J Risk Uncertain (2012) 44:161–180
DOI 10.1007/s11166-011-9133-1

Belief elicitation in the presence of naïve respondents: An experimental study

Li Hao · Daniel Houser

Published online: 23 December 2011

Abstract It is often of interest to elicit beliefs from populations that may include naïve participants. Unfortunately, elicitation mechanisms are typically assessed by assuming optimal responses to incentives. Using laboratory experiments with a population that potentially includes naïve participants, we compare the performance of two elicitation mechanisms proposed by Karni (Econometrica 77(2):603-606, 2009). These mechanisms, denoted as "declarative" and "clock," are valuable because their incentive compatibility does not require strong assumptions such as risk neutrality or expected utility maximization. We show that, theoretically and empirically, with a sufficient fraction of naïve participants, the clock mechanism elicits beliefs more accurately than the declarative. The source of this accuracy advantage is twofold: the clock censors naïve responses, and participants are more likely to employ dominant strategies under the clock. Our findings hold practical value to anyone interested in eliciting beliefs from representative populations, a goal of increasing importance when conducting large-scale surveys or field experiments.

Keywords Belief elicitation · Declarative mechanism · Clock mechanism · Proper scoring rules · Laboratory experiment

JEL Classification C91

It is often of interest to elicit beliefs (or subjective probabilities) from populations that might include naïve participants. Unfortunately, belief elicitation mechanisms

L. Hao (✉)
Department of Economics, Sam Walton College of Business, University of Arkansas, Fayetteville, AR 72701, USA
e-mail: lhao@walton.uark.edu

D. Houser
Interdisciplinary Center for Economic Science and Economics Department, George Mason University, Fairfax, VA 22030, USA
e-mail: dhouser@gmu.edu

are typically assessed under the assumption that all people respond optimally to incentives. This paper relaxes that assumption, and evaluates the performance of two elicitation mechanisms using laboratory experiments with a subject population that potentially includes naïve participants. These mechanisms, proposed by Karni (2009), are valuable because their incentive compatibility does not require the strong assumptions that (i) utility functions are linear in money (i.e., risk-neutrality) or (ii) probability weighting functions are identity transformations (i.e., expected utility).

We denote Karni's (2009) mechanisms as "declarative" and "clock." In the declarative mechanism, the respondent declares directly her subjective probability. In the clock mechanism, she participates in an ascending clock, where the only other participant is a computerized player who exits stochastically. The clock stops when either the human or computer participant exits, or when the clock reaches its terminal value, whichever occurs first. The subject's stopping point represents her belief. Both mechanisms provide incentives to encourage respondents to report beliefs truthfully.

An important advantage to comparing mechanisms using laboratory experiments is that subject populations typically include mixtures of "sophisticated" and "naïve" types, in the sense that some but not all subjects respond optimally to the incentives of the mechanism (see, e.g., Andreoni 1995; Houser and Kurzban 2002; Houser et al. 2004). In an environment with a mixture of types, we show that the clock can demonstrate an advantage in elicitation accuracy (under certain conditions) because it censors naïve responses. Indeed, our experiment's results confirm that in our environment the clock mechanism does elicit beliefs significantly more accurately than the declarative mechanism. Moreover, our analysis reveals that the advantage of the clock mechanism is not solely due to censoring. Rather, we find respondents are more likely to adopt dominant strategies under the clock. These findings have substantial practical value to anyone interested in eliciting beliefs from representative populations, a goal of increasing importance, for example, when designing large-scale surveys.

The Karni (2009) mechanisms we study depart from the widely used proper scoring rules (e.g., Nyarko and Schotter 2002; Palfrey and Wang 2009).[1] Although popular, the incentive compatibility of proper scoring rules does not hold when the respondent is not risk-neutral or when s/he deviates from expected utility maximization. In light of this, Offerman et al. (2009) proposed that one first estimate risk attitudes and then adjust elicited beliefs accordingly. Alternatively, Andersen et al. (2010) suggested that one estimate jointly subjective beliefs and risk attitudes using maximum likelihood methods.

Allen (1987) was first to generalize the quadratic scoring rule by utilizing binary lottery payoffs to induce risk-neutrality.[2] McKelvey and Page (1990) implemented a similar generalized quadratic scoring rule in an early laboratory study of belief

[1] A scoring rule is "proper" if the respondent must report true beliefs to maximize expected score. It was first introduced by meteorological statistician Brier (1950), and later popularized by Savage (1971). Reported beliefs are compared against realized outcomes, so that proper scoring rules provide incentives for accuracy. The quadratic, spherical, and logarithmic scoring rules are examples of proper scoring rules.

[2] Allen's (1987) idea is to treat the scores as probabilities of winning a prize in a binary lottery. This technique was first investigated by Roth and Malouf (1979), Roth and Murnighan (1982), and Roth and Schoumaker (1983).

elicitation. Recently, Schlag and van der Weele (2009) extended Allen's (1987) mechanism to elicit parameters characterizing subjective probability distributions. Hossain and Okui (2011) provided empirical results showing that quadratic scoring rules with and without binary lottery payoffs perform equally well.

To the best of our knowledge, no previous empirical comparison of Karni's (2009) mechanisms has appeared. Despite the potential equivalence of the two mechanisms,[3] substantial evidence suggests theoretically equivalent mechanisms can perform differently in practice (e.g., second-price sealed-bid and English clock auctions such as reported by Kagel et al. 1987; Kagel and Levin 1993, 2009; Rutström 1998; Harstad 2000). In light of this, it seems of significant practical importance to investigate whether and how elicitations from declarative and clock mechanisms might differ.[4]

It is worthwhile to note that the clock and declarative mechanisms are not isomorphic in theory. The reason is that the clock mechanism fails to elicit an exact probability whenever the computerized bidder is the first to exit, because the participant did not have an opportunity to respond. At first glance, this data censoring may appear to be a source of inefficiency. While this is possible, it turns out that if there exist both naïve and sophisticated respondents, the clock can be structured so that it is more likely to censor naïve than sophisticated responses. As a result, the clock can be more accurate than the declarative mechanism.

