Contents lists available at ScienceDirect

Journal of Economic Behavior & Organization

journal homepage: www.elsevier.com/locate/econbase

An update on Bayesian updating

Charles A. Holt[a,1], Angela M. Smith[b,*]

[a] Department of Economics, PO Box 400182, University of Virginia, Charlottesville, VA 22904-4182, United States
[b] Department of Economics, MSC 0204, James Madison University, 800 S. Main Street, Harrisonburg, VA 22807, United States

ARTICLE INFO

Article history:
Received 9 January 2007
Accepted 31 August 2007
Available online 20 September 2008

JEL classification:
C91
D83

Keywords:
Laboratory experiments
Baye's rule
Probability weighting

ABSTRACT

This paper reports an experiment in which subjects are asked to assess probabilities for unknown events, with treatments that vary the extremity of the prior information. Probabilities are elicited using a Becker–DeGroot–Marshak procedure that does not depend on assumptions about risk aversion. The focus is on the pattern of biases in information processing.

1. Introduction

Bayes' rule is the basic normative model of information processing that is used in economic analysis. The probability perceptions of human subjects in laboratory situations are known to deviate from Bayesian predictions in systematic ways. Laboratory experiments can be used to provide a picture of how well the Bayesian model predicts in standard (non-extreme) situations and to document the nature of biases in extreme situations (e.g. with very low-probability priors for some events). The hope is that the accumulated data from such experiments will lead to an improved model with predictions that are both tractable and behaviorally relevant.

Bayes' rule provides a way of combining prior information with new information that arises from some observed signal or test. The mathematical formula for this rule is familiar to economists and other social scientists, but it is useful to review the intuition with a simple example. Suppose that a particular medical condition is present in 25 percent of some target population and that a test always reveals the condition if is exists, but that the test will generate a "false positive" one-third of the time. If this test were administered to 100 randomly selected people, the "base rate" of 25 percent would lead one to expect that 25 of them actually have the disease and would receive a positive test result. In addition, the 1/3 false positive rate would lead one to expect that 25 out of the 75 remaining people would also receive positive test results. Thus with 25 true positives and 25 false positives, the probability of having the disease given the positive test result is $25/(25+25)=0.5$. The focus can be switched from frequencies to (less intuitive) probabilities by dividing numerator and denominator by 100, so that the new numerator (0.25) is the probability of getting a positive from someone with disease D, $\Pr(\text{positive}|D)\Pr(D)$, and the

* Corresponding author. Tel.: +1 540 568 3218; fax: +1 540 568 3010.
E-mail addresses: cah2k@virginia.edu (C.A. Holt), smith9am@jmu.edu (A.M. Smith).

[1] Tel.: +1 434 924 7894; fax: +1 434 982 2904.

0167-2681/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.jebo.2007.08.013

denominator is the overall probability of getting a positive from either type: Pr(positive|D)Pr(D) + Pr(positive|∼D)Pr(∼D).[2] Hammerton (1973) used the medical context to present a similar problem to groups of British housewives. The actual probability of the disease, conditional on a positive test result, was 0.5, but the typical assessed probability was above 0.8, possibly indicating an underweighting of the prior "base rate" information.

Kahneman and Tversky (1973) report a classic experiment with even more dramatic results. Subjects were given brief descriptions of individuals who were either lawyers or engineers. One group of subjects was told that the individuals had been selected randomly from a population that contained 70 percent lawyers and 30 percent engineers, and another group was told that the descriptions had been selected from a population that contained 30 percent lawyers and 70 percent engineers. Subjects were asked to assess the chances out of 100 that each person being described was a lawyer. Some descriptions revealed clear indications of the person's profession, but other descriptions were neutral ("he is highly motivated" or "he will be successful"). The modal response in these neutral cases was near one half, regardless of whether the population had more than twice as many lawyers as engineers, or vice versa. This tendency to ignore prior information is known as "base rate bias."

Grether (1978) noted that the descriptions in this experiment were hypothetical and wondered whether the subjects truly believed the stated population proportions. He also noted that the nature of the monetary payoffs was unclear since subjects were told that they would be paid a bonus to the extent that their estimates were close to those submitted by an "expert panel." Grether's experiment was in the tradition of the "book bags and poker chips," where subjects could observe the random draws directly. In addition, he provided a financial incentive by asking subjects which event was more likely and paying them a monetary reward of $15 if they guessed the actual event or container from which draws were made, versus a $5 payment for an incorrect guess. Grether was interested in a "representativeness bias" that arises when the sample draws look like the overall population. For example, suppose that three draws are made with replacement from either cup *A*, with two red marbles and one blue, or cup *B*, with one red and two blues. If the draws are two reds and a blue, this sample is representative of cup *A*, even though it would be more likely to have come from cup *B* if the prior on cup *B* were high enough. Grether cleverly varied the prior probabilities so that Bayes' rule and representativeness indicated the same guess in some cases, and a different guess in other cases. Observed guesses were consistent with Bayes' rule about 80 percent of the time when representativeness would suggest the same answer, but this rate fell to about 60 percent when representativeness suggested a different answer.

