


# Experimental guidance for eliciting beliefs with the Stochastic Becker–DeGroot–Marschak mechanism

Ingrid Burfurd<sup>1</sup> · Tom Wilkening<sup>1</sup> 

Received: 1 November 2017 / Revised: 28 January 2018 / Accepted: 28 February 2018 /  
Published online: 15 March 2018  
© Economic Science Association 2018

**Abstract** We compare different implementations of the Stochastic Becker–DeGroot–Marschak (SBDM) belief elicitation mechanism, which is theoretically elegant but challenging to implement. In a first experiment, we compare three common formats of the mechanism in terms of speed and data quality. We find that all formats yield reports with similar levels of accuracy and precision, but that the instructions and reporting format adapted from Hao and Houser (J Risk Uncertain 44(2):161–180 2012) is significantly faster to implement. We use this format in a second experiment in which we vary the delivery method and quiz procedure. Dropping the pre-experiment quiz significantly compromises the accuracy of subject’s reports and leads to a dramatic spike in boundary reports. However, switching between electronic and paper-based instructions and quizzes does not affect the accuracy or precision of subjects’ reports.

**Keywords** Beliefs · Elicitation · Prediction accuracy · Methodology

**JEL Classification** C91 · D81 · D83

We thank Amy Corman, Laboratory Manager at the University of Melbourne’s Experimental Economics Lab. We gratefully acknowledge the financial support of the Australian Research Council through the Discovery Early Career Research Award DE140101014 as well as the Faculty of Business and Economics at the University of Melbourne.

**Electronic supplementary material** The online version of this article (<https://doi.org/10.1007/s40881-018-0046-5>) contains supplementary material, which is available to authorized users.

✉ Tom Wilkening  
tom.wilkening@unimelb.edu.au

Ingrid Burfurd  
ingrid.burfurd@gmail.com

<sup>1</sup> Department of Economics, The University of Melbourne, Melbourne, Australia

## 1 Introduction

Most theories of decision-making assume that choices are based on an individual's preferences and probabilistic beliefs. Economists who want to test the descriptive validity of these theories are hindered by the fact that preferences and beliefs are typically unobservable. An advantage of economic experiments over other sources of empirical data is that secondary measures such as probabilistic beliefs can be elicited. These secondary measures supplement choice data and allow for stronger identification of the forces governing the decision-making process.

A challenge faced by practitioners is that there is a potential tradeoff between practical considerations—such as speed—and data quality considerations, such as accuracy and separability between subjects' beliefs and preferences. To help practitioners assess the relative merits of different experimental techniques, we explore the practicality-quality trade-off with regard to the Stochastic Becker–DeGroot–Marschak (SBDM) belief elicitation mechanism. The SBDM mechanism has been chosen for two reasons. First, the SBDM mechanism is incentive-compatible for all subjects whose preferences respect probabilistic sophistication and dominance (Karni 2009).<sup>1</sup> These properties are desirable because heterogeneous risk preferences have been well documented in the laboratory (see, for example, Holt and Laury 2002) and there is evidence that some subjects are not well described by the expected-utility model of decision-making (Harrison and Rutström 2009). Second, the SBDM mechanism is quite complex.<sup>2</sup> This complexity has prompted practitioners to experiment with quite different instructions, reporting interfaces, and training methods. Given the current absence of standard procedures, we believe it is important to identify which format offers the best balance of practicality and data quality.

We report results from two experiments. In the first experiment, we compare three isomorphic presentations of the SBDM mechanism, which are adapted from Holt and Smith (2009), Hao and Houser (2012) and Trautmann and van de Kuilen (2015). The first format presents careful and detailed descriptive instructions, the second introduces a simple analogy to explain a complex probabilistic concept, and the third uses a list-based format for reporting beliefs. Each has desirable features. To get at the practicality-quality trade-off, we compare the formats in terms of accuracy, precision, and the time it takes for subjects to work through instructions, a quiz, and each iteration of the belief-elicitation task. We find that all formats yield reports with similar levels of accuracy and precision, but that the instructions and

---

<sup>1</sup> The term “probabilistic sophistication” is used as per Machina and Schmeidler (1992)—that is, that the subject ranks lotteries based purely on the implied probability distribution over outcomes. The practical implication is that a subject will rank bets with subjective probabilities over outcomes in the same manner as he would rank lotteries with an objective probability distribution. Epstein (1999) defines ambiguity neutrality as a decision-maker for which the probability is sophisticated. Thus, SBDM is not in general incentive compatible when decision-makers are ambiguity averse. “Dominance” is the condition that a subject has preference relation  $\succeq$  over lotteries such that  $HpL \succeq H_pL$  for all  $H > L$  if and only if  $p \geq p'$ .

<sup>2</sup> Ducharme and Donnell (1973) present the first experimental test of the mechanism and observe that while it is “basically simple”, the SBDM mechanism task “seems complicated at first exposure”.

reporting format adapted from Hao and Houser (2012) is significantly faster to implement.

