The "Confirmation Bias": Does it Result from Mental Accounting? An Exploratory Analysis

Günter Molz
Justus-Liebig-University Giessen, Department of Psychology,
Otto-Behaghel-Str. 10, 35394 Giessen, Germany
phone: +49 (0) 641 / 99 - 2 61 22 (2 61 21)
e-mail: Guenter.Molz@psychol.uni-giessen.de

(Handout for a poster prepared for SPUDM'16 in Leeds, August 1997)

An experiment was conducted to explore whether the confirmation bias can be interpreted in terms of prospect theory: Useful but disconfirming information is assessed more negatively than information confirming own assumptions. Consequently it was predicted that people value utility of relevant information less than satisfaction about irrelevant information which is compatible with their preconceptions. This hypothesis was tested within the Wason selection task paradigm. After selecting the cards to be turned, subjects were told whether the card was consistent with the given statement. Next they had to assess their satisfaction about this information and its utility. Satisfaction ratings on cards consistent with the statement were higher than utility ratings on relevant results. Utility scores on relevant and irrelevant cards as well as for consistent and falsifying results were assessed rationally. However, all subjects failed to select only relevant cards. Explanations for these inconsistent findings are discussed.

This paper focuses on the "confirmation bias", i. e. the tendency to test a hypothesis by checking examples confirming it. Wason (1966) investigated this problem by means of the so called Wason selection task. Subjects had to test a conditional statement: *If a card has a vowel on one side, then it has an even number on the other side.* Subjects were given four cards, each of which had a letter on one and a number on the other side. The visible sides showed a vowel, a consonant, an even and an odd number. The task was to decide which cards had to be flipped over to test the statement. The logical form of this statement is "if p, then q". The correct solution is to test the instances showing the antecendent p (vowel) and the denied consequent non-q (odd number). These instances may show the only combination falsifying the statement: vowel (p) and odd number (non-q). Only a minority (4 %) selected the relevant cards showing the vowel and the odd number; almost every second subject (46 %) chose the relevant card with the vowel and the irrelevant card with the even number (Wason and Shapiro, 1971). Even if the subjects experienced the problem before the vast majority of them was not able to solve the problem correctly. These findings were interpreted such that subjects want to confirm the statement "if p, then q" by selecting the instances p and q (Wason and Johnson-Laird, 1972). After a closer look at the logic of this task it does not seem to be reasonable to apply to the term "confirmation bias" here. Looking at the instance p may also lead to confirmation (q on the card's other side) as to disconfirmation (non-q on the card's other side). Looking at q cannot confirm the hypothesis since it is an irrelevant instance. Therefore it is always consistent with the statement. Thus the term *confirmation* is not appropriate here. Hence the term "confirmation bias" in this paper is put in quotation marks and the notion of Klayman and Ha (1987) will be preferred: In their terms, looking at p and q is a positive test strategy. Positive testing means to test instances fitting the rule: "If p" determines the target set, whereas "then q" describes the target property.

Also the notion *bias* is problematic for two reasons. First there is a rational argument: Bias implies that this strategy leads to irrational behavior. This is the case regarding instance q, but it is correct to test p. A second argument can be characterized as "meta-rational". According to Jungermann (1986), the meta-rationality argument states that people take into account also the costs of conducting rationally. Regarding this argument it seems rational to test a statement of the form "if p, then q" by checking the instances p and q. - Why? - It is reasonable to formulate an "if ..., then ..."-rule, if two conditions are met:

First there must be a substantial content. This is the case if (a) the base rate for the occurrence of the antecedent p is high, (b) the probability for the consequent q is low. The following simple example shall illustrate this point: "If someone is coughing (p), then he has a cold (q)." The content of this rule would be impaired if the probability for q increased (e. g.: then he has a cold **or** asthma.) In this case the base rate for q would be higher than in the first version of the rule which had more content. Normatively one has to check the non-q instances (i. e. people who do not have a cold). Since there are in case of if-than-relationships with much content more non-q- than q-instances, it is more difficult to search for all cases of non-q in order to test them than looking for q-instances.

The second criterion is the rule's validity. Supposing that a staff manager puts forward the following rule: If someone is sucessful on the job, then he must have had a good school report. Thus he engages only applicants with

good school reports. He might know that a minority of the applicants with good school reports will not be successful on the job and that some applicants with no good school reports are successful in other companies. From a purely logical point of view the staff manager should also test negatively by engaging applicants with average or bad school reports. However, because of a not perfect but substantial positive correlation between school marks and job performance this would increase his portion of wrong decisions (i. e. unsuccessful employees in his company). Consequently, he avoides the negative test
strategy to engage applicants with no good school reports (non-q), although this strategy is logically indicated.