The "book bag and poker chip" approach may seem abstract, and the alternative of providing context has been shown to help subjects process complex information in other logic-based inference tasks. Providing context, however, carries its own risks in this type of experiment. In Hammerton's (1973) experiment where the Bayes' probability of an event was 0.5, the average probability assessment of about 0.8 in the context of a test giving information about a disease fell to about 0.6 when the same problem was presented in terms of information being provided by an auto mechanic about a possible mechanical defect in a car. The context provided in these problems seems to have had an effect on the subjects' perceptions of the reliability of the test.[3] Since the goal of the experiment reported in this paper was not to study context effects, the context was presented in a neutral manner.

In order to obtain a richer set of choices with which to identify biases, subjects in this study were asked to assess probabilities, in terms of the chances out of 100, as was done in the Kahneman and Tversky (1973) and Hammerton (1973) experiments. This required a modification of Grether's payoff procedure based on an event guess. One approach would be to use a "scoring rule" to assess probabilities. For example, a quadratic scoring rule lets the subject report a probability, and the payoff is a quadratic function of the reported probability and of the actual event. The rule, together with the participant's subjective beliefs, determine an expected payoff function that is maximized by a reported probability that equals the person's subjective probability (see Davis and Holt, 1993, Chapter 8, for examples). This approach has been used effectively in some cases (Offerman and Sonnemans, 2004; Offerman et al., 2007), but a possible problem is that it relies on an assumption that the subject is risk neutral (Holt, 1986). This risk-neutrality assumption is inconsistent with laboratory evidence, even with low payoffs (Holt and Laury, 2002). The next section describes an elicitation procedure based on the Becker–DeGroot–Marshak (1964) procedure for the truthful elicitation of valuations, which has been used by Grether (1992).[4]

2. Elicitation

The elicitation procedure that was used in these experiments is based on the intuition that if you want someone else to make a choice for you, then you should tell them your true preferences so that they can make the best choice. The setup in the experiments involves two "cups," *A* and *B*, with different proportions of colored marbles. For example, in the baseline treatment, one cup contained two marbles of one color and one marble of the other color, with the proportions exactly reversed for the other cup. Each round of the experiment begins with one of the cups being selected randomly according to

[2] See Gigerenzer and Hoffrage (1995), Anderson and Holt (1996), and Holt (2006) for tips on how to teach Bayes' rule using intuitive frequency-based arguments.

[3] We are indebted to David Grether for suggesting the Hammerton reference.

[4] This approach was first suggested to Holt by the late Morris DeGroot, an accomplished Bayesian statistician, who was his thesis advisor at Carnegie-Mellon.

a pre-announced probability, which determines the prior probabilities Pr(A) and Pr(B) = 1 − Pr(A). Each subject sees a series of draws, *with replacement*, from the selected cup; these draws are used to assess the probability that cup A is being used for the draws. For notational simplicity, let the individual's subjective probability of cup A be denoted by P. Then the subject is asked to report the "chances out of 100" that cup A is being used. The answer determines a reported probability, which will be denoted by R.

The elicitation mechanism is designed so that the optimal report is truthful in the sense that $R = P$. The incentives are based on a comparison of "cup A lottery," which pays a fixed prize amount \$$V$ if cup A is in fact being used, with an "N lottery" that pays the same amount (\$$V$) with probability P_N and \$0 otherwise. Since the payoffs are the same, the N lottery would be preferred to the cup A lottery if it yields a higher probability of the payoff, regardless of risk attitudes, as long as the subject's preferences satisfy a stochastic dominance assumption. The actual value of P_N is determined randomly after the subject makes a report of R for the probability of cup A; then, the experimenter makes the choice that is best for the subject, based on the reported probability. Thus the N lottery would be used to determine the person's payoffs if $P_N \geq R$, and the cup A lottery would be used otherwise.

The actual procedures used were not couched in probability terms, but rather, the subject was asked to report the "chances out of 100" that the A cup was being used, with a report of 0 indicating no chance, a report of 50 indicating an equal chance, and a report of 100 indicating a certainty that cup A is used. Similarly, the N lottery was determined by a random cutoff number (N) with equal chances of being 0, 1, 2,. . ., 99. If the N lottery was used, a second random draw was made and the prize payoff was obtained if the second draw was strictly less than the cutoff. The second draw was uniform on the 100 integers between 0 and 99, so that the prize probability for the N lottery is $P_N = N/100$. The N lottery was used if N turned out to be greater than or equal to the reported chances out of 100 that the draw(s) were coming from the A cup. It is straightforward to show that the chances of obtaining the \$$V$ payoff are maximized by making a report that equals one's subjective probability, so anyone whose preferences satisfy stochastic dominance should report truthfully. The result is intuitive since the only way that the experimenter can choose the option with the highest chance of a \$$V$ payoff is for the subject to report a probability that represents their true subjective beliefs. This fact was explicitly noted in the instructions, and subjects did not seem to have trouble grasping the idea, especially after several decision rounds.

