of Matched Pairs tests (SMP). Median estimates are assumed to be valid if and only if we fail to reject the null hypothesis characterizing this test. The WMP test' results suggest that "real incentives-random questions" and "hypothetical incentives-random questions" treatments provide valid estimates, while "real incentives-sequential questions" and "hypothetical incentives-sequential question" treatments do not. The SMP test almost produces the same results, except for the fact that also "hypothetical incentives-sequential question" treatment provides valid estimates (Table 3). The discrepancy between WMP and SMP's results about the "hypothetical incentives-sequential question" treatment suggests that the interpretation of these results is problematic, and thus, we conclude that only "real incentives-random questions" and "hypothetical incentives-random questions" treatments provide valid subjective estimates.

Testing *Hypothesis 2* for each experimental group of subjects, allows us to investigate whether subjects, belonging to diverse experimental treatments, provide valid estimates of the median, first quartile, and third quartile values of individual CDFs or not. In the "real incentives-sequential questions" treatment we have 143 matched pairs of observations, in the "real incentives-random questions" 167, in the "hypothetical incentives-sequential questions" 136, and in the "hypothetical incentives-random questions" 115 (Table 4). Again, the validity of median, first quartile, and third quartile estimates of individual CDFs ($g_{1/2}$, $a_{1/2}$, $r_{1/2}$, $g_{1/4}$, $a_{1/4}$, $r_{1/4}$, $g_{3/4}$, $a_{3/4}$, and $r_{3/4}$) is

determined by testing *Hypothesis 2* via both the WMP and the SMP tests. Estimates are assumed to be valid if and only we fail to reject the null hypothesis characterizing this test. The WMP test's results show that the "real incentives-sequential questions" treatment and the "hypothetical incentives-random questions" treatments do not provide valid estimates, while the "real incentives-random questions" and the "hypothetical incentives-sequential questions" treatments do. However, the validity of WMP test's results about the "hypothetical incentives-sequential question" treatment may be compromised because all assumptions behind the test are not completely satisfied. As the SMP test's results suggest that also the "hypothetical incentives-sequential questions" treatment does not provide valid estimates, we conclude that the "real incentives-random questions" is the only treatment providing valid estimates (Table 5). Considering the whole set of subjective estimates, and not just median estimates, our results support the hypothesis, under which only subjects provided with real monetary incentives along with random questions return valid subjective probabilities. This result demonstrates that when the chaining structure of the elicitation mechanism is not perceived by experimental subjects, monetary incentives increase the chance of eliciting valid subjective estimates. However, it does not prove that validity depends on whether subjects perceive the Exchangeability Game to be incentive compatible or not.

Above, we predicted that subjects who perceived the EM to not be incentive compatible strategically play the game and provide invalid subjective probabilities. Our prediction is supported if the percentage of rewarded subjects is higher in experimental treatments where real incentives are associated with sequential questions rather than with random questions. This is due to the fact that subjects, who face sequential questions, perceive the chaining, and, thus, strategically play (or, at least try to) the incentive incompatible elicitation mechanism to get better rewarded at the end of the experiment. We test this hypothesis by taking into account the subjects, who belong to

real incentive treatments, and simulating the rewards that each subject should have gained if she/he was the randomly drawn subjects and the third question he/she answered was the randomly drawn question[13].

We found that the chance of being rewarded is 50 percent for subjects playing sequential questions and 34.78 percent for those playing random questions (Table 6). This finding supports the hypothesis that subjects who are aware of the chaining play the elicitation mechanism to get better rewarded, but, unfortunately, also provide invalid subjective probabilities.

7.2. The validity rate

For each treatment, we calculate the validity rate (V) which is simply the percentage of valid estimates within each treatment. According to the previous findings, we found that "real incentives-random questions" treatment provides the highest validity rate (39.13%), then the "hypothetical incentives-random questions" (29.86%), "real incentives-sequential questions" (26.26%), and "hypothetical incentives-sequential questions" (22.22) follow (Table 7).

Again, according to our predictions, we found that subjective probabilities, elicited providing real monetary incentives and using random questions, are likely more valid than those elicited providing real monetary incentives and using sequential questions. As demonstrated above, the low validity rate we found for the "real incentives-sequential questions" treatment depends on the fact that the provision of sequential questions along with monetary incentives makes the overall incentive incompatibility of the Exchangeability Game clear.

[13] We have chosen the third question because it was the randomly drawn question at the end of the "real incentives-sequential questions" session.