The contributions of this paper are threefold. First, by focusing on novice respondents who are likely to submit naïve responses, we provide practical guidance for belief elicitation in contexts including large-scale survey environments. Indeed, in recent years large-scale belief elicitation (presumably from novice respondents) has become a particularly active area (e.g., Manski 2004; Bellemare et al. 2008). Our investigation sheds light on how to implement incentive-compatible belief elicitation efficiently in environments that include noisy responses.[5]

Second, our results contribute to the applied mechanism design literature by demonstrating another environment where the (ascending) clock maintains its advantage in inducing truth-revealing dominant strategies. For example, it is well known that the equivalence of sealed-bid second-price auctions and English clock auctions quickly breaks down in practice, in the sense that bids in English clock auctions are much closer to the truth in both induced (Kagel et al. 1987; Kagel and Levin 1993, 2009; Harstad 2000) and home-grown value settings (Rutström 1998).

The third contribution of this research is that it studies belief elicitation when participants are endowed with objective beliefs.[6] This allows us to shed light on the performance of the mechanisms absent noise due to variations in subjects' abilities

[3] On page 604, Karni (2009) introduces the clock mechanism by saying, "An equivalent probability-elicitation auction mechanism is as follows…"

[4] Previous studies including Grether (1992), Möbius et al. (2011) and Holt and Smith (2009) have implemented the declarative mechanism in laboratory experiments; Köszegi and Rabin (2008) also discussed this mechanism.

[5] However, there exist practical difficulties when using incentive compatible mechanisms in the field to elicit one's beliefs regarding, for example, the chance of changing jobs within the next 5 years. As we discuss in the conclusion, future research could address these important issues.

[6] Some subjects may fail to adopt the beliefs with which they are endowed. However, this would occur in both mechanisms, and thus presents no difficulty for our comparative analysis (other than adding noise and leaving it more difficult to discover differences).

to predict uncertain events. Garthwaite et al. (2005) point to the importance of doing this when they write, "it is important to distinguish between the quality of an expert's knowledge and the accuracy with which that knowledge is translated into probabilistic form," a distinction Winkler and Murphy (1968) refers to as "substantive goodness" and "normative goodness" respectively. Our subjects are perfectly informed and thus possess "substantive goodness," so that our data imply the clock mechanism is better than the declarative at facilitating probabilistic formulations by novice respondents.

The paper proceeds as follows. Section 1 reviews Karni's (2009) theory, Section 2 formulates our hypothesis, Sections 3 and 4 report experimental design and results, Section 5 discusses practical details regarding field implementation of the mechanisms, and Section 6 concludes.

1 Review of Karni's (2009) theory

This section briefly reviews Karni's (2009) mechanisms. In Savage's (1954) framework, an individual holds the belief that an event E will occur with probability $\pi(E)$. If E occurs, the individual receives the prize x; otherwise, she receives y ($x>y$). We call a mapping between the occurrence (and non-occurrence) of an event and monetary payoffs a "bet," denoted by $\beta: = x_E y$.

Consider a lottery that pays x with probability r or y with probability $1-r$; denote this lottery by $L(r)$. The number r is randomly selected from a uniform distribution on [0, 1]. The individual knows the distribution, but she does not know r when she makes her decision.

1.1 Declarative mechanism

The individual submits a decision, $\mu\in[0, 1]$, which is compared with the random number r. If $\mu\geq r$, she plays the bet β; if $\mu<r$, she plays the lottery $L(r)$.

Dominant strategy Karni (2009) demonstrates that the unique dominant strategy in this mechanism is to report truthfully: $\mu = \pi(E)$. Doing so guarantees that the individual obtains either the bet β or the lottery $L(r)$, whichever has the higher probability of winning the prize x. The individual has no incentive to report a number greater than the truth, because as soon as the random number r falls between the truth and her report, $\pi(E) < r < \mu$, she receives the bet β, and forgoes the lottery $L(r)$ that has a higher winning probability. The same logic applies when her report is smaller than the truth.

1.2 English clock mechanism

In the English clock auction mechanism, the individual competes with a dummy bidder and knows that the dummy bidder exits the auction at (unknown) number r. The clock starts at 0 and rises continuously as long as both the individual and the truth-revealing dummy bidder are "in the auction." The clock stops when at least one bidder drops out, or when the clock reaches 1, whichever occurs first. If the individual is the first to exit, she receives the lottery $L(r)$; if the dummy bidder exits first, the individual receives the bet β.

Dominant strategy Following Karni's (2009) argument, the dominant strategy is to stay in the auction as long as the clock is below $\pi(E)$, and exit exactly at $\pi(\mathrm{E})$.

1.3 Assumptions

For Karni's (2009) mechanisms, as well as proper scoring rules, a necessary condition for incentive compatibility is the no-stakes condition (Kadane and Winkler 1988). The no-stakes condition requires the wealth of an individual, excluding elicitation-related payoffs, to be independent of the occurrence (or nonoccurrence) of the event.[7]

Incentive compatibility in Karni's (2009) mechanisms requires also that an individual's preferences exhibit dominance and probabilistic sophistication (Machina and Schmeidler 1992).[8] This condition is weaker than the requirement that preferences satisfy expected utility.

2 Comparison of the mechanisms

2.1 Censoring can lead to greater accuracy in elicitations

In this section we demonstrate that, under certain conditions, the clock's censoring can lead to accuracy advantages over the declarative mechanism. We say that a mechanism is more accurate if its fraction of truthful elicitations is greater. Assuming that people use identical decision strategies with the two mechanisms,[9] the general conditions under which the clock displays greater accuracy are that (i) the population includes a sufficient fraction of naïve decision makers and (ii) the clock censors sufficiently few "optimal" decisions. We show below that these specifics can vary depending on the nature of the environment.

To develop this point, suppose there are two types of respondents: i) truth-revealers who report $\pi(\mathrm{E})$ and ii) naïve agents whose responses are characterized by a distribution with c.d.f. $F_n(\bullet)$. Let the proportions of the optimal and naïve types be α and $1-\alpha$ respectively ($0 < \alpha < 1$).[10] Suppose also that the random number r

[7] As noted by Kadane and Winkler (1988), the elicited probabilities intertwine with utilities "not just through the explicit or implicit payoffs related to the elicitation process, but also through other stakes the individual may have in the events of interest." Previous work including Karni (1999) and Jaffray and Karni (1999) proposed elicitation procedures when the no-stakes condition is violated. However, we are not aware of any evidence informing the extent to which this violation matters empirically.

[8] In essence, probabilistic sophistication means that the individual ranks bets with subjective probabilities over outcomes in a similar fashion as she would rank lotteries with an objective probability distribution.