The Becker–DeGroot–Marshak procedure is intuitive but complicated to implement. Therefore, we provided practice to help subjects understand the procedures. In the "pencil and paper" experiment reported in the next section, there were several practice decision rounds in which the subjects became familiar with the throwing of dice, draws from the relevant cup, and the calculation of earnings. Similarly, actual cups, dice, and sample draws were used after the reading of instructions for the web-based experiment reported in Section 4 so that subjects would have a clear mental model of the physical process being implemented with the virtual draws of dice, and so on. Participants did have some questions, but the answers always ended up stressing why it is best for people to report their true "chances out of 100" that cup A was being used.[5]

3. Experiment I: playing with dice and marbles

Twenty-two subjects were recruited from economics classes at the University of Virginia in several groups with an announcement that promised a payment of \$6 plus all earnings obtained. Draws were made with replacement from either cup A, with two light marbles (L) and one dark marble (D), or from cup B with two darks and one light.

$$\text{Cup}\,A : L, L, D, \qquad \text{Cup}\,B : L, D, D.$$

With a prior of 1/2, the posterior probability for cup A after a draw of L would be: $\Pr(A|L) = 0.67$, since ex ante all 3 light marbles are equally likely to be drawn and two-thirds of them are in cup A. Bayes' rule can be used to calculate the posterior probabilities for other draw sequences in a straightforward manner. In particular, an imbalance caused by draws of L, D, and L would also imply a posterior of 2/3 for cup A, with the intuitive reason being that the first two draws cancel each other out, returning the prior to 1/2, and then the third draw of L moves the posterior to 2/3.

In each session, one person was chosen at random to be paid a flat fee and serve as a "monitor" to select the relevant cup with the throw of a die and to observe the draws. The contents of the selected "cup" were emptied into a single container that subjects could see so that any scratches on that container would not convey information about whether the contents were from cup A or cup B. Then the experimenter would go from one person to the next and make the draw(s) with replacement, shaking the container between draws. The experimenter would watch to be sure that the subject recorded the draw correctly on their decision sheet as L or D. After subjects filled in the chances out of 100 for cup A, the experimenter returned to each person's position to throw a 10-sided die twice to determine N. The first throw served as the tens digit while the second throw represented the ones digit so that each number between 0 and 99 was equally likely. If N was greater than or equal to the reported chances in 100 for that subject, then the N lottery was used and a second random number was generated

[5] One perspective on the amount of residual confusion can be gleaned from the proportion of subjects who gave correct assessments in the initial decision rounds of the web-based experiment when no sample draws were available, when no rounds with actual draws had been done previously, and when the prior was not 1/2 (since confusion in this case may still lead to a correct report). When the prior was 1/3, we coded a report as correct if it was between 30 and 35 (inclusive) and when the prior was 2/3, we coded a report as correct if it was between 65 and 70 (inclusive). Exactly three-fourths of the reports given in rounds characterized by the conditions noted were within these intervals.

Table 1
Numbers of draws in the short treatment (5 subjects in this order, and 5 subjects with sequences II and III reversed).

I. Pr(A) = 1/2	0	0	0	1	2	3	1	2	3
II. Pr(A) = 2/3	0	0	0	1	2	3	1	2	3
III. Pr(A) = 1/2	0	0	0	1	2	3	1	2	3

Table 2
Numbers of draws in the extended treatment (6 subjects in this order, and 6 subjects with sequences II and III reversed).

I. Pr(A) = 1/2	0	0	0	1	2	3	1	2	3	–
II. Pr(A) = 2/3	0	0	1	2	3	4	1	2	3	4
III. Pr(A) = 1/2	0	0	1	2	3	4	1	2	3	4

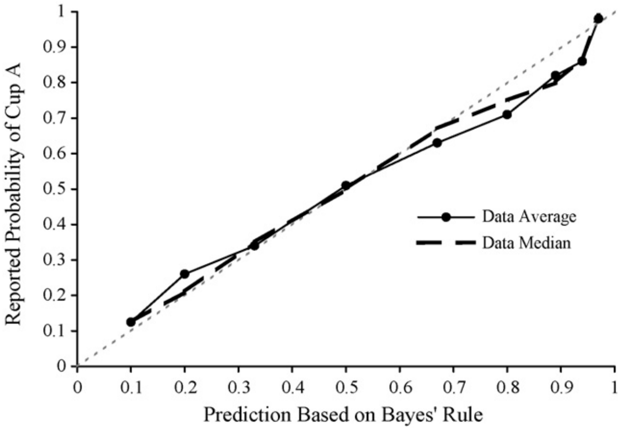

Fig. 1. Predictions versus average and median elicited probabilities for 22 subjects.

(with a draw from a bucket with 100 marked ping pong balls) to determine whether the N lottery resulted in a payment of $1. If N was less than the reported chances in 100, then the subject received a payment of $1 if cup A was used, which was announced at the end of the round. Subjects filled in their own earnings, which were checked to ensure that they understood the procedure. Instructions are provided in Appendix A (appendices available on the JEBO website).

Ten of the subjects made 27 decisions in three sequences, each of which began with three rounds where no draws were made, so the only available information was the prior information about the die throw that determined the cup, A or B. These no-draw rounds were followed by rounds with 1, 2, and 3 draws, as shown in each row of Table 1. Notice that the prior probability of cup A was raised to 2/3 in the second sequence (middle row) and then reduced to 1/2 in the final sequence (bottom row). For half of the subjects, the order of the second and third sequences was reversed.

In addition, another 12 subjects made decisions for the three sequences shown in Table 2 with only two no-draw rounds in the final two sequences, which allowed time for rounds with 4 draws, as shown. Again, sequences II and III were reversed for half of the subjects. All draws and elicited probabilities are reported in Appendix B.