Even when monetary incentives are not provided to subjects, we found that random questions perform better than sequential questions in terms of validity. This result may be due to the fact that, in the part of the Exchangeability Game related to the elicitation of the third quartile estimates, subjects are asked to choose between prospects that they have already ruled out in the elicitation of the first and second quartile estimates. For example, a subject who has expressed median and first quartile estimates, respectively equal to $g_{1/2} = 72$ and $g_{1/4} = 68$, by answering the first and second set of binary questions, is then asked to express the third quartile estimate $g_{3/4}$. She does so by answering a third set of binary questions which involve outcomes greater than 72, and, thus, in conflict with outcomes she has just chosen in previous questions.

While this is clear to subjects who belong to the "hypothetical incentives-sequential questions" treatment, this is not clear to the subjects who belong to the "hypothetical incentives-random questions" treatment. Thus, chaining may induce subjects to reduce the effort invested in the tasks, as they may believe that questions related to the elicitation of the third quartile are somewhat meaningless. Our hypothesis here is supported by the fact that validity rate of third quartile estimates in the "hypothetical incentives-sequential questions" treatment (about 22 percent) is lower than that founded in the "hypothetical incentives-random questions" treatment (almost 30, see Table 7).

Our prediction is also confirmed by the fact that while, in the "hypothetical incentives-sequential questions" treatment, the validity rate of third quartile estimates (almost 18 percent) is lower than that of first quartile (almost 23 percent), in the "hypothetical incentives-random questions" treatment, the validity rate of third quartile estimates (about 27 percent) is greater than that of first quartile (25 percent – again, see Table 7). The issue of meaningless sequential questions does not arise when monetary

incentives are provided because subjects are assumed to put more mental effort into trying to earn as much monetary reward as they can.

Unfortunately, we found relatively low validity rates for all our treatments. However, we do not believe this is due to the elicitation mechanism per se, but rather, to a series of different issues that we discuss below. First, such low validity rates may be due to the particular uncertain outcomes we investigated. As many subjects were, at least, initially unlikely to be familiar with the pesticide risk issue addressed in the experiment, the validity of elicited probabilities may be undermined by the sense of insecurity that subjects have likely felt during the tasks (Frisch and Baron, 1988). In contrast, something simple and familiar to all, such as uncertainty about temperature, might yield higher validity.

An alternative potential reason, as to why our subjects' responses have such low validity rates, involves the test we have used to investigate the validity of elicited probabilities. Recall that to calculate the validity rate, we assume that each estimate is valid if and only if the certainty equivalent for Lottery 1 was equal to that for Lottery 2. This procedure seems to be quite constraining as it does not imply any statistical test, but is just a simple check of the equality. Unfortunately, here, we cannot either measure or disentangle the effect of such influencing factors, but only speculate on them.

Given the large proportion of invalid probabilities, we also investigate their level of invalidity. Using the invalidity scale described above, we found that about 31 percent of the invalid probability measures are characterized by a very low level of invalidity, about18 percent by a low level, approximately 12 percent by a medium level, about 8 percent by a high level, and about 31 percent by a very high level (Table 8 and Figure 2). The fact that invalid observations are concentrated at the two extreme levels of invalidity emphasizes that subjects were either rather sophisticated about their

probability estimates or not at all, with a smaller portion of the subjects falling in-between.

7.3. The econometric analysis

In this paper, we hypothesize that, not only experimental designs, but also socio-economics characteristics of subjects and their degree of familiarity with the problem influence individual performances in terms of validity. This hypothesis is econometrically tested by estimating a discrete model in which the dependent variable *VALID* represents the validity of each estimate provided by each subject. The dependent variable takes the value 1 if and only if the estimate is valid according to Hypothesis 2, and thus $CE\left[L\left(x : G_j^i\right)\right] = CE\left[L\left(x : G_j^k\right)\right]$, with $k \neq i$, $k \leq j$[14]. Given that each subject *i* provides 9 estimates ($g_{1/2}$, $a_{1/2}$, $r_{1/2}$, $g_{1/4}$, $a_{1/4}$, $r_{1/4}$, $g_{3/4}$, $a_{3/4}$, and $r_{3/4}$), we should have a panel data of 720 observations. However, we have 142 missing values for the dependent variable *VALID* because the Certainty Equivalent Game investigating the validity of each estimate was not always displayed to subjects during the experiment depending on their choice behavior.

In our model (Equation 1), the probability that each individual estimate is valid, depends on a set of explanatory variables available from survey-type questions given in the laboratory: the experimental treatment that subjects belong to, the socio-economics status of subjects themselves, and subjects' degree of interest in the issue of food safety (see Table 9 for details about the explanatory variables).

$$VALID_i = \beta_0 + \beta_1 T_i + \beta_2 RV_i + \beta_3 P_i + \beta_4 S_i + \beta_5 I_i + \beta_6 TR_i \qquad \text{(Equation 1)}$$

[14] Our dependent variable relies only on Hypothesis 2, but not on Hypothesis 1, because, while the latter only test the validity of median estimates, the former takes into account also first and third quartiles.