In a second experiment, we restrict attention to the Hao and Houser (2012) format and run three treatments that focus on the practicalities of implementation. One treatment drops the pre-experiment quiz, one delivers the instructions and quiz on paper, and the third delivers the instructions and quiz electronically. We find that dropping the pre-experimental quiz significantly compromises the accuracy of subjects' reports and leads to a dramatic spike in boundary reports. We also find that switching between electronic and paper-based instructions and quizzes does not affect the accuracy or precision of subjects' reports.

This paper contributes to the small but growing literature on belief-elicitation methodologies. Existing work has compared the quality of reports under different belief-elicitation mechanisms, including Huck and Weizsäcker (2002), Palfrey and Wang (2009), Massoni et al. (2014), Trautmann and van de Kuilen (2015) and Hollard et al. (2016).<sup>3</sup> There has, however, been little work on the practicalities of implementation. The notable exception is Holt and Smith (2016), which is closest to our paper. Holt and Smith use a Bayesian-updating task to compare direct elicitation and a list-based format for implementing the SBDM mechanism. Our paper partially replicates their list of formats but also tests analogy-based instructions which are promising in both speed and accuracy. We also provide guidance on the importance of quizzes and instruction format when implementing the SBDM mechanism.

## 2 The Stochastic Becker–DeGroot–Marshak mechanism

The Stochastic Becker–DeGroot–Marschak mechanism is based closely on the Becker–DeGroot–Marschak (BDM) mechanism (Becker et al. 1964), which was originally conceived as a method for eliciting certainty equivalents for lotteries. In its original context, the BDM mechanism works as follows: Let  $H_pL$  denote the lottery that pays  $H$  with probability  $p$  and  $L$  otherwise. In the first stage of the mechanism, the subject is asked to report a price  $r$ , which he is prepared to pay to acquire the lottery  $H_pL$ . In stage two, a number  $z$  is realised from the distribution of random variable  $Z$ , which has distribution  $PZ$  with support  $[0, H]$ . The subject receives the outcome of lottery  $H_pL$  if  $z \leq r$  and payment  $z$  otherwise.

For all expected-utility maximizing agents, it is a dominant strategy to report one's certainty equivalent (CE). The intuition for this result is straightforward: a subject who reports  $r > CE$  runs the risk that  $CE < z < r$ . He will be paid according to the outcome of the lottery which he values at CE, but would prefer to receive payoff  $z$ . If the subject under-states their CE, with  $r < CE$ , this is also costly: if  $r < z < CE$ , the subject will receive  $z$  but would prefer to receive the lottery.

The Stochastic Becker–DeGroot–Marschak mechanism adapts this approach to elicit the probability  $p$  of a particular stochastic event  $A$ . As per the deterministic case, the subject is endowed with a lottery that pays  $H$  if event  $A$  occurs and  $L$

<sup>3</sup> For excellent reviews of belief elicitation in the lab, see also Schlag et al. (2013) and Schotter and Trevino (2014).

otherwise. Given a true belief  $p$ , this lottery corresponds to a lottery  $HpL$ . The subject reports his belief  $r$  about  $p$ . A number  $z$  is realised from the distribution of random variable  $Z$ , which has distribution  $PZ$  on support  $[0, 1]$ . If  $z \leq r$ , the subject retains his original lottery; if  $z > r$ , the agent exchanges his original lottery for a new lottery  $H_zL$ . The lottery payoffs are identical, with the two lotteries distinguished only by their probabilities of winning. Not only is this mechanism robust to heterogeneous risk preferences but also to preferences that do not conform with expected-utility maximisation. For subjects who do not have a stake in the event of interest (i.e., they have no incentive to hedge) and whose preferences are consistent with probabilistic sophistication and dominance, it is in their interest to report  $r = p$ , as they otherwise risk receiving their less-preferred lottery (Karni 2009).

## 2.1 The SBDM in practice

The SBDM mechanism is a complex procedure. Its incentive compatibility requires subjects to have a thorough understanding of the mechanism, or at least to trust a researcher who tells them that it is in their best interests to report beliefs accurately. Experimental economists have broadly taken one of three approaches when implementing the SBDM, varying in the ways they explain the SBDM and the way subjects report their beliefs.

Early implementations of the SBDM mechanism such as Ducharme and Donnell (1973) and Grether (1992) explained the SBDM mechanism rigorously and precisely, often alongside descriptions of probabilistic concepts and incentive-compatibility. They then ask subjects to report  $r$  directly—that is, to issue a numeric report about their belief. We refer to this as a “descriptive” approach to capture the faithful depiction of the underlying SBDM mechanism.