[9] Our analysis suggests people use different strategies, but our development below shows that the clock can have an accuracy advantage even when this is not the case.

[10] If all decisions are optimal (or if all decisions are naïve), the clock mechanism has no advantage from censoring.

(the probability of winning the lottery prize) has continuous and strictly increasing c.d.f. $F_r(\bullet)$.

By definition, the accuracy of the declarative mechanism equals α, the fraction of optimal decisions in the population.

The accuracy of the clock mechanism is given by the following expression:

$$\frac{\alpha(1-F_r(\pi))}{\alpha(1-F_r(\pi))+(1-\alpha)\left(1-\int_0^1 F_r(u)dF_n(u)\right)}$$

The numerator is the fraction of optimal decisions that are not censored by the clock. The denominator is the total fraction of non-censored decisions. Thus, the value of the expression is the fraction of truthful elicitations in the population.[11]

It immediately follows that the clock is more accurate than the declarative mechanism when the following condition holds:

$$\frac{\alpha(1-F_r(\pi))}{\alpha(1-F_r(\pi))+(1-\alpha)\left(1-\int_0^1 F_r(u)dF_n(u)\right)}-\alpha>0 \tag{1}$$

Inequality (1) makes clear that whether the clock holds an accuracy advantage in relation to the declarative mechanism depends critically on the fraction of optimal decision makers α as well as the value of the true belief π in relation to the distribution of the random number r (which influences the fraction of censored optimal decisions).

2.2 Experiment hypothesis

Our interest is in testing the following hypothesis:

Hypothesis Beliefs elicited using the clock mechanism are more likely to be accurate than beliefs elicited using the declarative mechanism, especially with novice participants.

To develop this, suppose now that $F_r(\bullet)$ and $F_n(\bullet)$ are both U(0,1).[12] Under these assumptions it is easy to show that inequality (1) simplifies to

$$\frac{\alpha(1-\pi)}{\alpha(1-\pi)+(1-\alpha)/2}-\alpha>0 \tag{2}$$

Given $0<\alpha<1$, (2) holds if and only if

$$0<\pi<0.5 \tag{3}$$

[11] Note that this approach does not make any use of censored observations. An alternative is to take a censored belief as the interval between the clock's stopping point and the clock's upper limit. However, the transformation from interval estimates to point estimates is arbitrary (e.g., using the mean of the interval as the estimate of the belief), and results would generally be sensitive to this transformation. We do not pursue this approach here.

[12] Karni specified the uniform distribution on [0, 1] for $F_r(\bullet)$, but we note that properties of the mechanisms remain the same for any continuous and strictly increasing distribution $F_r(\bullet)$.

Thus, the clock has an advantage in this environment when true beliefs are less than 0.5, and not everybody makes optimal decisions. This result guides our experiment design, as we detail below.

3 Experiment design and procedures

A key feature of our design is that we endow subjects with objective beliefs. We made this choice for two reasons. First, as noted earlier, our goal is to assess the mechanisms' ability to facilitate accurate translation from participants' knowledge to probabilities, so we eliminate differences in quality of their knowledge. In addition, doing this better connects our research to second-price and English clock auctions, where participants typically make decisions using known (induced) values. A transcript of subjects' instructions can be found in the Appendix.

3.1 Declarative mechanism

Endowed belief= 0.2 The subject is presented with two opaque bags (physical; made of cloth), bag A and bag B. She knows that bag A has 10 chips in total: 2 white chips and 8 black chips. She also knows that bag B also has a total of 10 chips of the two colors, but the number of white chips (denoted by R) is equally likely to be any integer from 1 to 9.

The participant submits a number between 1 and 9 (inclusive; integer only[13]) on a computer terminal.[14] This number is then compared with R, i.e., the number of white chips in Bag B. If the submitted number is greater than R, the subject draws a chip from bag A; otherwise, she draws a chip from bag B. In either case, the subject is paid \$10 if she draws a white chip, and is paid \$1 if she draws a black chip.[15]

Endowed belief= 0.3 This proceeds exactly as the above procedure, except now there are 3 white chips (and 7 black chips) in bag A.

Dominant strategy Take bag A as the default choice; the declarative mechanism is effectively asking the subject, "*What is the minimum number of white chips in bag B so that you are willing to switch to bag B?*" The dominant strategy is to declare either the number of the white chips in bag A, or one more than the number of white chips in bag A.[16] The presence of two equally advantageous actions stems from our using a discrete state space.

[13] Since both lotteries are presented using integers only, we chose to constrain the decisions also to integers for simplicity and transparency. Moreover, words such as "probabilities" or "distributions" were not used during our experiment.

[14] To maintain symmetry with the clock procedure, all decisions were submitted via computers.

[15] The subject physically drew a chip from the appropriate cloth bag. See Section 3.4 for details.

[16] The two dominant strategies are equivalent because the individual ends up with the same bag except when R is the same as the number of white chips in bag A, as the former strategy leads to bag A while the latter leads to bag B. However, in this case the two bags have the same number of white chips, so the chance of winning \$10 is identical.

3.2 Clock mechanism

Endowed belief=0.2 Bag A and bag B are exactly the same as in the declarative mechanism with the endowed belief of 0.2. Instead of declaring a number, the subject participates in a computerized clock auction. Similar to an English clock auction (e.g., Kagel et al. 1987), it starts as the computer screen displays the number 1 for 5 seconds, and then the number 2 for 5 seconds, and so on. The subject exits the auction by pressing the space key. The clock stops when the subject exits, or after reaching and displaying the number R for 5 seconds, whichever comes first.[17] If the clock stops due to reaching number R, the subject draws a chip from bag A; if the clock stops due to the subject's exit, she draws a chip from bag B.[18] In either case, the subject is paid $10 if she draws a white chip and $1 if she draws a black chip.

Endowed belief=0.3 This proceeds exactly as the above procedure, except now there are 3 white chips (and 7 black chips) in bag A.

Dominant strategy Considering bag A as the default choice, the clock mechanism is effectively asking the subject, "*The number displayed on the screen is the minimum number of white chips in bag B; do you want to switch to bag B now?*" The dominant strategy is to indicate the willingness-to-switch by exiting as soon as the displayed number is the same as the number of the white chips in bag A, or one more than the number of white chips in bag A (see footnote 17).