We estimate this model by using the generalized linear model estimation with and without robust standard errors. Hereafter, we focus on the estimation with robust standard errors that allows for clustering effects.

Our first aim is to test again whether the probability of providing valid estimates depends on the provision of monetary incentives and the ordering of questions. The set of variables *T* consists of four dummies (*TRS*, *TRR*, *THS*, and *THR*) which take the value 1 if and only if the subjects belong to the experimental treatment that the variable represents. We observe that only subjects who, belong to the "real incentives-random questions" treatment (*TRR*), have a statistically significant higher probability of providing valid estimates than those, who belong to the "hypothetical incentives-sequential questions" treatment (*THS*), which is used as baseline (Table 10). This result supports our previous findings from non-parametric testing and validity rate's analysis.

Other two sets of dummy variables have been included in our model, the first, *RV*, to capture whether the probability of providing valid estimates depends on the variable that subjects have to consider in playing the Exchangeability Game (*G*, *A*, or *R*), the second, *P*, to capture whether the validity of stated estimates is statistically different among median ($g_{1/2}$, $a_{1/2}$, and $r_{1/2}$), first quartile ($g_{1/4}$, $a_{1/4}$, and $r_{1/4}$), and third quartile estimates ($g_{3/4}$, $a_{3/4}$, and $r_{3/4}$). However, we found no statistical difference in terms of validity between estimates related to diverse variables and diverse percentiles (Table 10).

Then, we also investigate the effects of socio-economic variable *S* on the probability that subjects provide valid estimates. We take our cues from extensive psychological research on the role that several factors can play in the determination of perceived risks. The variables under study are age (*AGE*), gender (*FEMALE*), education (*SECONDARY, HIGH_SCHOOL,* and *UNIVERSITY*), and the type of education (*SCIENTIFIC*). We expected that the probability of providing valid estimates would

possibly increase for high educated and younger subjects, but we found that older subjects' estimates are more likely to be valid than the others (even though at 10 percent significance level) and education does not affect the validity of individual estimates, at least in our sample (Table 10).

Furthermore, we consider also the interest of subjects on apples and food safety by including in the model a set of dummy variable (*I*) such as being an apple farmer (*PRODUCER*), being an apple consumer (*CONSUMER*), being a member of a consumer association (CONS_ASS), and being resident in the Province of Trento (*TRENTINO*). Although we expected to observe that subjects who reside in the Province of Trento and consume and/or produce apples perform better than the other in terms of validity, perhaps, because they are more interested than the others in the topic, our empirical results suggest no significant explanatory effects for these variables (Table 10).

Finally, we add in our model another set of dummy variables (*TR*) which capture whether subjects trust the predictions of IPCC about temperature and precipitation in 2030 (IPCC_TRUST), the predictions of Edmund Mach Foundation (EMF) about the fire blight's infestation risk in 2030 (EMF_TRUST), and our statement that apple farmers will continue to use the chemical control against apple disease in the future (SCENARIO_TRUST). In this case, we predict that subjects who trust the information we gave them during the experimental instructions more likely provide valid estimates than the others. This is due to the fact that the truster plays the game more carefully. Despite our predictions are confirmed overall, we found the trust in EMF's predictions reduces the probability of providing valid estimates (Table 10).

The consistency of our econometric results with those obtained from non-parametric tests and validity rate's analysis suggests that the provision of real monetary incentives along with random questions increases the validity of elicited estimates.

Moreover, we found that socio-economic variables and the interest of subjects in the topic do not influence the likelihood of providing valid estimates. Only age and trust affect subjects' ability to state valid estimates.

8. Conclusion

This paper has considered the influence of monetary incentives and question ordering on elicitation of subjective probabilities via the Exchangeability Method. In particular, we have shown that incentive compatibility of elicitation mechanisms determines the validity of elicited beliefs, at least in our study. In fact, when subjects are provided with monetary incentive along with sequential questions, which make subjects aware of the chaining and, thus, of the incentive incompatibility of the game, they try to strategically behave in order to get better rewarded at the end of the task and provide invalid subjective probabilities. On the other hand, when subjects are provided with monetary incentive, but random questions, which make the chaining and, thus, the incentive incompatibility of the game less transparent to subjects, they state their real beliefs and return valid subjective probabilities. Non-parametric tests demonstrate that only subjects provided with real monetary incentives and random questions state valid subjective probabilities.