Our benchmark for the descriptive format is Holt and Smith (2009) (HS). Subjects are told that they must report their  $r$ -in-100 belief that a particular event (“Event A”) has occurred. This event is worth \$ $x$ . HS explains that belief  $r$  is equivalent to a belief that a lottery has an  $r$ -in-100 chance of winning \$ $x$ . Subjects are then introduced to a stochastic “payoff lottery”, in which the subject can win \$ $x$ . Subjects are told that the probability of winning the payoff lottery is  $t$ -in-100, with  $t$  drawn from a uniform distribution between 0 and 100. If the subject’s reported belief  $r$  is above cutoff  $t$ , the subject will be paid \$ $x$  if Event A has occurred. If  $r$  is less than or equal to  $t$ , the subject’s payoff will be determined by the payoff lottery. Both lotteries potentially pay \$ $x$ , and—according to their reported belief  $r$ —the subject will play whichever game gives him a higher probability of winning.

Möbius et al. (2007), Hollard et al. (2016) and Möbius et al. (2011) also use direct reporting, but use analogies to explain the stochastic payoff mechanism. Our “analogy-based” format is adapted from the instructions presented in Hao and Houser (2012) (HH), which use a ‘chips-in-a-bag’ analogy to explain the stochastic payoff mechanism. Subjects are asked to report a belief  $r$  about the probability of an event occurring (with the event associated with payoff \$ $x$ ). They are told that a number between 0 and 100 will be randomly selected, with each number equally

likely to be chosen. If this number “?” is larger than  $r$ , the subject’s payoff will be determined by the draw of a chip from a bag. This bag contains 100 chips: ? are black and the remainder are white. A black chip is worth \$ $x$ . Subjects are told that after they report belief  $r$  they will be paid either according to the realisation of the event or the draw of a chip from the bag—whichever has a higher payoff according to their reported belief. Hao and Houser’s subjects see a physical bag filled with chips; our chips-in-a-bag are computerised.

Trautmann and van de Kuilen (2015) and Holt and Smith (2016) move away from direct reporting and explore an alternative list-style reporting format for the SBDM mechanism. The format is similar to the lists that are common in risk and time-preference elicitation tasks: a subject is presented with a list of choice tasks in which he indicates his preference over two lotteries. In Trautmann and van de Kuilen (2015) (TK), the subject indicates whether he prefers to be paid according to “Asset A”—which makes a payment if a particular event is realised—or Option B, which offers an objective probability of winning with the outcome determined by the role of a die. Following (TK), our variant of the “list” format requires subjects to choose whether they would prefer to be paid according to the outcome of Event A, or alternatively according to the outcome of the Dice Lottery. Similar to Holt and Smith (2016), we use a two-step titration procedure. In step one, subjects nominate the support for their switch-point, with supports expressed as ranges of 10 percentage points (e.g. “51–60%”). On a second screen, subjects indicate precisely when they switch from preferring one lottery to the other. The experiment does not allow subjects to nominate more than one switch-point.<sup>4</sup>

### 3 Experiment 1

Experiment 1 was conducted at the University of Melbourne’s Experimental Economics Laboratory in July 2015 and consisted of 125 subjects. Each subject was paid a \$15 show-up fee, and won \$15 or \$0 in the experiment.<sup>5</sup> The experiment used deliberately high stakes to ensure that rewards were salient. Payment was based on one period chosen from the fifteen periods at random.

We use an “induced probability” approach in our design. Subjects are given a Bayesian-updating task and asked to report their beliefs about a posterior which has an objective probability that is known to the researcher. The task is modelled on Holt and Smith (2009). Subjects are told that there are two buckets: Buckets A and B. Bucket A contains two dark balls and one light-colored ball, while Bucket B contains two light balls and one dark ball. Subjects are informed that each bucket is equally likely to be selected, and that a ball will then be drawn from this unknown bucket. Each ball is equally likely to be chosen. Subjects are shown the color of the

<sup>4</sup> By restricting subjects to a single switch point we might prevent subjects from reporting their true preferences and/or imposing consistency when subjects are actually confused. However, as we did not allow for multiple reports in the other two mechanisms the cleanest comparison is to preserve a single switch point.

<sup>5</sup> Subjects’ total completion time for Experiment 1 varied between 16 and 58 min, and subjects received an average payoff of \$25.95.

ball and asked to nominate their belief that the ball has been drawn from Bucket A. We make minor adjustments to the instructions to accommodate our computerised format and the belief-formation task is called the “Bucket Game” for easy and consistent reference throughout the instructions.

Subjects all received identical instructions regarding the Bucket Game and the pay-one-period payment protocol. Subjects then read one of the three SBDM mechanism instructions. The HH and TK instructions are adapted to the context of the HS “Bucket Game”, and all instructions use the same language. In particular, this means that probabilities are expressed as the “chance in 100” of an event occurring. Following HS, all instructions tell subjects to “think carefully” about their beliefs because it will affect the selection of payoff method.

After reading their instructions, all subjects completed a computerised pre-participation quiz. Quiz formats for the HS and HH treatment were identical, while the TK format was slightly modified to give subjects practice in making reports via the two-step procedure.

Neither the instructions nor quiz use verbal interactions. This is to minimise experimenter effects and so that the instructions can be easily used across experiments and laboratories. We also use “portable” instructions—that is, instructions that avoid reference to the experiment itself.