Fig. 1 shows a plot of the reported probabilities as a function of the Bayes' prediction for all 22 subjects in all rounds of the final two sequences (a total of 420 decisions). Overall, the reported probabilities are quite close to Bayesian predictions, with a slight upward bias when the Bayes' prediction is low and a downward bias when the Bayes prediction is high. These biases are stronger for means than for medians since means are affected by large deviations, which are more likely to be on the up side when the Bayes' prediction is low and on the down side when the Bayes' prediction is high.

Consistent with Grether's earlier results, there is some evidence of "representativeness bias" that would cause the reported probability for cup A to be higher when a sample of three draws matches the proportions of cup A. For example, in the symmetric prior rounds, a posterior of 2/3 follows a draw of L or a representative draw sequence of L, L, and D. The average reported probability for cup A was 0.61 in the former case and 0.66 in the latter case. Here, representativeness raises the assessed probability in a manner that cancels the downward bias when the Bayes' prediction is high. Similarly, with a single D draw, the assessed probability was 0.42, clearly above the Bayes prediction of 0.33 for the symmetric prior treatment, but a draw sequence of two darks and one light that is representative of cup B lowered the average elicited probability for cup A to 0.31. The next two sections report experiments with a wider range of prior probabilities, which can be used to generate clearer bias patterns.

4. A web-based experiment with a between-subjects design

In addition to the hand-run experiment, we also ran computerized sessions of this experiment using the Bayes' Rule program on the Decision Menu of the Veconlab Experiment Selection webpage (http://veconlab.econ.virginia.edu/admin.php).

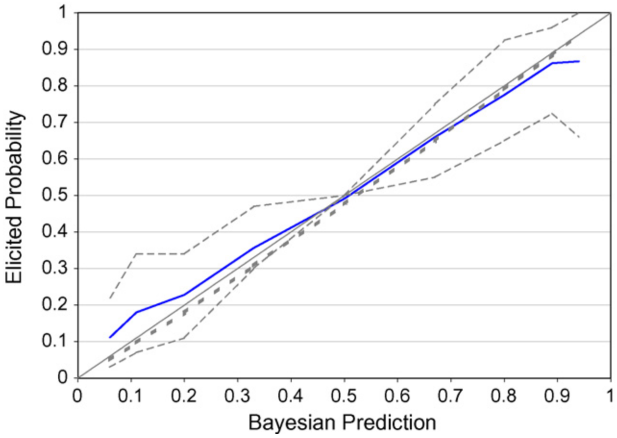

Fig. 2. Average and bounds on 2/3 of the elicited probabilities (solid line and thin dashed lines) and fitted values based on estimation (thick dotted line) for the treatment with a prior of 1/2.

In these sessions, the information was still presented in terms of a (simulated) die throw to select the relevant cup and simulated random draws from that cup (with replacement).[6] The results of the random draws were presented simultaneously, which would tend to mitigate the tendency for subjects to overweight recent observations (Grether, 1992). The computer interface reminded subjects of the payoff procedure at various points. For example, the message on their Results screen was

> The first throw of the dice yielded a value **N = 28** so the "N-lottery" offers 28 chances in 100 of obtaining the high payoff. You reported a belief that there are **50** chances in 100 that the cup used is the red cup, and since this number is at least as large as N, we will let your payoff determined by whether or not the red cup was used.

The web-based experiment was set up with a between-subjects design. Each subject made 30 decisions under the same prior probability for cup *A*, which was 1/3 for 24 subjects, 1/2 for 24 subjects, and 2/3 for 24 subjects. There were 3 sequences of 10 decisions, and in each sequence, the numbers of draws were 0, 0, 1, 2, 3, 4, 1, 2, 3, 4.

Fig. 2 shows a smoothed line of the average elicited probabilities that correspond to each possible Bayesian prediction generated by sequences of draws in the symmetric treatment with a prior of 1/2. The thin dashed lines bound the middle two-thirds of choices such that only 1/6 of the distribution of choices lies above the upper dashed line and 1/6 of the distribution lies below the lower dashed line. In the aggregate, decisions are well approximated by Bayes' rule, with less variability for the balanced cases in the center. As with the earlier hand-run experiment, there is a slight tendency for average predictions to be too high on the left and too low on the right, which is consistent with the asymmetry of the spread represented by the dashed lines.

Prediction behavior was a little noisier in the asymmetric treatments, shown in Figs. 3 and 4. There is a tendency to predict too high for low probabilities when the prior is 1/3, but this is not the case when the prior is 2/3 where predictions tend to be low across the board.

The downward bias in predictions when the prior is 2/3 suggests that prior information is not being fully incorporated, a conjecture that we will address in the subsequent estimation. As a quick check, we calculated what the incorrect prediction would be if a prior of 1/2 had been used in the asymmetric treatments (i.e. if prior information were ignored). These incorrect predictions can be plotted as a function of the correct Bayesian prediction. The upper dashed line in Fig. 5 shows the incorrect Bayes' prediction obtained by using a prior of 1/2 for the treatment with a true prior of 1/3, and the lower dashed line shows the analogous prediction where the true prior of 2/3 is replaced with 1/2. The data averages for the two asymmetric treatments, represented by the solid lines in Fig. 5, tend to fall between these lines, which indicates that prior information is in fact being used to some degree. Finally, note that the subjects are obviously not ignoring sample information since the data lines would not have positive slopes in this case.