3.3 Treatment design

Each subject participated in two independent elicitation tasks, which occurred in round one and round two, respectively. The second round was a "surprise" as we announced it only after completing the first round.[19] Table 1 summarizes our two-by-two treatment design.

Each session consisted of 4 to 8 participants and a heterogeneous belief environment: half were endowed with beliefs equal to 0.2 and the other half with beliefs equal to 0.3. Subjects were given new instructions in the second round.

3.4 Procedures

All sessions were conducted between April and October 2009 at the Interdisciplinary Center for Economic Science (ICES) laboratory of George Mason University in Fairfax, VA. Subjects were invited via emails sent to a large undergraduate subject pool, and a total of 130 participated. Subjects were

[17] The screen displays number R for 5 seconds, and the subject were aware that they were able to drop out at R and obtain bag A.

[18] The dummy bidder in our experiment is implemented by stopping the clock when it reaches number R.

[19] Subjects were not told there was a second round at the beginning of the experiment. Upon finishing the first round, the experimenter announced, "That was the end of the experiment. However, we still have some time left; let us do another experiment so you can make more money."

Table 1 Treatment design

Treatment	Number of Subjects (total=130)	First Round	Second Round
D+D	24	Declarative	Declarative
D+C	29	Declarative	Clock
C+D	37	Clock	Declarative
C+C	40	Clock	Clock

paid a guaranteed $5 plus their earnings in the experiment. Average total earnings were $16, and sessions lasted between 30 and 60 minutes. The experiments were partially computerized: we used physical bags and chips to illustrate the lotteries and perform the random draws, whereas the ticking clock was programmed using E-prime.[20]

To facilitate understanding of the mechanisms, subjects were first given abundant time to read the instructions. Following this, the experimenter read the instructions aloud to them. Each subject then took a quiz,[21] and their answers were recorded. The majority of the subjects correctly answered all questions. The experimenter then announced and explained the correct answers. Throughout the experiment, we did not use words such as "probability" or "percent chance," and subjects made all decisions in whole numbers.[22]

Also, we generated the random number *R* (number of white chips in bag B) using the following three steps. In step one, the experimenter showed subjects a deck of nine cards, numbered from 1 to 9. The experimenter then put each card into one of nine opaque envelopes and shuffled the envelopes thoroughly. In step two, each subject was asked to pick an envelope and immediately return it to the experimenter without opening it. At this point, the experimenter wrote the subject's ID on the envelope. Finally, in step three, the experimenter publicly opened each envelope from a distance (so that no subject could read the numbers inside), transcribed the random number *R* for each subject ID, and then sealed the envelope. The reason for these steps was to demonstrate that the number *R* was determined prior to the subjects' decisions, as well as to make clear that *R* was an integer between 1 and 9, each with equal probability. For the surprise second round, a new random number was generated for each subject from a new set of nine opaque envelopes, using exactly the same procedures as described above.

Finally, we implemented the payment procedure individually to ensure that subjects knew they were making independent decisions. After the bag of payoff was determined, the experimenter went to a subject with the appropriate bag and chips, which the subject examined. The experimenter put the 10 chips into the opaque cloth bag. The subject then drew one chip from the bag while keeping his or her head turned away. The subject earned $10 if he or she drew a white chip and $1 otherwise.

[20] E-Prime is a commercial software for computerized experiment design commonly used in psychology research: http://www.pstnet.com/eprime.cfm

[21] The quiz was designed to test whether subjects understood how various hypothetical scenarios and decisions are translated into payoffs. The quiz is available upon request.

[22] Hoffrage et al. (2000) showed that natural frequencies are much better than percent chances at facilitating statistical reasoning of people including experts and non-experts.

4 Results

We organize our results in two subsections. The first describes decisions from the first round, and our first result is that novice subjects are more likely to report their endowed beliefs in the clock mechanism. Our second result shows that the subjects use different strategies between the two mechanisms, and that the distribution of clock data more accurately characterizes the distribution of endowed beliefs.

The second subsection presents decisions from the second round, and our results show that the clock and declarative mechanisms are equally likely to elicit endowed beliefs; however, the distribution of clock data continues to reflect more closely the distribution of endowed beliefs.

4.1 Responses from novice participants

With heterogeneous beliefs of 0.2 and 0.3, Table 2 describes individual decisions in the first round.[23]

Among the 53 and 77 observations in the declarative and clock mechanisms respectively, the proportions of optimal decisions are 47% versus 39%, and non-optimal decisions are 53% versus 22%. *In the declarative mechanism, fewer than half of novice responses are optimal*. This suggests that the dominant strategies in this environment are not trivial to subjects.[24] This is especially significant in light of our explicit effort for simplicity and transparency.[25]

Overall, the clock mechanism censors 39% of decisions. In comparison, if the population consisted of only optimal decisions, the proportion would be between 25% and 35%. This suggests that not everybody in our experiment makes optimal decisions.

The mean deviations, as well as the mean absolute deviations from optimal strategies, are significantly different from zero in both mechanisms.[26] They are smaller in the clock mechanism than in the declarative mechanism, but the difference is insignificant.

We now take a closer look at the two mechanisms by excluding censored decisions. Note that we do not consider alternative approaches that make use of information in censored decisions (see footnote 12). Our first result is as follows:

> **Result 1.** *With novice respondents, beliefs elicited using the clock mechanism are more likely to be accurate than beliefs elicited using the declarative mechanism.*

Evidence Among elicited beliefs, the proportions of optimal decisions are 64% and 47% in the clock and declarative mechanisms respectively (Fig. 1). A two-sided

[23] The two equally dominant strategies are both set as deviation of 0. In particular, when endowed belief is 0.2, decisions {0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9} are converted into deviations {−0.1, 0, 0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6}; when endowed belief is 0.3, they are {−0.2, −0.1, 0, 0, 0.1, 0.2, 0.3, 0.4, 0.5}.

[24] Post-experiment surveys indicate many subjects did not recognize the existence of optimal strategies.

[25] Note the proportion of optimal decisions in our declarative mechanism is consistent with proportions of optimal decisions in the second-price auction in Cooper and Fang (2008, p. 1583).

[26] Censored decisions are excluded from these statistics (see footnote 12).

