Behavioral biases and cognitive reflection☆

Eva I. Hoppe⁎, David J. Kusterer

Department of Economics, University of Cologne, Albertus-Magnus-Platz, 50923 Köln, Germany

A B S T R A C T

In a large-scale laboratory experiment, we investigate whether subjects’ scores on the cognitive reflection test (CRT) are related to their susceptibility to the base rate fallacy, the conservatism bias, overconfidence, and the endowment effect.

© 2010 Elsevier B.V. All rights reserved.

1. Introduction

Only recently, researchers have started to investigate the impact of cognitive ability on judgment and decision making. Frederick (2005) introduces the cognitive reflection test (CRT) which is a simple, three-item test to measure a person’s mode of reasoning and cognitive ability.1 Frederick (2005) shows that people with high CRT scores are generally more patient and more willing to gamble in the domain of gains.2 In a related study, Oechssler et al. (2009) replicate the findings regarding time and risk preferences and in addition they study the relationship between cognitive abilities and the conjunction fallacy, conservatism, and anchoring.3 One central result is that individuals with low cognitive abilities tend to be significantly more affected by behavioral biases.

In the present study, we investigate whether the incidence of further behavioral biases is related to cognitive abilities. Specifically, we study the base rate fallacy, overconfidence, and the endowment effect. Moreover, we replicate the finding of Oechssler et al. (2009) related to the conservatism bias in order to investigate an interesting question that was brought up in their paper. Are people that exhibit the conservatism bias (i.e., overweight the base rate) less susceptible to the base rate fallacy (i.e., to underweight the base rate)? We observe the contrary. In particular, we find that individuals with lower cognitive abilities are significantly more likely to exhibit both, the base rate fallacy and the conservatism fallacy. With regard to overconfidence, we find that subjects with higher CRT scores have a significantly more precise self-assessment. Finally, test scores do not affect the occurrence of the endowment effect which is striking in both, low and high CRT groups.

2. Experimental design

The experiment was conducted in July 2009.4 Using ORSEE (Greiner, 2004), we recruited the participants from the subject pool of the Cologne Laboratory for Economic Research. In total, 414 students from the University of Cologne participated in the experiment. Following several socio-demographic questions (concerning gender, age, field of...
study, and length of study), the subjects had to fill in a questionnaire consisting of three questions building the CRT and several questions related to the behavioral biases mentioned in the introduction. Participants were given 15 min to fill in the questionnaire and the experimenter stopped the experiment after the time was over.\footnote{Subjects found a calculator, a pen, and a piece of paper in their cabin.}

Subjects were paid 0.40€ for each CRT question they answered correctly. Moreover, for the decision problems related to the base rate fallacy and conservatism, they received 0.40€ if their answer did not deviate more than 15 percentage points from the correct answer. Regarding overconfidence, subjects had to answer five general knowledge questions and they had to assess how many of these they answered correctly. For each correct answer (including the assessment question) they received 0.20€.\footnote{Note that only three participants did not complete the questionnaire within the given time limit so that our analysis is based on 411 completed questionnaires.} Finally, with regard to the endowment effect, subjects could receive additional 0.20€ or, alternatively, take a highlighter home. In total, subjects earned between 0 and 2.80€, and the average payoff was 1.24€. Moreover, 180 subjects left the lab with a brand new highlighter. The experiment was programmed and conducted with z-Tree (Fischbacher, 2007).

### 3. Cognitive reflection test

To measure cognitive ability, we use the three-item cognitive reflection test (CRT) that was introduced by Frederick (2005). The three questions are designed such that they have an intuitive but wrong answer that comes to mind quickly and a correct answer that is easy to understand when explained. Hence, the test is supposed to measure a person’s ability to engage in cognitive reflection and thus to resist reporting the spontaneous but wrong answer. In particular, the three questions read as follows.

1. A bat and a ball together cost 110 cents. The bat costs 100 cents more than the ball. How much does the ball cost? (Spontaneous answer: 10 cents; correct answer: 5 cents).
2. If it takes 5 machines 5 min to make 5 widgets, how long would it take 100 machines to make 100 widgets? (Spontaneous answer: 100 min; correct answer: 5 min).
3. In a lake, there is a patch of lily pads. Every day, the patch doubles in size. If it takes 48 days for the patch to cover the entire lake, how long would it take for the patch to cover half of the lake? (Spontaneous answer: 24 days; correct answer: 47 days).