Another way to explore the updating behavior of subjects is to examine how the average reported probability changes when uniform signals are reinforced by subsequent draws. For example, how are reported probabilities affected when subjects observe two red signals (*RR*) as opposed to just one red signal (*R*)? Although some of this information is contained in the previous figures, it is highlighted in Tables 3 and 4, which show the average elicited probabilities in all three treatments for rounds where all draws were of identical color.

[6] There were several other minor differences between the hand-run and web-based experiments. There was no monitor in the web-based experiments since there were no procedures to be observed. In addition, the web-based experiments allowed values of *N* between and including 0 and 100 and the *N* lottery was only selected whenever *N* was strictly greater than the reported probability, as opposed to the hand-run experiments where the *N* lottery was used if *N* was greater than or equal to the reported probability. Lastly, red and blue balls were used in the computerized sessions instead of light and dark marbles.

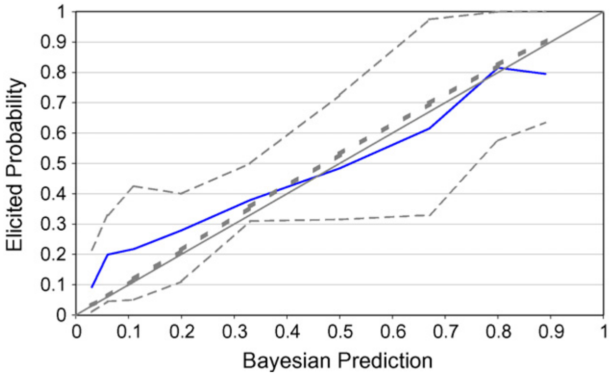

Fig. 3. Average and bounds on 2/3 of the elicited probabilities (solid line and thin dashed lines) and fitted values based on estimation (thick dotted line) for the treatment with a prior of 1/3.

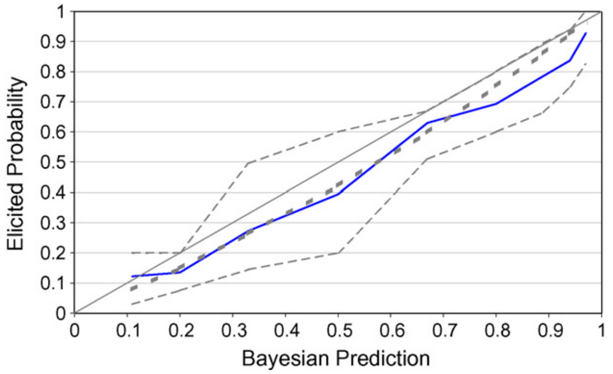

Fig. 4. Average and bounds on 2/3 of the elicited probabilities (solid line and thin dashed lines) and fitted values based on estimation (thick dotted line) for the treatment with a prior of 2/3.

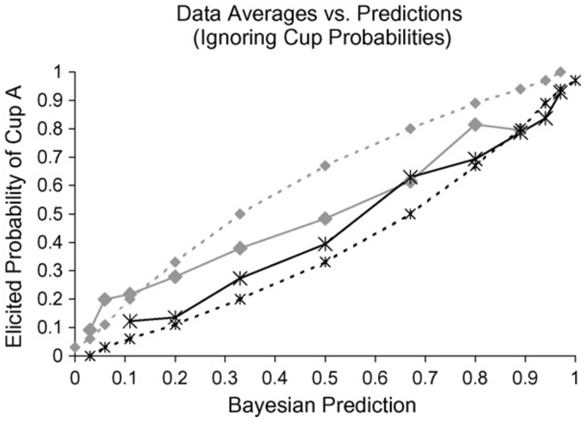

Fig. 5. Average elicited probabilities (solid lines) and Bayesian predictions based on an incorrect prior of 1/2 (dashed lines) for treatments with actual prior probabilities of 1/3 (diamonds) and 2/3 (stars).

Table 3
Average elicited probabilities for the red cup with a reinforcing sequence of red signals.

		Draws			
		R	*RR*	*RRR*	*RRRR*
Prior	1/3	.46	.60	.81	.80
	1/2	.63	.74	.86	.87
	2/3	.68	.79	.84	.93

Table 4
Average elicited probabilities for the red cup with a reinforcing sequence of blue signals.

		Draws			
		B	*BB*	*BBB*	*BBBB*
Prior	1/3	.30	.24	.20	.09
	1/2	.37	.24	.18	.11
	2/3	.46	.29	.14	.12

Table 5
Average elicited probabilities for the red cup with single draws versus mixed draws representative of red cup.

Prior	Bayesian posterior	Draws	
		R	2 reds, 1 blue
1/3	.50	.46	.53
1/2	.67	.63	.71
2/3	.80	.68	.71

Table 6
Average elicited probabilities for the red cup with single draws versus mixed draws representative of blue cup.

Prior	Bayesian posterior	Draws	
		B	2 blues, 1 red
1/3	.20	.30	.25
1/2	.33	.37	.32
2/3	.50	.46	.31

Subjects seem to be incorporating some information from sample draws, as evidenced by the overall trend (with one exception) of elicited probabilities to increase from left to right in each row of Table 3 and to decrease from left to right in each row of Table 4. That is, subjects update their beliefs in the correct direction when they observe reaffirming draws of the same color.