Table 2 Novice responses: descriptive statistics

	Declarative	Clock
Observations	53	77
Optimal decisions	25/47%	30/39%
Non-optimal decisions	28/53%	17/22%
Censored decisions	–	30/39%
Mean of deviation from truth	0.0604[b] (0.0264)	0.0362[a] (0.0203)
Mean of absolute deviation from truth	0.1208[c] (0.0221)	0.0787[c] (0.0175)

In parentheses are standard errors. All tests are two-sided one-sample *t*-tests, [a] is 10% significance, [b] is 5%, and [c] is 1%.

Wilcoxon-Mann–Whitney test found a statistical difference at $p=0.096$ (binary data: 1 if a decision is optimal, 0 otherwise).[27]

However, is the accuracy of the clock mechanism driven by data censoring? If respondents use identical pre-determined strategies in the two mechanisms, then beliefs elicited using the clock mechanism should be identical to beliefs elicited using the declarative mechanism after applying a filter that is equivalent to clock censoring. That is, we should obtain the same data if they are "naturally" censored during the experiment in the clock mechanism or "artificially" filtered after the experiment in the declarative mechanism.

In particular, for each decision in the declarative mechanism, we randomly select a number that is equally likely to be 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, or 0.9. The decision is filtered and discarded if it is strictly greater than the random number. The sample of 53 independent decisions from the declarative mechanism is filtered 10,000 times, and the frequency of each decision remaining is calculated. Doing this yields our second result,

> **Result 2**. *With novice respondents, filtered declarative data differ significantly from clock data.*

Evidence As shown in Fig. 2, the distribution of filtered declarative data is significantly different from the distribution of clock data ($p=.014$, Chi-squared test).[28]

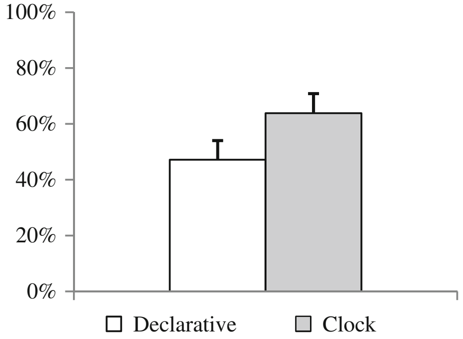
In addition, Fig. 2 plots the distribution that would emerge under only dominant strategies.[29] It has a single mode at 0.3. In comparison, beliefs elicited using the clock mechanism also have a single mode at 0.3, and seem to characterize the dominant strategy distribution reasonably well. In contrast, the mode of filtered declarative data is 0.1, far less accurate regarding the underlying beliefs.

[27] The one-sided test is significant at 5% level; we have a clear ordered hypothesis as illustrated in Section 2.

[28] The distribution of unfiltered declarative data is marginally significantly different from the distribution of the clock data ($p=0.063$, Chi-squared test).

[29] We assume the two dominant strategies are equally likely to be chosen. Hence, the ratios of subjects who choose decisions 0.2, 0.3 and 0.4 are 1:2:1.

Fig. 1 Proportion of optimal decisions in first round



Note: Error bar is one s.e. of the mean.

4.2 Responses from one-time experienced participants

We next report an analysis of decisions made in the second round. Our third result is as follows.

> **Result 3**. *With one-time experienced participants, the declarative and clock mechanisms are equally accurate.*

Evidence Figure 3 shows that proportions of optimal decisions in the declarative and clock mechanisms are 57% and 60% respectively, and these values are not significantly different (p=.84, two-sided Wilcoxon-Mann–Whitney).[30]

Similarly, we apply clock-equivalent filtering to second-round decisions from the declarative mechanism and compare them with the second-round clock data:

> **Result 4**. *With one-time experienced participants, filtered declarative data differ significantly from clock data.*

Evidence As shown in Fig. 4, the distribution of filtered declarative data is significantly different from the distribution of clock data (p=.052, Chi-squared test).[31]

Combining results 3 and 4, we observe that the proportions of truthful elicitations are not distinguishable in the second round, but the distribution of clock data continues to characterize underlying beliefs more accurately than does the distribution of filtered declarative data.

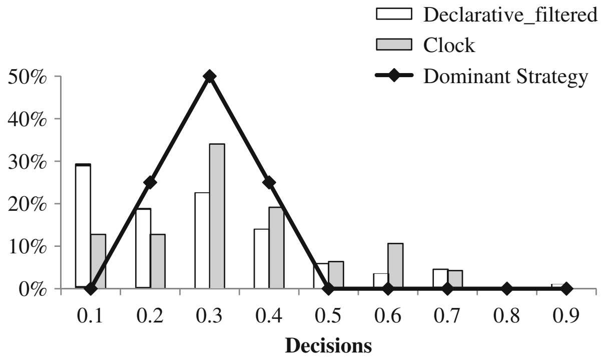
5 Implementing the mechanisms

The practical design of incentive-compatible belief elicitation mechanisms can be important, for example, to the rapidly emerging literature reporting

[30] In the declarative mechanism, deviations from dominant strategies in the second round are significantly smaller than they are in the first round (p=.04, two sided Wilcoxon-Mann–Whitney).

[31] However, the distribution of unfiltered declarative data is not different from the distribution of the clock data (p=.392, Chi-squared test).

Fig. 2 Distributions of first-round decisions



Note: Half hold beliefs equal to 0.2 and the other half hold beliefs equal to 0.3.

experiments from large-scale surveys or field experiments (see, e.g., Andersen et al. 2008; Dohmen et al. 2009; Bellemare et al. 2008). Below we briefly discuss how one can use these incentive-compatible mechanisms to elicit beliefs in the framework of the Bellemare et al. (2008) experiment. We chose this experiment for three reasons. First, this experiment uses belief elicitation instrumentally, in the sense that it is part of an investigation of broader economic questions. The mechanisms discussed in this paper may hold particular value for such studies. Second, the Bellemare et al. experiment was conducted online, an environment where extensive training may be impractical. Third, subjects were a large representative sample,[32] a group which could in principle include mixtures of naïve and sophisticated respondents.

The goal of Bellemare et al. (2008) is to structurally estimate preferences for inequity aversion. Their approach is to combine choice data from ultimatum games (Güth et al. 1982) with beliefs elicited from proposers. In the Ultimatum game, a proposer offers to a randomly matched responder a split of 10 euros. The responder either accepts or rejects the offer; in the case of rejection both players earn zero.