In our sample, 13% of the subjects answered none of the questions correctly, 24% knew the correct answer to one question, 27% to two questions, and 36% answered all three questions correctly. On average, the subjects answered 1.84 of the CRT questions correctly.\footnote{There is a strong gender difference: male subjects have an average score of 2.12, while female subjects have an average score of 1.61 only. The difference is highly significant (p=0.0001, two-sided Mann–Whitney U test). This gender difference has also been found in other studies using the CRT, e.g. Frederick (2005) and Oechssler et al. (2009).}

Table 1 shows the distribution of the answers to the CRT questions. For each question, the majority of the subjects gave the correct answer. Among the subjects who did not submit the correct answer, the intuitive answer was given most frequently.

<table>
<thead>
<tr>
<th>Question</th>
<th>Correct</th>
<th>Intuitive</th>
<th>Other</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bat and ball</td>
<td>56.7%</td>
<td>39.9%</td>
<td>3.4%</td>
</tr>
<tr>
<td>Widgets</td>
<td>58.9%</td>
<td>28.2%</td>
<td>12.9%</td>
</tr>
<tr>
<td>Lily pads</td>
<td>68.9%</td>
<td>16.1%</td>
<td>15.1%</td>
</tr>
</tbody>
</table>

### 4. Results

The central results of our study are summarized in Tables 2 and 5. Depending on their CRT score, we divide the participants in two groups. The “low” group consists of individuals who answered zero or one of the questions correctly, while the “high” group consists of participants that gave the correct answer to two or three questions.\footnote{This categorization was used by Oechssler et al. (2009). We also considered the categorization of Frederick (2005) who assigned subjects with zero correct answers to the “low” group and those with three correct answers to the “high” group. However, with regard to our data this would imply not to analyze more than 53% of the observations. Note that the latter categorization would not change our results qualitatively.} We refer to subjects in the “high” group as the more analytical decision makers, while we describe subjects in the “low” group as relatively intuitive decision makers.

#### 4.1. Base rate fallacy

When people are asked to judge the probability of an event, they often have to take into account information about the base rate probability and at the same time, they have to consider specific evidence about the case at hand (Tversky and Kahneman, 1982). In such a context, they exhibit the base rate fallacy if they follow the representativeness heuristic and neglect the base rate probability.

In analogy to the mammography problem in Eddy (1982), subjects in our study faced the following problem: “In a city with 100 criminals and 100,000 innocent citizens there is a surveillance camera with an automatic face recognition software. If the camera sees a known criminal, it will trigger the alarm with 99% probability; if the camera sees an innocent citizen, it will trigger the alarm with a probability of 1%. What is the probability that indeed a criminal was filmed when the alarm is triggered?” The correct answer is $\approx 5\%$, but in both CRT groups, a large fraction of the subjects stated a probability larger than 90%. These subjects exhibit the base rate fallacy since they do not or barely consider the low base rate of criminals in the population. However, compared to the low CRT group, subjects in the high CRT group are considerably less susceptible to this bias and state the correct probability more often (see Fig. 1).\footnote{In the high CRT group, 19.1% of the subjects choose 9 or 10% as their answer, while in the low CRT group, this answer is stated in only 9.7% of the cases (p=0.01, two-sided $\chi^2$ test). Moreover, the average probability assessed by the subjects in the high CRT group equals 61.5%, which is significantly smaller than 77.4%, the average probability assessed in the low CRT group.}

It is also striking that the average CRT score of subjects who correctly take into account the small base rate is considerably larger than the average CRT score of subjects who exhibit the base rate fallacy (see Table 3).