As mentioned previously, there is also another interesting updating phenomenon known as representativeness bias where a group of mixed draws that represents the population in the red (blue) cup pulls the elicited probabilities up (down) from the situation with only one draw that shares the same Bayesian prediction. As with the hand-run experiment, data in these computerized sessions continue to demonstrate this bias. Tables 5 and 6 that show the average elicited probabilities for single draws and mixed "representative" draws along with the relevant Bayesian posterior probabilities for all three treatments. The bias occurs in all treatments where, starting from the situation with a single draw, the mixed signals bring up elicited probabilities when they are representative of the red cup, and they bring down elicited probabilities when they are representative of the blue cup.

Many of the documented violations of Bayes' rule involve very low prior probabilities (e.g. in cases of rare diseases where individuals overweight test results). Therefore, we ran a fourth treatment with a much lower prior for cup *A*. The prior was communicated by the use of a simulated 100-sided die, with 4 sides that result in cup *A*, which only contained red marbles, and 96 sides that result in cup *B*, with a mix:

$$\text{Cup}\,A : R, R, R, R, R, R, R, R, \qquad \text{Cup}\,B : R, R, R, B, B, B, B, B.$$

Obviously, $\Pr(\text{cup } A|B) = 0$. The frequency interpretation of Bayes' rule can be used to derive the other posterior for a single draw: in 100 repetitions, cup *A* would tend to be used 4 times on average, and this would generate 4 "true positives." To get a posterior of 0.1 after an *R* draw, we would need to balance the four true positives with 36 false positives, which results if cup *B* is used 96 times with a 3/8 chance of *R*, since $(3/8) \times 96 = 36$. Thus a test that has no false negatives and only a 3/8 false positive rate would yield a posterior of only 1/10 for cup *A* following a positive test. As expected, the average elicited probabilities for this experiment are higher than the Bayes' predictions, except at the extremes, as shown by Fig. 6.

5. Estimation

The overestimation of low probabilities and underestimation of high probabilities that appears in Figs. 1 and 2 is consistent with the notion of "probability weighting". A comparison of these figures with Fig. 6 indicates that probability weighting applied to a Bayes' posterior probability cannot be the whole story since the degree of overweighting of low probabilities is much more severe for the low-prior treatment shown in Fig. 6. In this section, we estimate a model in which the degree of probability weighting may be different for prior probabilities than for likelihoods based on observed samples.

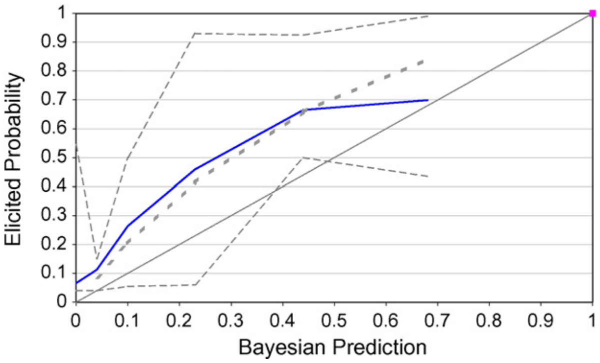

Fig. 6. Average and bounds on 2/3 of the elicited probabilities (solid line and thin dashed lines) and fitted values based on estimation (thick dotted line) for the treatment with a prior of 4/100.

Table 7
Ordinary least-squares estimates.

Model	Constant	Log prior odds	Log likelihood ratio	Representative of cup *A*	Representative of cup *B*	R^2	*N*
1	–.098 (.028)	.713 (.024)	1.027 (.024)	–	–	.504	2395
2	–.089 (.029)	.732 (.025)	.995 (.026)	.234 (.098)	–.387 (.112)	.507	2395

We start with a one-parameter specification used by Wu and Gonzalez (1996), and others:

$$w(p) = \frac{p^\gamma}{(p^\gamma + (1-p)^\gamma)^{1/\gamma}}, \tag{1}$$

where $w(p)$ is the weighted or perceived probability, p is the actual probability, and γ is a weighting parameter. Note that weighted and actual probabilities are equal if $\gamma = 1$. This formula generates over-weighting of low probabilities and under-weighting of high probabilities when $\gamma < 1$. It follows from the symmetric nature of the parameterization of probability weighting in (1) that the odds ratio for the perceived probability is just a power function of the true odds ratio:

$$\frac{w(p)}{w(1-p)} = \frac{p^\gamma}{(1-p)^\gamma} = \left(\frac{p}{(1-p)}\right)^\gamma. \tag{2}$$

The probability p on the right side of (2) is the Bayes' posterior resulting from the sample, S, of draw(s) if any:

$$\Pr(A|S) = \frac{\Pr(S|A)\Pr(A)}{\Pr(S|A)\Pr(A) + \Pr(S|B)\Pr(B)}. \tag{3}$$

It follows from (2) and (3) that the perceived odds ratio can be expressed as power functions of the likelihood ratio and the prior odds ratio:

$$\frac{w(p)}{w(1-p)} = \left(\frac{\Pr(A)}{1-\Pr(A)}\right)^\gamma \left(\frac{\Pr(S|A)}{\Pr(S|B)}\right)^\gamma. \tag{4}$$

A natural generalization is to allow the degree of probability weighting for the priors and the likelihood probabilities to differ, and therefore, we estimated a log-linear equation of the following form:

$$\ln\left(\frac{r}{1-r}\right) = \tau + \gamma_1 \ln\left(\frac{\Pr(A)}{1-\Pr(A)}\right) + \gamma_2 \ln\left(\frac{\Pr(S|A)}{\Pr(S|B)}\right), \tag{5}$$

where r is the reported probability, as determined by weighting the prior and likelihood odds ratios. The constant in (5) actually results from multiplying a constant times the middle and right parts of (2) so that the log weighted odds ratio is a linear function of the true odds ratio plus a constant, which produces a two-parameter probability weighting function discussed in Gonzalez and Wu (1999).