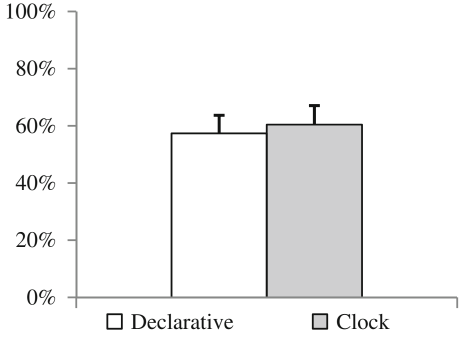
Bellemare et al. elicited proposers' beliefs by asking, "how many out of 100 people do you think would accept this offer?" Subjects were not rewarded based on the accuracy of their answers to this question.[33] To pursue an approach based on the incentive-compatible mechanisms studied in this paper, one could instead provide the following straightforward instructions to participants.

> You now can draw a ball to earn more money. If you draw a red ball you earn an additional 10 euros; otherwise you earn nothing. You will draw from either Bag A or Bag B. Each bag contains 100 balls, some red and some white. You do not know the exact number, but do know the following. The number of red balls in Bag A is equal to the number of people—out of 100 total people—who played this game and actually accepted the amount you offered to your

[32] Subjects are members of the CentERpanel, consisting of about 2,000 households, who answer questions every weekend. See www.centerdata.nl for more.

[33] In a laboratory study, Palfrey and Wang (2009) report that probabilities elicited using the linear scoring are biased towards 0 and 1 to a greater degree than with proper scoring rules. Their finding is consistent with theory predictions. On the other hand, Sonnemans and Offerman (2001) show that a flat-rate incentive does just as well as the quadratic scoring rule.

Fig. 3 Proportion of optimal decisions in second round



Note: Error bar is one s.e. of the mean.

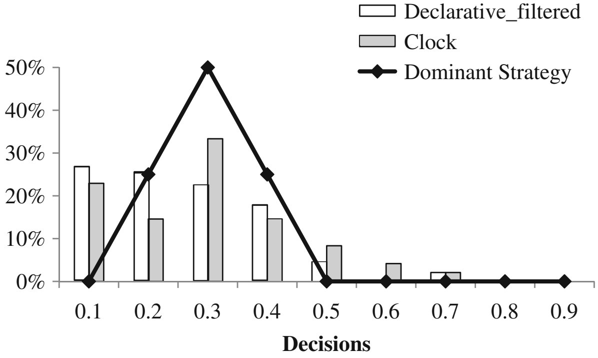
responder. The number of red balls in Bag B is between 0 and 100, all equally likely.

Declarative Mechanism: Please write down a number between 0 and 100 (any number you like). If the number you write down is smaller than the number of red balls in Bag B, then you will draw a ball from Bag B. Otherwise, you will draw a ball from Bag A.

Clock Mechanism: You will see a number on your screen that starts at 0 and counts up by one each second. The counting will stop when the number reaches its maximum, which is the number of red balls in Bag B. At any point before the counting stops, you can hit the "switch" button on the screen, in which case you will draw a ball from Bag B. If you do not "switch" before the counting stops, you will draw a ball from Bag A.

In any practical application, one must of course first decide between the two mechanisms. Our results indicate the clock holds advantages when analyzing populations that include a mixture of sophisticated and naïve participants. One can be confident in inferences from the clock mechanism in the presence of novice participants, while confidence with the declarative mechanism is greater as the fraction of experienced (sophisticated) participants increases. The reason is that the clock censors noise from naïve respondents and thus improves elicitation accuracy, while the declarative procedure makes use of all of the data. This also suggests that

Fig. 4 Distributions of second-round decisions



Note: Half hold beliefs equal to 0.2 and the other half hold beliefs equal to 0.3.

the declarative approach could be advantageous when extensive subject training on mechanism incentives is feasible, or when individual responses are desired by the investigator, such as when eliciting experts' opinions.

6 Conclusion

In a laboratory study using mixtures of naïve and sophisticated participants, we compared the declarative and clock belief elicitation mechanisms proposed by Karni (2009). These mechanisms are of interest because their incentive compatibility does not require strong assumptions such as risk neutrality or expected utility maximization. We found that, in relation to the declarative mechanism, with the clock mechanism (i) elicited beliefs are more likely to be accurate and (ii) the distribution of elicited beliefs more accurately characterizes the underlying (endowed) beliefs. Our findings complement an auction literature providing evidence that English clocks outperform second-price mechanisms in inducing truth-telling, and have implications for the practical design of incentive-compatible belief elicitation mechanisms.

Despite similarity in accuracy between the two mechanisms with experienced participants, we were surprised to find that differences in decision strategies were apparent. This finding resonates with the benefits of the clock reported in the auction literature (e.g., Kagel et al. 1987; Kagel and Levin 1993; Harstad 2000), raising the important fundamental questions of why and how a clock presentation improves decision making. Distilling the source of the clock's advantage might allow one to implement procedures to improve decision making in a wide variety of environments, even those that do not easily admit a clock representation of decision alternatives.

A limitation of our study stems from our choice of parameters. We induced beliefs that are nearer to zero than one, and we explained that doing this provides a favorable environment for the clock mechanism. While the clock may perform less well when actual beliefs are closer to the relevant upper bound (e.g., when true beliefs are above 0.5 in our experiment), this is not necessarily a problem in practice. In particular, the investigator is free to choose the clock's range and increments arbitrarily, and can always include extra ticks at larger values. In doing so, one can be more confident that actual beliefs are near the clock's starting point and thus minimize data loss.

Future research might investigate whether the truth-inducing advantage of a clock procedure persists in environments with subjective beliefs, or where incentive-compatible mechanisms are difficult to implement. This includes cases where outcomes are impractical or impossible to verify. Of particular interest here are large-scale surveys of respondents' beliefs regarding life-style choices and consequences related to, for example, changes in health or job status. In addition, recent work by Charness et al. (2007, 2010) suggests that the frequency of mistakes declines when subjects can consult with others. Hence, extending our investigation to environments where decisions are made by groups is a profitable next step to further our understanding of accurate belief elicitation.

Acknowledgements We thank Glenn Harrison, Edi Karni, Eric Danan, Jacob Sagi, Ron Harstad, James Andreoni, Yan Chen, Soo Hong Chew, Nat Wilcox, colleagues at ICES, and seminar participants at ESA international 2009 (Washington DC), Advanced Workshop in Experimental Economics 2009 (Sydney, Australia), SEA 2009 (San Antonio, TX), and FUR 2010 (Newcastle, England) for helpful discussions.