#### 4.2. Conservatism bias

While people that exhibit the base rate fallacy underestimate base rates, there are also situations where base rates are overweighted relative to sample evidence. In such situations, subjects are too conservative in adapting prior probabilities to new evidence, and hence this fallacy is

<table>
<thead>
<tr>
<th>Category</th>
<th>CRT group</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>low</td>
</tr>
<tr>
<td>Base rate fallacy</td>
<td>Avg. prob. stated (correct prob.: 9%)</td>
</tr>
<tr>
<td>Conservatism</td>
<td>Avg. prob. stated for urn A (correct prob.: 97%)</td>
</tr>
<tr>
<td>Overconfidence</td>
<td>% overconfident</td>
</tr>
<tr>
<td></td>
<td>% correct self-assessment</td>
</tr>
<tr>
<td></td>
<td>% underconfident</td>
</tr>
</tbody>
</table>

The p-values regarding the base rate fallacy and the conservatism bias result from two-sided Mann–Whitney U tests, while the p-value regarding overconfidence is obtained using a two-sided $\chi^2$ test.
called conservatism bias. In order to test whether the tendency to exhibit this fallacy is related to a person’s CRT score, we confronted the subjects with the following problem that was first studied by Edwards (1968): “There are two urns; each one contains ten balls. Urn A contains 7 red and 3 blue balls, while urn B contains 3 red and 7 blue balls. One urn is randomly chosen by flipping a fair coin. 12 balls are now drawn from this urn with replacement. The result is the following: 8 red and 4 blue balls were drawn. What is the probability that the randomly drawn urn is urn A when observing this result (8 red and 4 blue balls)?” The correct answer is 97%, but many subjects (34% in the low CRT group and 28% in the high CRT group) simply entered the base rate of 50% as their answer (see Fig. 2). The average answer in the low CRT group is 56.8%, while it is 60.1% in the high CRT group. The fact that the intuitive decision takers’ average answer is closer to the base rate than the analytical decision takers’ average answer shows that the former group is more heavily affected by the conservatism fallacy than the latter group.

Again, it is remarkable that subjects whose answer is close to the correct one have a larger CRT score on average (see Table 4). This indicates that more reflection and analytical reasoning is helpful to prevent falling for the bias.

### 4.3. Overweighting and underweighting of base rates

The results in Sections 4.1 and 4.2 show that subjects with a low CRT score are on the one hand more inclined to neglect base rates when facing a representativeness problem, but on the other hand they also tend to overemphasize prior information when confronted with conservatism problems. In this context, Oechsler et al. (2009) raise the question whether subjects who exhibit a conservatism bias (i.e., who overweight base rates) are less likely to neglect base rates in representativeness problems. We cannot find evidence for this conjecture. It turns out that subjects who neglect the base rate also tend to overemphasize prior information stronger than subjects who are not susceptible to that bias. It seems to be simply the framing of these problems which leads low CRT subjects to focus more heavily on very salient pieces of information and not to process all available information correctly.

#### 4.4. Overconfidence

We asked the subjects five questions related to general knowledge. After answering the questions, they were also asked to estimate the number of general knowledge questions they had answered correctly. 11

Within each CRT group, we divided the participants into three subgroups. Subjects were classified as being overconfident (underconfident) when the estimated number of correct answers was larger (smaller) than the actual number, and otherwise they were classified as assessing the number of right answers correctly.

Compared to the low CRT group, relatively more subjects in the high CRT group assessed their number of right answers correctly (see Table 2). The difference is statistically significant. While the more intuitive decision takers are relatively less successful in assessing the right number of correct answers, there is no clear tendency that they are more overconfident than the analytical decision takers (in the low CRT group, the share of subjects being overconfident and also the share of underconfident subjects are larger than in the high CRT group).

#### 4.5. Endowment effect

In experimental economics, it is often observed that subjects’ willingness to pay differs substantially from their willingness to accept (see e.g., Horowitz and McConnell, 2002). This phenomenon is often explained by the endowment effect (Kahneman et al., 1991). In our experiment, half of the subjects found a yellow ‘Stabilo Boss’ highlighter on the desk in their cabin (scenario 1) while the other half of the subjects was given the opportunity to buy such a highlighter from the experimenter when leaving the laboratory (scenario 2). At the end of the questionnaire, subjects in scenario 1 faced the following decision: “From now on, the yellow highlighter on your desk belongs to you. You can decide between the following alternatives: (1) I take the highlighter home when leaving the laboratory; (2) When the payment is conducted, I sell the highlighter to the experimenter at a price of 0.20 €. In case 2, 0.20 € are added to your profit realized in the experiment.” In contrast, subjects in scenario 2 saw a bowl with highlighters when entering the lab and at the end of the questionnaire they were confronted with the following decision: “When the