If the two criterions content and validity are met it is reasonable to communicate the relationship between p and q by means of an if-then-statement. Furthermore, in this paper it is suggested that a rule "if p, then q" itself implies content and validity. Thus the positive test strategy will be favoured because it is functional. Hence the evolutionary perspective on human reasoning of several authors is shared that the human mind has been shaped by organizing forces of evolution (e. g. Cosmides, 1989; Gigerenzer and Hug, 1992; Liberman and Klar, 1996). Thereafter in case of an "if-then"-statement it seems reasonable to assume that during a phylogenetic learning process, humans have been sensitized that these statements reflect contentful and valid relationships which can be effectively controlled by positive testing.

A Decision Theoretic Explanation

Mental accounting (Thaler, 1985) refers to the process of categorizing and evaluating options and outcomes. One essential component with which mental accounting can be described is the value function from prospect theory (Kahneman and Tversky, 1979). Davidsson and Wahlund (1992) interpret the "confirmation bias" in terms of prospect theory. Thereafter the loss to have disconfirmed own preassumptions is valued more negatively than the gain from falsifying wrong preconceptions (Davidsson and Wahlund, 1992, p. 351). Regarding the above argument that testing p may lead to confirmation as well as to disconfirmation this explanation is not convincing. An analysis of possible combinations of options and outcomes shows though that the positive test strategy is advantageous in comparison to the negative strategy (see figure 1).

			OUTCOME	
			"consistent with rule"	**"not consistent with rule"**
O	**positive**	**p**	**(1) confirmation**	**(2) disconfirmation**
P T	**test**	**q**	**(3) pseudo-certainty**	**(4)** not possible
I O	**negative**	**non-p**	**(5) irrelevant confirmation**	**(6)** not possible
N	**test**	**non-q**	**(7) confirmation**	**(8) disconfirmation**

Figure 1: Options and outcomes in the Wason selection task

Positive and negative tests potentially lead to confirmation (cells 1 and 7) as well as to disconfirmation (cells 2 and 8). In case the irrelevant instances non-p and q are tested, there is one crucial difference between the positive and the negative test strategy: The positive testing with q always results in an ambiguous confirmation, i. e. this instance is always in line with the rule. Hence, one can conclude that the statement "if p, then q" is correct. This is not the case for the negative test with the instance not-p. It is obvious that for testing a statement ""if p, then ..." testing with the instance not-p is not indicated. Thus to decide on the truth of the statement the result for non-p is irrelevant, since it neither defines the target set nor is it a constituent of it. So the positive test strategy implies the chance to confirm preconceptions by finding a representative of the rule (after testing p) and the sure outcome to find a member of the target set (after testing q). Applying a negative strategy would surely yield an irrelevant result (after testing non-p) and the chance either to falsify the rule (after testing non-q and getting the result p) or to find a card which does not belong either to the hypothesized or to the target set (after testing non-q and getting the result non-p). So in case of a confirming instance p the positive test strategy leads to a pseudo-certainty for the correctness of the statement: It has not been falsified and if the two criteria for if-then relationships content and validity (as mentioned above) were not met, this statement would not have been formulated.

Hypotheses

In order to explore whether the given decision theoretic explanation can account for the preference of the positive test strategy although the correct solution of the task has been experienced before, the following hypotheses were tested:
Hypothesis 1: More positive tests than negative tests are performed.

At first glance setting this hypothesis may seem to be trivial. If, however, in this study there should not be any tendency to positive testing there would not be any necessity to explain this behavior decision theoretically.

As mentioned above Davidsson and Wahlund interpret the modal testing of p and q as avoidance strategy. The central motive is to evade the disappointment following a falsification of one’s assumptions, it is not to find the correct solution. This consideration leads to the next hypothesis.
Hypothesis 2: Satisfaction ratings on results consistent with the statement are higher than utility ratings on relevant results.
If the tendency for positive testing is an avoidance strategy there must be at least a latent understanding of the task. This implies that subjects can estimate the informative value of the cards correctly. Hence, the third hypothesis should find empirical support:
Hypothesis 3: Relevant results lead to a higher confidence for the correctness of the statement than irrelevant results.

Method

40 first year undergraduate students recruited by poster advertising in the University of Giessen took part in the experiment. They had no previous experience with the Wason selection task.