Appendix

Instructions for Declarative mechanism with endowed belief of 0.2

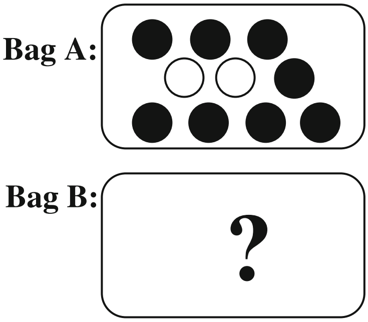
Welcome to this experiment! In addition to the $5 for showing up on time, you will be paid in cash based on your decisions in the experiment. Please note that no other participant's decisions in this experiment will affect your earnings, and vice versa. Please read these instructions carefully. Raise your hand if you have any questions, and the experimenter will come to assist you.

Overview

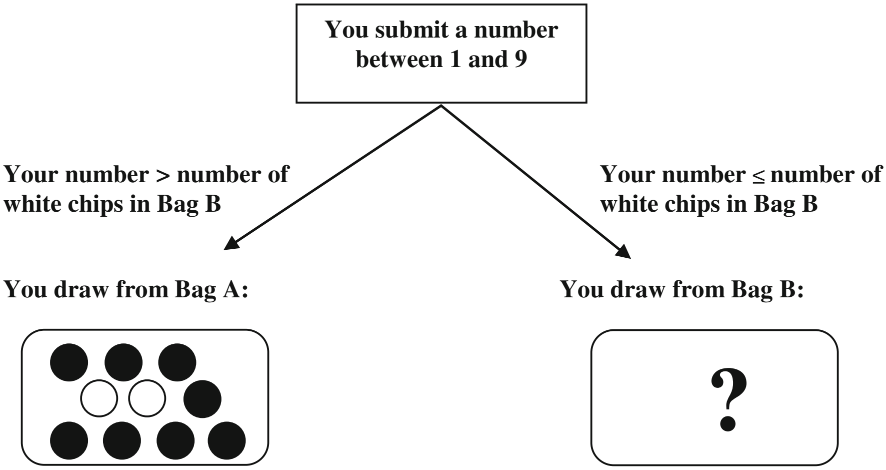
The procedure is simple. You will first submit a number, and then you will draw a chip from one of two bags. If the chip you draw is white you will earn $10, and if it is black you will earn $1.

Details

Bag A has **2** white chips and **8** black chips for a total of **10**. **Bag B** also has **10** chips, some white, some black, but you do not know how many of each. *The number of white chips in Bag B* is on the card in the sealed envelope at your desk. This card was drawn in advance from a deck of 9 cards, labeled from 1 to 9. Please do not open the envelope until you are told to do so.



To determine the bag you'll draw from, you will first submit a number between 1 and 9. If the number you submit is less than or equal to *the number of white chips in Bag B*, you will draw from **Bag B**, otherwise you will draw from **Bag A**.


You submit a number between 1 and 9
Your number > number of white chips in Bag B
Your number ≤ number of white chips in Bag B
You draw from Bag A:
You draw from Bag B:
?

Your payment: If you draw a white chip you earn \$10; a black chip earns you \$1.

Instructions for Clock mechanism with endowed belief of 0.3

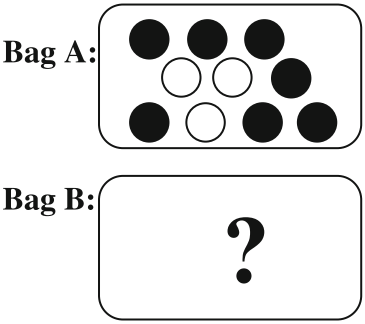
Welcome to this experiment! In addition to the \$5 for showing up on time, you will be paid in cash based on your decisions in the experiment. Please note that no other participant's decisions in this experiment will affect your earnings, and vice versa. Please read these instructions carefully. Raise your hand if you have any questions, and the experimenter will come to assist you.

Overview

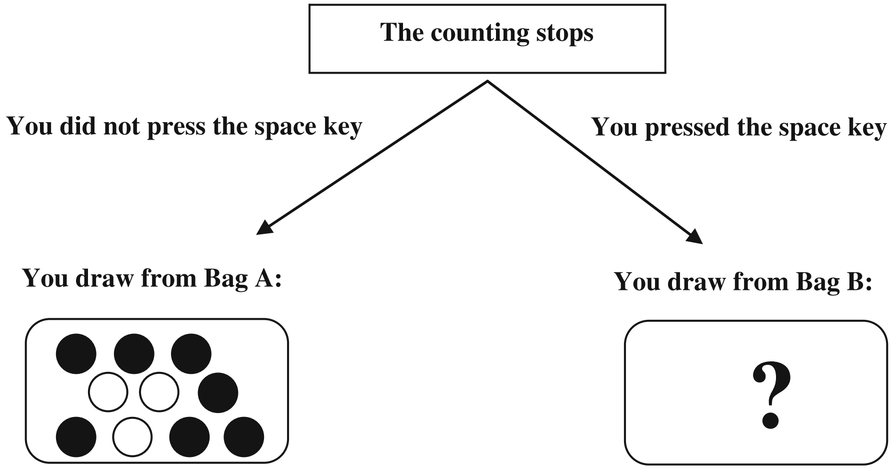
The procedure is simple. You will first participate in an exercise, and then you will draw a chip from one of two bags. If you draw a white chip you will earn \$10, and if it is black you will earn \$1.

Details

Bag A has **3** white chips and **7** black chips for a total of **10**. **Bag B** also has **10** chips, some white, some black, but you do not know how many of each. *The number of white chips in Bag B* is on the card in the sealed envelope at your desk. This card was drawn in advance from a deck of 9 cards, labeled from 1 to 9. Please do not open the envelope until you are told to do so.



To determine the bag you'll draw from, you will first participate in an exercise. The computer screen in front of you will start counting from number **1**, and increase by **1** every 5 seconds until it reaches the number in the sealed envelope. You can stop the counting at any point by pressing the space key. If you press the space key before the counting stops, you draw from **Bag B**, otherwise you draw from **Bag A**.



Your payment: If you draw a white chip you earn $10; a black chip earns you $1.

References

Allen, F. (1987). Discovering personal probabilities when utility functions are unknown. *Management Science, 33*(4), 542–544.