Although non-parametric tests have shown that subjects, who are not provided with monetary incentives, return invalid estimates, we demonstrated, by investigating validity rates, that subjects provided with random questions performs better than those provided with chained questions in terms of validity, given or not monetary incentives. In fact, validity rate for "hypothetical incentives-random questions" are substantially higher than that for "hypothetical incentives-sequential questions". This result is likely due to the fact that sequential questions generate less meaningful tasks where subjects are asked to choose between two prospects that they have just ruled out in previous

questions. This in turn may affect the validity of elicited probabilities, as subjects may invest less cognitive effort in playing the game.

Those interested in using the Exchangeability Method can thus walk away with important messages here. First, incentive compatibility of elicitation mechanisms may affect the validity of elicited beliefs. Subjects are indeed more likely to provide valid estimates, over more of an entire distribution (than one measure of central tendency), if they are rewarded with real monetary incentives based on their performances and presented with experimental design where the chaining is hidden through a particular randomization of the questions. Second, and more disappointing perhaps, is that only a relatively small portion of stated estimates (almost 40%) can be considered valid under the definition we have applied here, which relates to behavioral axioms. The latter implication may be of little surprise to skeptics, but is relevant in our goal to continue to improve ways to provide reliable information about people's subjective probabilities.

Further researches on the validity of subjective probabilities elicited via Exchangeability Method might address these issues at the individual level. Instead of investigating the validity of each single observation, one might investigate the ability of each subject in providing valid estimates. This would be possible by collecting, for each subject, a number of observations large enough to test the validity of her/his stated probabilities by using non-parametric tests.

Acknowledgments

We thank Roberta Raffaelli for help in organizing sessions and running the experiment at the Computable and Experimental Economics Laboratory (CEEL) at University of Trento. We also thank Marco Tecilla for help on designing and organizing the experiment. We appreciate comments on the experimental wording from Ilaria Pertot; on the approach from Matteo Ploner; and on the paper from Roberta Raffaelli, Mary Riddel, and Richard Woodward. This research was funded by Autonomous Province of Trent, project ENVIROCHANGE, Call for Major Project 2006. Shaw acknowledges funding from the U.S.D.A. Regional/Hatch project.

References

Abdellaoui, M., Baillon, A., Placedo, L., Wakker, P.P., 2011. The Rich Domain of Uncertainty: Source Functions and Their Experimental Implementation. *American Economic Review* 101, 695-723.

Andersen, S., Fountain, J., Harrison, G.W., Rutström, E.E., 2010. Estimating Subjective Probabilities. Working Paper 09-01, Department of Economics, College of Business Administration, University of Central Florida.

Alavanja, M.C.R., Hoppin, J.A., Kamel, F., 2004. Health Effects of Chronic Pesticide Exposure: Cancer and Neurotoxicity. *Annual Review of Public Health* 25, 155-197.

Baillon, A., 2008. Eliciting Subjective Probabilities Through Exchangeable Events: an Advantage and a Limitation. *Decision Analysis* 5(2), 76-87.

Brier, G.W., 1950. Verification of Weather Forecasts In Terms of Probability. *Monthly Weather Review* 78(1), 1-3.

Buzby, J.C., Fox, J.A., Ready, R.C., Crutchfield, S.R., 1998. Measuring Consumer Benefits of Food Safety Risk Reductions. *Journal of Agricultural and Applied Economics* 30(1), 69–82.

Cerroni, S., Shaw, W.D., 2012. Does climate change information affect stated risks of pine beetle impacts on forests? An application of the exchangeability method. In press, *Journal of Forest Policy and Economics,* http://dx.doi.org/10.1016/j.forpol.2012.04.001

Corso P.S., Hammit J.K., Graham J.D., 2001. Valuing mortality-risk reduction: Using visual aids to improve the validity of contingent valuation. *Journal of Risk and Uncertainty* 23, 165-184.

Chew S.H., Sagi, J., 2006. Event exchangeability: Probabilistic sophistication without continuity or monotonicity. *Econometrica* 74, 771–786

de Finetti, B, 1937. La prévision: Ses lois logiques, ses sources subjectives. Annales de l'IHP 7, 1–68

de Finetti, B., 1974a. The Value of Studying Subjective Evaluations of Probability. In: Stael Von Holstein C.A.S. (Eds.). The Concept of Probability in Psychological Experiments. Dordrecht-Holland/Boston-U.S.A: Reidel Publishing Company, 1-14.

de Finetti, B., 1974b. The True Subjective Probability Problem. In: Stael Von Holstein C.A.S. (Eds.). The Concept of Probability in Psychological Experiments. Dordrecht-Holland/Boston-U.S.A: Reidel Publishing Company, 15-23.

Edmund Mach Foundation, 2006. Colpo di Fuoco. IASMA Notizie 5, 1-4. [accessed 31.10.2011].