Each subject completed 15 repetitions of the belief-elicitation task. At the end of each period, subjects learned whether they earned \$15 or \$0. Participants in the HS and HH treatments learned  $z$ , were reminded of their report  $r$ , and were told whether they were paid according to the Bucket Game or Dice Lottery/Lottery Bag Game. Subjects in the TK treatment were told which of their choices was randomly selected, and were reminded about their preferred payoff option. All subjects were told the outcome of the stochastic payoff lottery, or alternatively whether their ball was drawn from Bucket A or B.

Experiment 1 was conducted across eight sessions and two days, with four sessions held on each day. In each session, roughly a third of subjects participated in each treatment. Subjects drew a numbered ball from a jar and were seated at the corresponding computer station, with a third of the laboratory’s computers devoted to each treatment. A summary of treatments is shown in Table 1.

### 3.1 Outcome measures and statistical tests

We consider three outcome measures when assessing the trade-offs that exist across formats: accuracy, precision, and brevity. Our measure of **accuracy** is the mean of

**Table 1** Summary of Experiment 1

Treatment	Periods	Computerised	$n$	Instructions		Quiz
				Word count	Z-tree screens	Reporting format
HS	15	Yes	42	936	6	Direct
HH	15	Yes	41	397	2	Direct
TK	15	Yes	42	391	4	List-style

the absolute error of a subject's reports, relative to the objective Bayesian posterior. Between-treatment variations in accuracy provides an indication of the incentive-compatibility characteristics of each treatment.

As a measure of **precision**, we use the standard deviation of absolute errors for each individual.<sup>6</sup> The experiment centers around an objective Bayesian-updating task, and there is no reason to suspect that individuals should vary systematically on their understanding of this task across formats. Between-format variation in precision may thus be a sign of different degrees of learning about the incentive properties of the mechanism and would suggest differences in initial understanding.

Finally, our measure of practicality is **brevity**, and we use the total time it takes a subject to go through the entire experiment. This includes the time taken to read the instructions, complete the quiz, and answer all 15 decision problems. It does not include time taken to complete the post-experiment questionnaire.

Throughout the analysis, we perform the Kruskal–Wallis test over all three formats with each individual treated as a single observation. This test is the natural extension of the Mann–Whitney–Wilcoxon test when there are more than two treatments. The null hypothesis is that a random observation from subjects in each treatment is equally likely to be larger or smaller than an observation drawn from a different treatment. As a post hoc test, we also use Dunn's test for stochastic dominance to compare pair-wise treatments and we adjust errors using the Benjamini–Hochberg procedure to adjust for multiple hypotheses. All results in the paper have also been assessed using randomisation tests identical to those in Holt and Smith (2016). Any differences between the two approaches are noted in the main text.

## 4 Experiment 1: Results

**Result 1** The accuracy and precision of reports achieved with adaptations of the Holt and Smith (2009) format, the Hao and Houser (2012) format, and the Trautmann and van de Kuilen (2015) format are not significantly different from one another. The Hao and Houser format is significantly faster to run than the other two formats.

Support for Result 1 is provided in Table 2, which provides summary statistics for our three outcome measures and the  $p$  values from all treatment-level statistical tests. As can be seen in the first row, average accuracy in the HS, HH and TK is similar, with no apparent difference between the three formats. The Kruskal–Wallis test cannot reject the null hypothesis and there is no significant difference found in any of the pairwise tests. As can be seen in the second row, the precision of reports is similar across the three formats and there is no statistical evidence that the three formats differ at the aggregate level.

As can be seen in the third row of the table, the HH format takes subjects 850 s on average to complete, while the HS format takes 1089 s and the TK format takes 1212 s. The difference in time is significant according to the Kruskal–Wallis test.

---

<sup>6</sup> Typically precision is defined as the inverse of the variance. However, since some subjects have zero variance, this measure is unbounded.

**Table 2** Summary statistics for the HS, HH and TK treatments

	Treatment means			KW-test	Pairwise Dunn tests		
	HS	HH	TK		HSvHH	HSvTK	HHvTK
Mean Abs. error	12.4 (1.53)	13.1 (1.43)	14.7 (1.61)	0.546	0.326	0.410	0.393
Within-Sub. SD errors	9.8 (1.11)	9.9 (1.17)	9.5 (1.44)	0.821	0.445	0.482	0.824
Total time (s)	1089 (63.1)	850 (47.2)	1213 (67.7)	<b>&lt; 0.001</b>	<b>0.005</b>	0.069	<b>&lt; 0.001</b>

The Kruskal–Wallis test is performed at the measure level and the Dunn pairwise tests are adjusted for multiple hypotheses using the Benjamin–Hochberg adjustment. Standard errors are reported in parentheses

Looking at the pairwise tests, response time in the HH format is significantly different from both the HS and TK formats. There is no significant difference in time between the HS format and the TK format.