Andersen, S., Harrison, G. W., Lau, M. I., & Rutström, E. E. (2008). Eliciting risk and time preferences. *Econometrica, 76*(3), 583–618.

Andersen, S., Fountain, J., Harrison, G. W., & Rutström, E. E. (2010). Estimating subjective probabilities. Working paper 2010–06, Center for the Economic Analysis of Risk, Georgia State University. http://cear.gsu.edu/files/Estimating_Subjective_Probabilities.pdf.

Andreoni, J. (1995). Cooperation in public-goods experiments: kindness or confusion? *American Economic Review, 85*(4), 891–904.

Bellemare, C., Kroger, S., & Van Soest, A. (2008). Measuring inequity aversion in a heterogeneous population using experimental decisions and subjective probabilities. *Econometrica, 76*(4), 815–839.

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

Charness, G., Karni, E., & Levin, D. (2007). Individual and group decision making under risk: an experimental study of Bayesian updating and violations of first-order stochastic dominance. *Journal of Risk and Uncertainty, 35*(2), 129–148.

Charness, G., Karni, E., & Levin, D. (2010). On the conjunction fallacy in probability judgment: new experimental evidence regarding Linda. *Games and Economic Behavior, 68*(2), 551–556.

Cooper, D. J., & Fang, H. (2008). Understanding overbidding in second price auctions: an experimental study. *The Economic Journal, 118*(532), 1572–1595.

Dohmen, T., Falk, A., Huffman, D., Sunde, U., Schupp, J., & Wagner, G. G. (2009). Individual risk attitudes: measurement, determinants and behavioral consequences. *Journal of the European Economic Association, 9*(3), 522–550.

Garthwaite, P. H., Kadane, J. B., & O'Hagan, A. (2005). Statistical methods for eliciting probability distributions. *Journal of the American Statistical Association, 100*(470), 680–701.

Grether, D. M. (1992). Testing Bayes rule and the representativeness heuristic: some experimental evidence. *Journal of Economic Behavior and Organization, 17*(1), 31–57.

Güth, W., Schmittberger, R., & Schwarze, B. (1982). An experimental analysis of ultimatum bargaining. *Journal of Economic Behavior and Organization, 3*(4), 367–388.

Harstad, R. (2000). Dominant strategy adoption and bidders' experience with pricing rules. *Experimental Economics, 3*(3), 261–280.

Hoffrage, U., Lindsey, S., Hertwig, R., & Gigerenzer, G. (2000). Communicating statistical information. *Science, 290*(5500), 2261–2262.

Holt, C. A., & Smith, A. M. (2009). An update on Bayesian updating. *Journal of Economic Behavior and Organization, 69*(2), 125–134.

Hossain, T., & Okui, R. (2011). The binarized scoring rule. Working paper: http://ssrn.com/abstract=1592082

Houser, D., & Kurzban, R. (2002). Revisiting kindness and confusion in public goods experiments. *American Economic Review, 92*(4), 1062–1069.

Houser, D., Keane, M., & McCabe, K. (2004). Behavior in a dynamic decision problem: an analysis of experimental evidence using a Bayesian type classification algorithm. *Econometrica, 72*(3), 781–822.

Jaffray, J.-Y., & Karni, E. (1999). Elicitation of subjective probabilities when the initial endowment is unobservable. *Journal of Risk and Uncertainty, 18*(1), 5–20.

Kadane, J. B., & Winkler, R. L. (1988). Separating probability elicitation from utilities. *Journal of the American Statistical Association, 83*(402), 357–363.

Kagel, J., & Levin, D. (1993). Independent private value auctions: bidder behaviour in first-, second-, and third-price auctions with varying numbers of bidders. *The Economic Journal, 103*(419), 868–879.

Kagel, J., & Levin, D. (2009). Implementing efficient multi-object auction institutions: an experimental study of the performance of boundedly rational agents. *Games and Economic Behavior, 66*, 221–237.

Kagel, J., Levin, D., & Harstad, R. (1987). Information impact and allocation rules in auctions with affiliated private values: a laboratory study. *Econometrica, 55*(6), 1275–1304.

Karni, E. (1999). Elicitation of subjective probabilities when preferences are state-dependent. *International Economic Review, 40*(2), 479–486.

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

Köszegi, B., & Rabin, M. (2008). Revealed mistakes and revealed preferences. In A. Caplin & A. Schotter (Eds.), *The foundations of positive and normative economics: A handbook*. New York: Oxford University Press.

Machina, M. J., & Schmeidler, D. (1992). A more robust definition of subjective probability. *Econometrica, 60*(4), 745–780.

Manski, C. (2004). Measuring expectations. *Econometrica, 72*(5), 1329–1376.

McKelvey, R. D., & Page, T. (1990). Public and private information: an experimental study of information pooling. *Econometrica, 58*(6), 1321–1339.

Möbius, M., Niederle, M., Niehaus, P., & Rosenblat, T. (2011). Managing self-confidence: theory and experimental evidence. NBER working paper no. 17104: http://www.nber.org/papers/w17014.

Nyarko, Y., & Schotter, A. (2002). An experimental study of belief learning using elicited beliefs. *Econometrica, 70*(3), 971–1005.

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*(4), 1461–1489.

Palfrey, T. R., & Wang, S. W. (2009). On eliciting beliefs in strategic games. *Journal of Economic Behavior & Organization, 71*(2), 98–109.

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

Roth, A. E., & Murnighan, J. K. (1982). The role of information in bargaining: an experimental study. *Econometrica, 50*(5), 1123–1142.

Roth, A. E., & Schoumaker, F. (1983). Expectations and reputations in bargaining: an experimental study. *American Economic Review, 73*(3), 362–372.

Rutström, E. E. (1998). Home-grown values and incentive compatible auction design. *International Journal of Game Theory, 27*(3), 427–441.

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

Schlag, K. H., & van der Weele, J. (2009). Eliciting probabilities, means, medians, variances and covariances without assuming risk-neutrality. Working paper, Universitat Pompeu Fabra, Barcelona.

Sonnemans, J., & Offerman, T. (2001). Is the quadratic scoring rule really incentive compatible? Working Paper, CREED, University of Amsterdam.

Winkler, R. L., & Murphy, A. H. (1968). “Good” probability assessors. *Journal of Applied Meteorology, 7*(5), 751–758.