Fiore, S.M., Harrison, G.W., Hughes, C.E., Rustrom, E.E., 2009. Virtual experiments and environmental policy. *Journal of Environmental Economics and Management* 57 (1), 65-86.

Frisch, D., Baron, J., 1988. Ambiguity and rationality. *Journal of Behavioral Decision making* 1, 149-157.

Gerking, S., Khaddaria, R., 2011. Perceptions of Health Risk and Smoking Decisions of Young People. *Health Economics* 21 (7), 865-877.

Gigerenzer, G., Hoffrage, U., 1995. How to improve Bayesian reasoning without instruction - frequency formats. *Psychological Review* 102, 684-704.

Hammit, J.K., Graham J.D., 1999. Willingness to Pay for Health Protections: Inadequate Sensitivity to Probability? *Journal of Risk and Uncertainty* 8, 33-62.

Harrison, G.W., 1986. An experimental test for risk aversion. *Economic Letters* 21, 7-11.

Harrison, G.W., List, J.A., 2004. Field Experiment. *Journal of Economic Literature* 42(4), 1009-1055.

Harrison, G.W., List, J.A., Towe, C., 2007. Naturally Occurring Preferences and Exogenous Laboratory Experiments: A Case Study for Risk Aversion. *Econometrica* 75(2), 433-458.

Jakus, P.M., Shaw, W.D., Nguyen, T.N., Walker, M., 2009. Risk Perceptions of Arsenic in Tap Water and Bottled Water Consumption. *Water Resource Research* 45, W05405, doi:10.1029/2008WR007427.

Kilka, M., Weber, M., 2001. What Determines the Shape of the Probability weighting function under uncertainty. *Management Science* 47, 1712–1726.

Kivi, P.A., Shogren, J.F., 2010. Second-order Ambiguity in Very Low Probability Risks: Food Safety Valuation. *Journal of Agricultural and Resource Economics* 35(3), 443-56.

Manski, C., 2004. Measuring Expectations. *Econometrica* 72 (5), 1329-1376.

Morgan, M.G., Henrion, M., 1990. Uncertainty: A Guide to Dealing with Uncertainty in Quantitative Risk and Policy Analysis. New York: Cambridge University Press.

Németh, J., 2004. Practice of Applying Streptomycin to Control Fireblight in Hungary. Bulletin O*EEP/EPPO* 34, 381-382.

Offerman, T., Sonnemans, J., Van de Kuilen, G., Wakker, P.P., 2009. A Truth Serum for NON-Bayesians: Correcting Proper Scoring Rules for Risk Attitudes. *Review of Economic Studies* 76, 1461-1489.

Provincia Autonoma di Trento. 2010. La Produzione Lorda Vendibile dell'Agricoltura e della Selvicoltura nella Provincia di Trento nel 2006 e nel 2007. Available at:

http://www.statistica.provincia.tn.it/binary/pat_statistica/produzione_lorda_vendibile/Pat_Agricoltura_bassa.1282904118.pdf [accessed 31.10.2011]

Raiffa, H., 1968. Decision Analysis. London: Addison-Wesley.

Riddel, M., Shaw, W.D., 2006. A Theoretically-Consistent Empirical Non-Expected Utility Model of Ambiguity: Nuclear Waste Mortality Risk and Yucca Mountain. *Journal of Risk and Uncertainty* 32(2), 131-150.

Shaw, W.D., Jakus, P.M., Riddel, M., 2012. Perceived Arsenic-Related Mortality Risks for Smokers and Non-smokers. *Contemporary Economic Policy* 30(3), 417-429.

Spetzler, C.S., Stael Von Holstein, C.A.S., 1975. Probability encoding in decision analysis. *Management Science* 22, 340–358.

Van de Kuilen, G., Wakker, P.P., Zou, L., 2006. A Midpoint Technique for Easily Measuring Prospect Theory's Probability Weighting. CREED, University of Amsterdam, The Netherlands.

Viscusi, V.K., 1990. Do Smokers Underestimate Risks? *Journal of Political Economy* 98 (6), 1253-68.

Vossler, C.A., Evans, M.F., 2009. Bridging the gap between the field and the lab: Environmental goods, policy maker input, and consequentiality. *Journal of Environmental Economics and Management* 58(3), 338-345.

Wakker P., Deneffe D., 1996. Eliciting von Neumann-Morgenstern Utilities When Probabilities are Distorted or Unknown. *Management Science* 42(8), 1131-1150.

Weinstein, N.D., Diefenbach, M.A., 1997. Percentage and Verbal Category Measures of Risk Likelihood. *Health Education Research* 12 (1), 139-41.

Williams, P.R.D., Hammitt, J.K., 2001. Perceived Risks of Conventional and Organic Produce: Pesticides, Pathogens, and Natural Toxins. *Risk Analysis* 21(12), 319-330.