Statistical tests at the aggregate level may mask the distributional features of subjects' reports that are likely to be of concern to practitioners. For instance, in many settings, direct reports lead to groupings at round numbers—such as 10 or 20—and larger clusters at 0, 50, and 100. These groupings are likely to be obscured when averaged over multiple periods. We, therefore, examine the distribution of subjects' reports and the corresponding absolute errors.

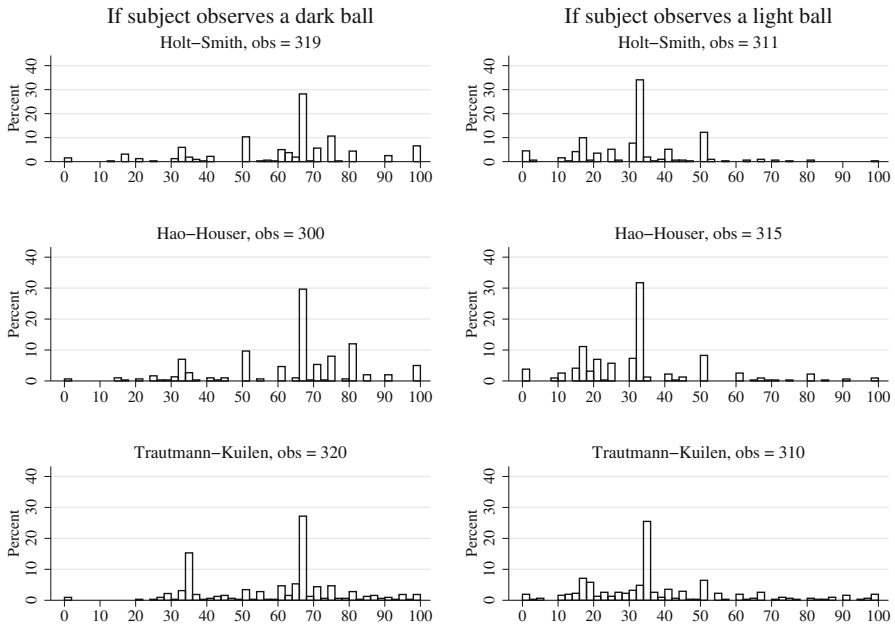
Figure 1 shows the distribution of subjects' reported beliefs. In all three treatments, there are pronounced spikes that are consistent with accurate Bayesian updating (posteriors of 67% in the wake of observing a dark ball, and 33% in the wake of a light ball). In the HS and HH treatments, there are also clusters of observations at each of the 10-point intervals nearest the true posterior and a small number of reports at 50. Boundary reports occur 5.7% of the time in the HS treatment and 4.8% of the time in the HH treatment. Clustering at 10-point intervals is less pronounced in the TK treatment and boundary reports occur in only 1.7% of cases.<sup>7</sup> However, after the observation of a dark ball, 17% of TK reports are 33—the posterior that should occur after observing a light ball. This suggests that some subjects might be losing track of the signal they have observed.<sup>8</sup>

Table 3 reports mean and median completion times for each major component of the experiment. Subjects in the HS and HH Treatments share the same quiz and period formats, and have similar mean and median completion times for these components of the experiment. Instruction times differ quite dramatically, however, with mean times of 480 (HS) versus 305 s (HH), and median times of 333 versus 288 s (Kruskal–Wallis test:  $p < 0.001$ ; Dunn test comparing HS and HH:  $p = 0.000$ ). The mean subject, therefore, takes nearly 3 min longer to work through the HS instructions than the HH instructions.

<sup>7</sup> As in Holt and Smith (2016), the difference in boundary reports is significant in a randomisation test at the 0.01 level. We note, however, that the proportion of these reports is much smaller in our sample than in theirs. This is due in part to restricting our Bayesian task to a single draw.

<sup>8</sup> Note that every screen in the TK format reminds subjects of the color of the ball they have observed. Thus, the reverse reporting is unlikely to be due to recall and is more likely due to distraction or a lack of salience.





**Fig. 1** Reported beliefs

Figure 2 presents subjects’ completion times across 15 periods. Subjects in the TK treatment exhibit greater dispersion in period completion times than their peers, particularly in early periods. Recall that these subjects have to indicate their preferences over multiple lottery choices, which is reflected in significantly longer mean and median period completion times. Focusing on subject-level mean period completion times, TK subjects have a mean period completion time of 40 s, versus 26 and 22 in HS and HH; the medians of subject-level means are 36 in TK, 22 in HS, and 17 in HH (Kruskall–Wallis test:  $p = 0.001$ ; Dunn test comparing TK and HS:  $p = 0.001$ ; Dunn test comparing TK and HH:  $p < 0.001$ ; Dunn test comparing HS and HH:  $p = 0.061$ ).