Table 7 shows the estimated ordinary least squares coefficients for the model in (5).[7] Following Grether (1992), elicited probabilities of 0 and 1 were converted to .01 and .99, respectively, to avoid undefined values of the reported odds ratio. The data include observations for all subjects in all four web-based treatments except for the rounds in treatment 4 when a blue

[7] This is essentially a panel dataset since each subject makes decisions in multiple rounds. Therefore, there is the possibility of a compound disturbance that represents unobservable individual heterogeneity that is persistent across rounds for the same subject in addition to the idiosyncratic errors that vary across both subjects and rounds. However, since the independent variables are logs of probability ratios determined solely by the treatment prior probability and the random sequence of draws, these independent variables are presumably uncorrelated with the potential unobserved individual heterogeneity, and therefore the least-squares estimates are consistent but not efficient.

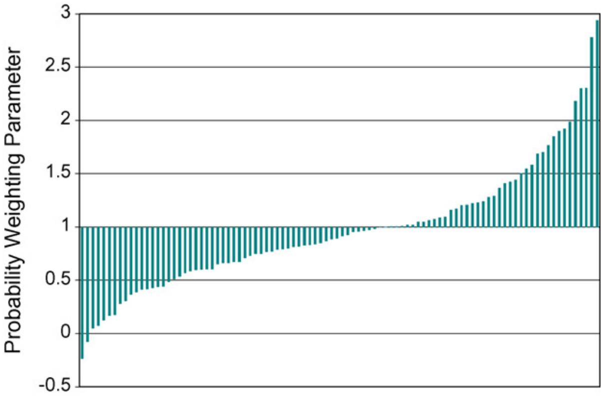

Fig. 7. Estimated probability weighting parameters (γ) for each individual.

draw was observed. These observations were omitted because not only is the probability of cup *A* is equal to 0, which makes the reporting task one of logic as opposed to probability inference, but the log likelihood ratio is also undefined.

The estimates for model 1 are all significant at conventional levels, and the coefficient on the likelihood ratio is not statistically different from 1, which indicates that subjects on average fully responded to observed sample draws.[8] In contrast, the coefficient on the log of the prior odds ratio, .713, suggests some tendency to overweight low prior probabilities and underweight high priors.[9] These estimated coefficients were used together with the log prior odds and log likelihood ratio associated with each possible Bayesian prediction to construct fitted values of the elicited probabilities based on model 1. These fitted values are shown as thick dotted lines in Figs. 2–4 and 6. Notice that the model does not pick up the slight overweighting of low probabilities and underweighting of high probabilities when the prior is 1/2 or 1/3, but it does pick up the downward deviations from Bayesian predictions for a prior of 2/3 and large upward deviations for a prior of 4/100. This may be due in part to the fact that this probability weighting function does not have much bite in the range from .3 to .5.

As discussed earlier with both the hand-run and computerized sessions, representativeness tended to raise the reported probability in the 3-draw sequence that matched the contents of cup *A* as compared with a single draw yielding the same posterior. An approximately symmetric downward bias was observed when the 3-draw sequence matched the contents of cup *B* compared to cases when only a single blue marble was observed, yielding the same posterior probability. For this reason, we also estimated a second model with data from the web-based experiment with dummy variables indicating representativeness with respect to cup *A* or cup *B*. In this estimation, representativeness is defined as the proportion of reds in the sample matching the proportion of reds in the cup. Thus, in treatment 4, any sequence that consisted of all red draws was deemed to be representative of cup *A*, which contained only red balls. As is apparent from the bottom row of Table 7, both of the dummy variables for representativeness are significant and of the expected sign. The coefficients on the log of prior odds and log likelihood ratio remain relatively close to those estimated in the first model.

There is some degree of individual heterogeneity in terms of updating behavior. Some individuals seem consistently to report probabilities very close to Bayesian predictions while others systematically differ in one way or another. In order to investigate these differences, we ran OLS log-linear regressions using the 30 decisions for each individual in the web-based experiment. As opposed to estimating a separate probability weighting parameter for the prior and likelihood probabilities as in (5), we estimate only a single parameter, γ, for each individual that represents the degree of weighting on the entire posterior probability.[10] Fig. 7 shows the estimates for each of the 96 subjects.

Note that there is a large degree of individual heterogeneity in updating behavior. While 45 percent (43 out of 96) of the estimated parameters are not statistically significantly different from the value of γ representing Bayesian behavior ($\gamma = 1$) at the 5 percent level, the range of estimated parameters is very large.[11]

[8] A similar model was estimated by Grether (1992) and Goeree et al. (2007), using data from information cascade experiments.

[9] With $\gamma < 1$, the overweighting of low probabilities and underweighting of high probabilities in (1) shows up as overweighting the odds ratio on the right side of (2) when that ratio is less than 1 and under-weighting it when it is greater than 1. The posterior odds ratio is factored into two terms in (4) and the effect of having the γ exponent be less than 1 is to bias the prior odds ratio in the same manner (upward when the odds ratio is below 1 and downward when it is below 1). This is why we interpret γ_1 in (5) as a measure of probability weighting for the prior probabilities.