Appendix A: Games' examples

Example 1. First question of the Exchangeability Game for the variable g

I prefer to bet 100€ on the fact that the number of days of April in which the *fire blight* infestation will occur with certainty in 2030 is:

□	□
smaller than g_a[a]	greater than or equal to g_a[a]

[a] $g_a=\{g_0 + [(g_1-g_0)/2]\}$

Example 2. First question of the Repeated Exchangeability Game Test for the variable $g_{1/2}$'

I prefer to bet 100€ on the fact that the number of days of April in which the *fire blight* infestation will occur with certainty in 2030 is:

□	□
greater than $g_{1/4}$ and smaller than $g_{1/2}$	greater than or equal to $g_{1/2}$ and smaller than $g_{3/4}$

Example 3. A question of the Certainty Equivalent Game for $g_{1/2}$

In each of the following question, do you prefer to play the lottery presented in Option A or do you prefer to take the amount of money presented in Option?

Option A			**Option B**
You win 100€ if the number of days of April in which the *fire blight* infestation will occur with certainty in 2030 is SMALLER THAN $g_{1/2}$ 0€, otherwise	□	□	0€
	□	□	25€
	□	□	49€
	□	□	51€
	□	□	75€
	□	□	100€

In each of the following question, do you prefer to play the lottery presented in Option A or do you prefer to take the amount of money presented in Option?

Option A			**Option B**
You win 100€ if the number of days of April in which the *fire blight* infestation will occur with certainty in 2030 is GREATER THAN OR EQUAL TO $g_{1/2}$ 0€, otherwise	□	□	0€
	□	□	25€
	□	□	49€
	□	□	51€
	□	□	75€
	□	□	100€

Appendix B: Definition, axioms and theorems of probability theory

Let G_j^i be disjoint events with $i = \{1,.., n\}$ and $j = n$ and S_G be a sample space, then:

Statement 1. $P(S_G) = 1$

Consider the sample space S_G, we impose that $S_G = G_1^1 = 1$ by telling respondents that the probability associated to the entire sample space is equal to 1, say $S_G = G_1^1 = 1$.

Statement 2. $P(G_j^i) \geq 0$

Consider $P(G_2^1)$ and $P(G_2^2)$, we impose that $P(G_2^1) \geq 0$ and $P(G_2^2) \geq 0$ by asking respondents to the lower (g_0) and upper (g_1) bounds of the event space outside of which they are essentially certain the outcome cannot happen at all. This is basically the first question of Exchangeability Game.

Statement 3. If $\{G_j^i\}$ is a sequence of disjoint sets in S_G, then

$$P\left(\bigcup_{i=1}^{n} G_j^i\right) = \sum_{i=1}^{n} P(G_j^i)$$

Consider $P(G_2^1)$ and $P(G_2^2)$, "*exchangeability*" assumption imposes that

$$P\left(G_2^1 \bigcup G_2^2\right) = P(G_2^1) + P(G_2^2) = 0.5$$

Statement 4. $P(G_j^i) = 1 - P(G_j^{iC})$

Consider $P(G_2^1)$ and $P(G_2^2)$, "*exchangeability*" assumption imposes that

$$P(G_2^1) = 1 - P(G_2^2) = 0.5 = 1 - 0.5$$

Statement 5. $P(\phi) = 0$

See Statement 2.

Statement 6. For each $G_j^i \in S_G$, then $0 \leq P\left(G_j^i\right) \leq 1$

See Statements 1 and 2.

Statement 7. If $G_j^i \subset G_n^i$ with $n = jk, k \in N, k \neq 0$, then $P\left(G_j^i\right) \geq P\left(G_n^i\right)$

Consider G_4^1 and G_2^1, "*exchangeability*" assumption imposes that $P\left(G_2^1\right) = 0.5 \geq P\left(G_4^1\right) = 0.25$

Tables

Table 1. McNemar's test of consistency		
Treatment	**Null Hypothesis**	χ^2
Real incentives-Sequential questions	P(AB)[a] = P(BA)[b]	1.60
Real incentives-Random questions	P(AB) = P(BA)	0.31
Hypothetical incentives-Sequential questions	P(AB) = P(BA)	0.82
Hypothetical incentives-Random questions	P(AB) = P(BA)	1.32

[a] P(AB) is the probability of choosing prospect A in the original question and prospect B in the repeated question.
[b] P(BA) is the probability of choosing prospect B in the original question and prospect A in the repeated question.
*1% significance level
**5% significance level
***10% significant level