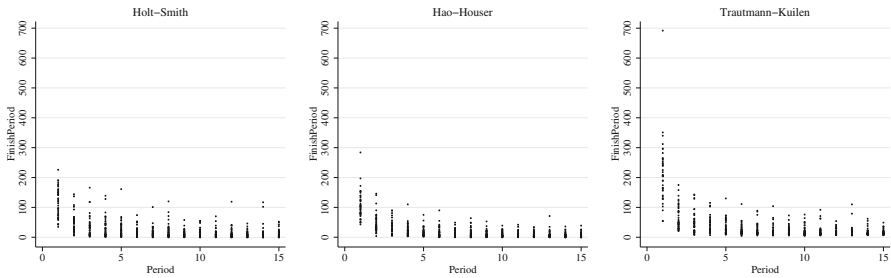
As a result of HS’s longer instructions and TK’s two-stage reporting interface, the total time taken to complete the experiment is significantly faster when subjects complete the HH treatment. The HH format, therefore, stands out as the most immediately appealing due to its improved speed and the lack of evidence that precision and accuracy are improved in either of the longer formats. We use this treatment as the basis of our second experiment, which tests whether the format of quizzes influences performance and speed.

## 5 Experiment 2

Part 2 of our study varies the implementation of the instructions adapted from Hao and Houser (2012). The Hao–Houser quiz treatment (abbreviated to Q) is identical to the Hao and Houser treatment from Experiment 1. The Hao–Houser no-quiz

**Table 3** Summary of completion times (s)

	Mean completion time			Median completion time		
	HS	HH	TK	HS	HH	TK
Period (mean)	26	22	40	22	17	36
Instructions	480	306	419	433	288	373
Quiz	212	212	187	160	178	155
Total time	1089	850	1212	988	815	1112

**Fig. 2** Completion times by period

treatment (abbreviated to NQ) drops the computerised quiz, and the paper treatment (P) administers the instructions and quiz in hard copy. The Q, NQ and P treatments are compared using the same criteria Experiment 1: accuracy, precision and brevity (Table 4).

Experiment 2 was conducted across 3 days. Three sessions were held on the first day, and one on each of the two subsequent days. Times were varied across the mornings and afternoons. The quiz- and no quiz treatments were both computerised and run jointly across three sessions with random assignment within each session.<sup>9</sup> Because of the need to distribute hard copies, the paper treatment was conducted in separate sessions so that subjects were not concerned that some participants might be completing different experiments.<sup>10</sup> Subjects' total completion time for Experiment 2 varied between 14 and 54 min, and subjects received an average payoff of \$23.55.

## 5.1 Results

**Result 2** Reports in the computerized quiz treatment are significantly more accurate than reports in the no-quiz treatment. Thus, using a quiz is important for

<sup>9</sup> The Q treatment is identical to the HH treatment and was repeated to allow for within-session randomization. Completion times in the Q treatment were slightly faster than the HH treatment with a mean session time of 766 s and a median session time of 695 s. However, the difference in session times is not significant using a Mann–Whitney–Wilcoxon test ( $p$  value = 0.13).

<sup>10</sup> Times for all treatments are measured precisely, with the exception of the paper treatment. When running the paper treatment, the laboratory assistant noted the times at which instructions were distributed, the time at which instructions were swapped for the quiz, and the time when the subject completed the quiz successfully. These times were noted in minutes rather than seconds, with all time-based analysis using the mid-point of the minute in question. There was 1 lab assistant and 15 subjects in each P treatment.

**Table 4** Summary of Experiment 2

Treatment	Periods	Computerised	<i>n</i>	Quiz Reporting format
Quiz	15	Yes	28	Direct
No Quiz	15	Yes	30	No Quiz
Paper	15	No	30	Direct

ensuring accuracy in the computerised analogy-based Hao and Houser format. There are no significant differences in the accuracy of reports in the computerized and paper-based quiz treatments, but the no-quiz treatment is significantly faster than both electronic treatments.

Table 5 reports accuracy, precision, and brevity for each of the three quiz treatments. Average accuracy in the quiz treatment is 11.2 and it is 13.4 in the paper treatment, and this difference is not significant. Accuracy in the no-quiz treatment is 20.4, which is significantly different from the quiz treatment at the 5% level ( $p = 0.04$ ) and from the paper treatment at the 10% level ( $p = 0.06$ ). The difference between the paper and no quiz treatments is significantly different at the 5% level when using the alternative randomization test.<sup>11</sup>

Precision in the no-quiz treatment is 15.2, while it is 10.4 and 8.6 in the quiz and paper treatments. The three-way Kruskal–Wallis test is not significant, but we note that a pairwise randomization test finds that the difference between the paper and no-quiz treatments is significant at the 0.05 level. We interpret this difference to be due to differences in initial understanding and learning: in the no quiz treatment, a large portion of individuals begin by making boundary reports and then revising their actions towards the objective probabilities. No such learning dynamic is observed in the other treatments.

As can be seen in the third row, the quiz increases the overall time of the experiment from an average of 613.7–766.3 s. Moving from an electronic quiz to a paper-based quiz increases the total time of the experiment to 1079.7 s. All differences are significant.

Figure 3 presents the aggregate distribution of subject's reports for the quiz, no-quiz-, and paper treatments. Accurate reports are much more common in the quiz treatment, while reports of 50 are more common in the paper-based treatment than the computerised quiz treatment. Reports in the no-quiz treatment are frequently inconsistent with Bayesian updating: while boundary reports are uncommon in the quiz and paper treatments, they occur 118 times in the no quiz treatment and account for 26.22% of observations (Table 6).