---

10 For this specific question, we restricted our analysis to a subset of subjects who, for each bias, either clearly exhibited the bias or stated the correct answer. Concerning the base rate neglect problem, subjects exhibit this fallacy when their answer is >90%, while they do not if their answer is 9% or 10%. Concerning the conservatism problem, subjects overweight prior information when their answer equals 50%, while they do not if their answer is >90%. There are 100 subjects who neglect the base rate of which 76 also overweight prior information, while of the 30 subjects who do not neglect the base rate, only 18 overweight prior information (two-sided χ² test, p = 0.086).

11 (1) What is the distance between earth and sun in astronomical units? a) 587; b) 1; c) 4553; d) 14. (2) How many inhabitants does the Saarland (a German federal state) have? a) 2,132,000; b) 1,670,000; c) 1,037,000; d) 890,000. (3) In which year did Albert Einstein die? a) 1955; b) 1947; c) 1961; d) 1938. (4) Who is the author of “Wilhelm Tell”? a) Johann Wolfgang von Goethe; b) Friedrich Schiller; c) Friedrich Hölderlin; d) Theodor Fontane. (5) Which metropolitan area has the largest number of inhabitants? a) Shanghai; b) Istanbul; c) Los Angeles; d) Moscow. (6) What do you think: how many of the preceding questions have you answered correctly?
payment is conducted, you can buy a yellow ‘Stabilo Boss’ highlighter at a price of 0.20 €. (1) I want to buy a yellow highlighter; (2) I do not want to buy a yellow highlighter. In case 1, you will receive 0.20 € less.

Overall, we find very strong evidence for the endowment effect. When the highlighter is in the cabin (scenario 1), 64.6% of the subjects decided to take the highlighter home, while only 22.9% of the subjects decided to buy it when the payment is conducted (scenario 2). This difference is highly significant (p ≈ 0.0001, two-sided χ² test).

However, higher cognitive ability does not reduce the susceptibility to the endowment effect (see Table 5).

To conclude, is subjects’ performance on the CRT a good predictor for their susceptibility to behavioral biases? While we have found strong evidence that this is indeed the case for the base rate fallacy and the conservatism bias, the susceptibility to the endowment effect does not vary with the CRT score. Hence, our results suggest that the CRT has strong predictive power only for biases that may arise in problems for which there exists a correct solution and where analytical skills are helpful to derive this solution.

Table 4
Average CRT scores for categorized answers to the conservatism problem.

<table>
<thead>
<tr>
<th>Prob. stated</th>
<th>Average CRT score</th>
<th>n</th>
</tr>
</thead>
<tbody>
<tr>
<td>0–49%</td>
<td>1.68</td>
<td>70</td>
</tr>
<tr>
<td>50%</td>
<td>1.73</td>
<td>124</td>
</tr>
<tr>
<td>51–89%</td>
<td>1.90</td>
<td>171</td>
</tr>
<tr>
<td>90–100%</td>
<td>2.20</td>
<td>46</td>
</tr>
</tbody>
</table>

Table 5
Endowment effect: fraction of subjects buying the highlighter per scenario and CRT group.

<table>
<thead>
<tr>
<th>CRT group</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>low</td>
<td>high</td>
</tr>
<tr>
<td>Scenario 1</td>
<td>61.0%</td>
</tr>
<tr>
<td>Scenario 2</td>
<td>20.6%</td>
</tr>
</tbody>
</table>

The p-values result from a two-sided χ² test.

Fig. 2. Distribution of probabilities stated for the conservatism problem. The dashed line indicates the correct answer.

References


12 Note that the endowment effect is the stronger, the larger is the difference in buying decisions between the two scenarios. In the high CRT group this difference is 42.8%, while it is 40.4% in the low CRT group.