Table 2. Summary statistics of median values obtained via EG ($X_{1/2}$) and REG ($X_{1/2}$')						
Treatment	**Variable**	**Obs**	**Mean**	**St.Dev.**	**Min**	**Max**
Real incentives-Sequential questions	$X_{1/2}$	24	44.37	27.69	7	94
	$X_{1/2}$'	24	44.96	27.87	7	94
Real incentives-Random questions	$X_{1/2}$	40	44.05	26.17	2	96
	$X_{1/2}$'	40	44.17	25.98	3	96
Hypothetical incentives-Sequential questions	$X_{1/2}$	22	54.91	28.03	5	94
	$X_{1/2}$'	22	55.91	28.08	7	94
Hypothetical incentives-Random questions	$X_{1/2}$	26	40.35	28.74	3	94
	$X_{1/2}$'	26	40.65	28.27	3	96

Table 3. Results at sample level obtained via EG ($X_{1/2}$) and REG ($X_{1/2}$')			
		Wilcoxon matched-pairs signed ranks test	**Binomial sign test**
Treatment	**Null Hypothesis**	**Z**	**P>Z**
Real incentives-Sequential questions	Median($X_{1/2}$) =Median($X_{1/2}$')	-2.234**	0.062
Real incentives-Random questions	Median($X_{1/2}$) =Median($X_{1/2}$')	-0.665	0.480
Hypothetical incentives-Sequential questions	Median($X_{1/2}$) = Median($X_{1/2}$')	-1.880***	0.125
Hypothetical incentives-Random questions	Median($X_{1/2}$) = Median($X_{1/2}$')	-1.174	0.266

*1% significance level
**5% significance level
***10% significant level

Table 4. Summary statistics of the Certainty Equivalents obtained via CEG						
Treatment	**Variable**	**Obs**	**Mean**	**St.Dev.**	**Min**	**Max**
Real incentives-Sequential questions	CE_{L1}	143	51.21	46.38	0	125
	CE_{L2}	143	76.95	44.69	0	125
Real incentives-Random questions	CE_{L1}	167	59.80	42.31	0	125
	CE_{L2}	167	68.22	41.72	0	125
Hypothetical incentives-Sequential questions	CE_{L1}	136	70.80	43.30	0	125
	CE_{L2}	136	75.86	42.14	0	125
Hypothetical incentives-Random questions	CE_{L1}	115	55.65	36.14	0	125
	CE_{L1}	115	73.17	37.11	0	125

Table 5. Results at sample level obtained via the CEG

		Wilcoxon matched-pairs signed ranks test	Binomial sign test
Treatment	**Null Hypothesis**	**Z**	**P>Z**
Real incentives-Sequential questions	Median(CE_{L1}) = Median(CE_{L2})	-3.713*	0.002
Real incentives-Random questions	Median(CE_{L1}) = Median(CE_{L2})	-1.513	0.304
Hypothetical incentives-Sequential questions	Median(CE_{L1}) = Median(CE_{L2})	-1.283	0.088
Hypothetical incentives-Random questions	Median(CE_{L1}) = Median(CE_{L2})	-3.005*	0.000

*1% significance level
**5% significance level
***10% significant level

Table 6. Percentage of rewarded subjects based on their answers to Question 3

Treatment	Number of Subjects	Number of Rewarded Subjects	Percentage of Rewarded Subjects
Real Incentives-Sequential Questions	22	11	50.00
Real Incentives-Random Questions	23	8	34.78

Table 7. Validity rates (*V*) for all treatments

Treatment	Variable	Number of observations	Number of valid observations	*V* (%)
Real incentives-Sequential questions	First Quartile	66	15	22,72
	Median	66	24	36,36
	Third Quartile	66	13	19,69
	Total	192	52	26,26
Real incentives-Random questions	First Quartile	69	25	36,23
	Median	69	34	49,27
	Third Quartile	69	22	31,88
	Total	207	81	39,13
Hypothetical incentives-Sequential questions	First Quartile	57	13	22,80
	Median	57	15	26,31
	Third Quartile	57	10	17,54
	Total	171	38	22,22
Hypothetical incentives-Random questions	First Quartile	48	12	25,00
	Median	48	18	37,50
	Third Quartile	48	13	27,08
	Total	144	43	29,86

Table 8. Percentage of probabilities per level of invalidity in each treatment

Level of Invalidity	Δ(CE)	TRC	TRU	THC	THU	Total
Very Low	< 27	23.71	31.18	30.93	40.54	31.02
Low	27-51	12.37	18.28	24.74	17.57	18.28
Medium	52-76	8.25	17.20	11.34	9.46	11.63
High	77-100	7.22	10.75	6.19	8.11	8.03
Very High	>101	48.45	22.58	26.80	24.32	31.02