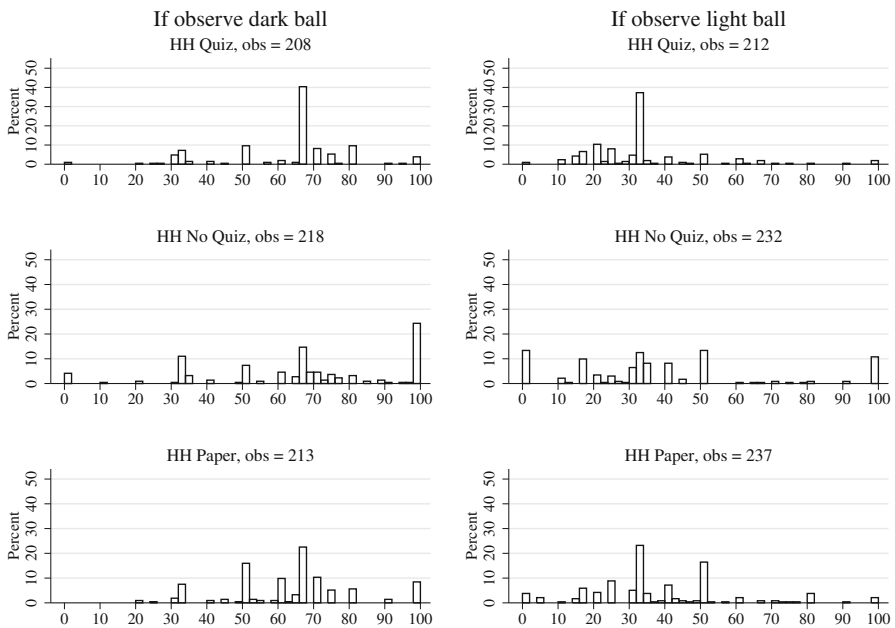
As seen in Table 3, subjects' mean period completion times do not differ significantly across the three treatments. This is not unexpectedly given that all treatments use the same computerised reporting interface. Subjects take significantly longer to complete the instructions if they participate in the paper treatment: the mean completion time is nearly 8 min, in contrast with about 5 min for the computerised instructions (quiz and no quiz treatments).

<sup>11</sup> All randomisation test results are included in the Appendix.

**Table 5** Summary statistics for the quiz (Q), no-quiz (NQ) and paper (P) treatments

	Treatment means			KW-test	Pairwise Dunn tests		
	Q	NQ	P		QvNQ	QvP	NQvP
Mean Abs. error	11.2 (1.66)	20.4 (2.78)	13.4 (1.81)	<b>0.042</b>	<b>0.017</b>	0.230	0.062
Within-subject SD of errors	10.4 (1.54)	15.2 (2.35)	8.6 (1.52)	0.158	0.238	0.187	0.082
Total time (s)	766.3 (62.5)	613.7 (37.7)	1079.7 (86.4)	<b>&lt; 0.001</b>	<b>0.035</b>	<b>0.004</b>	<b>&lt; 0.001</b>

The Kruskal–Wallis test is performed at the measure level and the Dunn pairwise tests adjusted for multiple hypotheses using the Benjamin–Hochberg adjustment. Standard errors are reported in parentheses



**Fig. 3** Reported Beliefs

**Table 6** Summary of completion times (s)

	Mean completion time			Median completion time		
	Q	NQ	P	Q	NQ	P
Period (mean)	18.6	20.5	20.4	15.2	17.5	19.6
Instructions	295	306	474	245	281	420
Quiz	191	0	300	157	0	240
Total time	766	613	1079	695	596	951

## 6 Conclusion

While belief elicitation is increasingly popular, there are no widely adopted or standard procedures. To help practitioners assess the relative merits of different experimental techniques, we explore the practicality-quality trade-off with regard to the SBDM belief-elicitation mechanism. We study behavior in three formats of the SBDM: a “descriptive” instruction format with direct reporting, adapted from Holt and Smith (2009); an “analogy-based” instruction format with direct reporting, adapted from Hao and Houser (2012); and a “list-style” format adapted from Trautmann and van de Kuilen (2015). We find that accuracy and precision of reports in the three formats are remarkably similar but that the format adapted from Hao and Houser (2012) is quicker to run than the other two formats. We use this format in a second experiment in which we vary the delivery method and quiz procedure. Dropping the pre-experiment quiz significantly compromises the accuracy of subjects’ reports and leads to a dramatic spike in boundary reports. Switching between electronic and paper-based instructions, however, does not affect the accuracy or precision of subjects’ reports.

Brevity and efficient communication in experiment instructions tends to be under-valued, and should be taken more seriously given the limited attention span of subjects. Our HH format is the shortest, yet it helps subjects make quick decisions without compromising accuracy. We thus view it as a promising format for eliciting beliefs.