[10] To be clear, we estimated the following log-linear equation for each individual:

$$\ln\left(\frac{r}{1-r}\right) = \tau + \gamma \ln\left(\frac{\Pr(A)\Pr(S|A)}{(1-\Pr(A))\Pr(S|B)}\right).$$

We estimate only one probability weighting parameter because the parameter γ_1 in (5) is not identified separately from the constant for each of these subjects, who only made decisions in a treatment with a single prior.

[11] It is important to note that while it is difficult to interpret a γ less than 0, the two estimates reported to be negative are not significantly different from 0 at any conventional significance level. It is also interesting to point out that the 5 estimates above 2.0 imply that those individuals always underweight posterior probabilities.

6. Conclusion

This paper is inspired by the classic experiments of Grether (1980, 1992). Following his lead, we conduct both web-based and hand-run experiments in which people observe random draws and make probability assessments with monetary rewards. The elicited probabilities were phrased in terms of "chances out of 100," and a Becker–DeGroot–Marshak was used to tie these assessments to the earnings that subjects received.

For two of the intermediate values of the prior, behavior in the aggregate was well approximated by Bayesian predictions. For example, if you took the prediction line from Fig. 1 and rotated it 45° to make it level and thought of that as a time-series in a market experiment, then an economist would say "it has converged". There are, however, systematic deviations from Bayesian predictions even in that figure (e.g. overweighting of low probabilities and underweighting of high probabilities). These deviations are much more dramatic when the prior for the event being forecast is increased to 2/3 or reduced to 0.04 (see Figs. 4 and 6). Therefore, it seems that Bayes' rule works well in some of these treatments but not in others, giving rise to the question of whether it should be abandoned as a predictive devise or modified.

A different approach to probability assessment is the notion of probability weighting. We estimate a model that allows probability weighting to differ for perceptions based on prior probabilities and those based on observed random draws. Estimates of this model indicate that subjects only tend to underweight high prior probabilities and overweight low prior probabilities, but they fully account for probabilities associated with observed draws. Although fitted predictions of this model track the large deviations from Bayes' rule for the highest and lowest priors, they fail to pick up minor deviations from Bayes' predictions in the other treatments. In addition, further estimation indicates that subjects are affected by "representative" draw sequences in which the sample color proportions match those of one of the cups. Lastly, while a number of subjects act closely in accordance with Bayesian predictions, there appears to be a substantial amount of individual heterogeneity in behavior.

Acknowledgements

We acknowledge partial financial support from the National Science Foundation (SBR 0094800) and the Bankard Fund for Political Economy at the University of Virginia. We would like to thank Lindsay Osco, Barkley Rosser, and David Zirkle for helpful suggestions.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jebo.2007.08.013.

References

Anderson, L.R., Holt, C.A., 1996. Classroom games: understanding Bayes' rule. Journal of Economic Perspectives 10, 179–187.
Becker, G.M., DeGroot, M.H., Marshak, J., 1964. Measuring utility by a single-response method. Behavioral Science 9, 226–232.
Davis, D.D., Holt, C.A., 1993. Experimental Economics. Princeton University Press, Princeton, NJ.
Gigerenzer, G., Hoffrage, U., 1995. How to improve Bayesian reasoning without instruction: frequency formats. Psychological Review 102, 84–704.
Goeree, J.K., Palfrey, T.R., Rogers, B.W., McKelvey, R.D., 2007. Self-correcting information cascades. Review of Economic Studies 74, 733–762.
Gonzalez, R., Wu, G., 1999. On the shape of the probability weighting function. Cognitive Psychology 38, 129–166.
Grether, D.M., 1978. Recent psychological studies of behavior under uncertainty. American Economic Review 68, 70–77.
Grether, D.M., 1980. Bayes' rule as a descriptive model: the representativeness heuristic. Quarterly Journal of Economics 95, 537–557.
Grether, D.M., 1992. Testing Bayes' rule and the representativeness heuristic: some experimental evidence. Journal of Economic Behavior and Organization 17, 31–57.
Hammerton, M., 1973. A case of radical probability estimation. Journal of Experimental Psychology 101, 252–254.
Holt, C.A., 1986. Scoring rules for eliciting subjective probability and utility functions. In: Goel, P.K., Zellner, A. (Eds.), Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti. North-Holland, Amsterdam, pp. 279–290.
Holt, C.A., 2006. Markets, Games, and Strategic Behavior. Addison-Wesley, Boston.
Holt, C.A., Laury, S.K., 2002. Risk aversion and incentive effects. American Economic Review 92, 1644–1655.
Kahneman, D., Tversky, A., 1973. On the psychology of prediction. Psychological Review 80, 237–351.
Offerman, T., Sonnemans, J., 2004. What's causing overreaction? An experimental investigation of recency and the hot-hand effect. Scandinavian Journal of Economics 106, 533–553.
Offerman, T., Sonnemans, J., van de Kuilen, G., Wakker, P.P., 2007. A truth-serum for non-Bayesians: correcting proper scoring rules for risk attitudes. Working Paper, University of Amsterdam (CREED).
Wu, G., Gonzalez, R., 1996. Curvature of the probability weighting function. Management Science 42, 1676–1690.