Table 9. Description of dependent and independent variables of Model 1					
Variable	**Definition**	**Mean**	**St.Dev.**	**Min**	**Max**
VALID	= 1 if valid, = 0 otherwise	.368	.482	0	1
TRS	= 1 if "Real Incentives-Sequential Questions" treatment, = 0 otherwise	.275	.446	0	1
TRR	= 1 if "Real Incentives-Random Questions" treatment, = 0 otherwise	.287	.452	0	1
THS	= 1 if "Hypo Incentives-Sequential Questions" treatment, = 0 otherwise	.237	.425	0	1
THR	= 1 if "Hypo Incentives-Random Questions" treatment, = 0 otherwise	.200	.400	0	1
G	Number of days when the infestation risk is extremely high in April	.333	.471	0	1
A	Number of apple containing at least one pesticide residue	.333	.471	0	1
R	Number of apple containing multiple pesticide residue	.333	.471	0	1
50^{th} PERCENTILE	Observations related to the median of G, A, and R	.333	.471	0	1
25^{th} PERCENTILE	Observations related to the I quartile of G, A, and R	.334	.471	0	1
75^{th} PERCENTILE	Observations related to the II quartile of G, A, and R	.333	.471	0	1
CONSUMER	= 1 if the subject eats at least 3 apples a week = 0 otherwise	.478	.500	0	1
CONS_ASS	= 1 if the subject is a member of a consumer association = 0 otherwise	.062	.242	0	1
PRODUCER	= 1 if the subject produces apples = 0 otherwise	.037	.190	0	1
TRENTINO	= 1 if the subject resides in the province of Trento = 0 otherwise	.737	.440	0	1
IPCC_TRUST	Trust in IPCC's predictions of the future temperature and precipitation (at 5 levels)[a]	2.950	.545	0	4
FEM_TRUST	Trust in FEM's predictions of fire blight's infestation risk in the future (at 5 levels)[a]	2.587	.684	0	4
SCENARIO_TRUST	Agreement with the fact that farmers will use the chemical	2.912	.778	0	4

	control in the future (at 5 levels)[b]				
AGE	Age in years	32.746	12.578	19	68
FEMALE	= 1 if female, = 0 otherwise	.4366	.4994	0	1
SECONDARY_SCHOOL	= 1 if the subject have this education level, = 0 otherwise	.1830	.3895	0	1
HIGH_SCHOOL	= 1 if the subject have this education level, = 0 otherwise	.5070	.5035	0	1
UNIVERSITY	= 1 if the subject have this education level, = 0 otherwise	.3098	.4657	0	1
SCIENTIFIC	= 1 if the subject have a scientific education = 0 otherwise	.487	.500	0	1

[a] From 0= very high trust to 4= very low trust
[b] From 0=strongly disagree to 4= strongly agree

Table 10. Generalized Linear Model Estimation of Models 1and 2

Dependent Variable: VALID

Variable	**Model 1**	**Model 2**[a]
TRS	.370**	.370
TRR	.648*	.648**
THR	.385**	.385
A	-.058	-.058
R	-.173	-.173
MEDIAN	-.077	-.077
25^{th} PERC	-.094	-.094
FEMALE	-.097	-.097
AGE	.019*	.019***
SEC_SCHOOL	-.086	-.086
HIGH_SCHOOL	-.016	-.016
SCIENTIFIC	.173	.173
PRODUCER	.584***	.584
CONSUMER	-.021***	-.021
CONS_ASS	.312	.312
TRENTINO	.067	.067
IPCC_TRUST	.359*	.359***
FEM_TRUST	-.355*	-.355**
SCEN_TRUST	.253*	.253***
CONSTANT	-2.160*	-2.160**
LOG L.HOOD	-347.702	-347.702

[a] Robust standard errors and clustering effects
*1% significance level
**5% significance level
***10% significant level

Figure 1. Scheme of the Exchangeability Game's bisection procedure

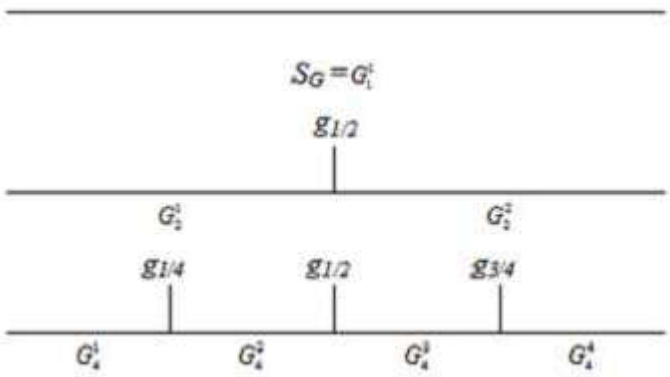

Figure 2. Histogram of the percentage of subjective probabilities per level of invalidity