Recent work by Holt and Smith (2016) also compares direct-elicitation and list-based formats of the SBDM. As with our experiment, they do not find significant differences in accuracy between formats. However, there is evidence that there is a difference in accuracy for simple situations when the true probability is .5. In the online appendix, we report on an additional robustness experiment, where we use a spinner task in which objective probabilities are easily calculated. We do not find differences in accuracy across formats using this alternative simple task.

Holt and Smith (2016) also find large differences in boundary reports across formats. When restricting the data in Holt and Smith (2016) to events that match ours, the overall boundary rate is only 2.8%, which is very similar to ours. The large difference in boundary reports in other segments of their data suggest that there may be an interaction between the complexity of the task and the importance of the elicitation format. In particular, it would be interesting to understand how subjects use the elicitation format to scaffold their probabilistic reasoning.

## Appendix A: Randomisation test results

Table 7 reports the results of pairwise randomization tests which compare outcomes from treatments in Experiments 1 and 2. All randomization tests are based on 500,000 simulations for comparability with the randomization test results reported in Holt and Smith (2016).

**Table 7** Results of pairwise randomisation tests comparing treatments

	Experiment 1			Experiment 2		
	HS-HH	HS-TK	HH-TK	Q-NQ	Q-P	NQ-P
Subject Abs. error.	0.735	0.299	0.457	0.007	0.356	0.042
Within-subject SD errors	0.949	0.846	0.806	0.105	0.395	0.022
Total time (s)	0.003	0.187	0.000	0.034	0.005	0.000

## References

- Becker, G. M., DeGroot, M. H., & Marschak, J. (1964). Measuring utility by a single-response sequential method. *Behavioral Science*, 9(3), 226–232.
- Ducharme, W. M., & Donnell, M. L. (1973). Intrasubject comparison of four response modes for “subjective probability” assessment. *Organizational Behavior and Human Performance*, 10(1), 108–117.
- Epstein, L. (1999). A definition of uncertainty aversion. *The Review of Economic Studies*, 66(3), 579–609.
- Grether, D. M. (1992). Testing bayes rule and the representativeness heuristic: Some experimental evidence. *Journal of Economic Behavior and Organization*, 17(1), 31–57.
- Hao, L., & Houser, D. (2012). Belief elicitation in the presence of naïve respondents: An experimental study. *Journal of Risk and Uncertainty*, 44(2), 161–180.
- Harrison, G. W., & Rutström, E. E. (2009). Expected utility theory and prospect theory: One wedding and a decent funeral. *Experimental Economics*, 12(2), 133–158.
- Hollard, G., Massoni, S., & Vergnaud, J.-C. (2016). In search of good probability assessors: An experimental comparison of elicitation rules for confidence judgments. *Theory and Decision*, 80(3), 363–387.
- Holt, C. A., & Laury, S. K. (2002). Risk aversion and incentive effects. *American Economic Review*, 92(5), 1644–1655.
- Holt, C. A., & Smith, A. M. (2009). An update on bayesian updating. *Journal of Economic Behavior and Organization*, 69(2), 125–134.
- Holt, C. A., & Smith, A. M. (2016). Belief elicitation with a synchronized lottery choice menu that is invariant to risk attitudes. *American Economic Journal Microeconomics*, 8(1), 110–39.
- Huck, S., & Weizsäcker, G. (2002). Do players correctly estimate what others do? Evidence of conservatism in beliefs. *Journal of Economic Behavior and Organization*, 47, 71–85.
- Karni, E. (2009). A theory of medical decision making under uncertainty. *Journal of Risk and Uncertainty*, 39(1), 1–16.
- Machina, M. J., & Schmeidler, D. (1992). A more robust definition of subjective probability. *Econometrica*, 60(4), 745–780.
- Massoni, S., Gajdos, T., & Vergnaud, J.-C. (2014). Confidence measurement in the light of signal detection theory. *Frontiers in Psychology*, 1455(5), 1–13.
- Möbius, M. M., Niederle, M., Niehaus, P., & Rosenblat, T. (2007). Gender differences in incorporating performance feedback. *draft, February*.
- Möbius, M. M., Niederle, M., Niehaus, P., & Rosenblat, T. S. (2011). *Managing self-confidence: Theory and experimental evidence*. Technical report, National Bureau of Economic Research.
- Palfrey, T., & Wang, S. (2009). On eliciting beliefs in strategic games. *Journal of Economic Behavior and Organization*, 71, 98–109.
- Schlag, K. H., Tremewan, J., & Van der Weele, J. J. (2013). A penny for your thoughts: A survey of methods for eliciting beliefs. *Experimental Economics*, 18(3), 1–34.
- Schotter, A., & Trevino, I. (2014). Belief elicitation in the laboratory. *Annual Review of Economics*, 6(1), 103–128.
- Trautmann, S. T., & van de Kuilen, G. (2015). Belief elicitation: A horse race among truth serums. *The Economic Journal*, 125, 2116–2135.