The experiment was conducted on IBM compatible PCs. In the experimental condition subjects (n = 20) were told the correct solution of the Wason Selection Task. Then some exercise trials had to be solved. A statement and two sides of one card were given. Subjects had to determine whether the card was consistent with the statement or not. If they gave correct answers in five consequent trials the next part of the session consisting of four trials was started. In each trial a set of two statements (if p, then q) and four cards (p, q, non-p, non-q) was given. Subjects had to choose one of the two statements for testing and two cards. Then they were told whether the back side of the first selected card was consistent with the statement or not. The card sets across all four trials were designed in such a manner that the instances p and non-q falsified the statement with a probability of .5. Next they had to rate their satisfaction and the perceived utility of this information. Following this procedure, they were told whether the second of the selected cards was in line with the statement. This information had also to be rated with regard to perceived satisfaction and utility. If both selected instances had been consistent with the statement, subjects had to assess their subjective probability for the correctness of the statement. This rating as well as the satisfaction and utility ratings was done on a nine point scale. The symbols on the cards in the training session as well as in the four data collecting trials were letters and numbers. The statement "if p, then q" was varied by exchanging the representatives for p and q across the statements. The instance p, for example, was either "a consonant", "a vowel", "an even number", "an odd number", "not a consonant", "not a vowel", "not an even number" or "not an odd number".

In order to explore the influence of the training this experiment was also conducted with a control group (n = 20). No training session was administered here. The data collecting in four trials took place in the same manner as in the experimental group. After the session subjects were debriefed about the correct solution of the Wason Selection Task.

Results

Before calling the results for the single analyses one remark in advance: The design of this exploratory experiment is a mixed between and within subject design. It was applied for economical reasons. Initial analysis showed that there is no effect of the trials´ subsequence. Because this design is neither a pure between nor a pure within subjects design, assumptions underlying the frequently used parametric tests are violated. Consequently non-parametric techniques were performed. The applied tests have a low asymptotic efficiency, i. e. they are very conservative (Bortz et al., 1992; Roussas, 1997). This can be confirmed by applying their less conservative parametric and non-parametric pendants, which yield lower alpha-error probabilities.

Each subject selected two cards in each of the four trials. Thus in each condition from 20 subjects 160 cards were selected and rated for the perceived satisfaction and utility after flipping them over. In the experimental condition a trial of one subject could not be analyzed because of an error in the protocol file. Consequently, there are only data concerning 158 cards in the experimental condition.

The first hypothesis received strong support. If subjects had selected the cards randomly or correctly they should have done 79 positive tests in the experimental and 80 in the control condition. In the experimental condition 102 cards represented a positive strategy, in the control group 116 cards. For both groups the tendency for positive testing was highly significant ($p < .001$, Chi-square test for comparison of proportions, (Fleiss, 1973)).

The second hypothesis was also supported for the experimental group. Pairs consisting of a satisfaction and a utility rating for a confirming p instance given by the same person were compared with a sign test ($z = 2.46$, $p = .007$) This effect could also be found in the control group ($z = 1.65$, $p = .050$). Table 1 presents the means of the ratings obtained in the single conditions.

	hypothesis	**H 2**				**H 3**	
	group	**experimental**		**control**		**experimental**	
	variable condition	**satisfaction** consistent cards	**utility** relevant cards	**satisfaction** relevant cards	**utility** irrelevant cards	**probability** relevant cards	**probability** irrelevant cards
	mean	7.40	5.50	6.04	5.85	6.20	5.16
	significance (one-tailed test)	p = .007		p = .050		p = .070	

table1: Results for H2 and H3

As predicted in the third hypothesis subjective probabilities for the correctness of the statement after testing relevant instances were distinctly higher than after selecting irrelevant cards (median test, z = 1.48, p = .070; see table 2). This analysis was not performed for the control group since there were only four trials in which relevant instances were selected confirming the statement.
Furthermore, some additional a posteriori analyses were performed (see table 2). Utility ratings for information falsifying the statement were higher than for consistent results (median test, z = 2.66, p = .008) . Also the utility scores for relevant cards consistent with the statement excelled the ratings for irrelevant cards (median test, z = 2.88, p = .004). These differentiations could not be replicated for the control group.

	variable	**utility**							
	group	**experimental**		**control**		**experimental**		**control**	
	condition	falsification	consistent result	falsification	consistent result	relevant cards	irrelevant cards	relevant cards	irrelevant cards
	mean	6.58	5.33	5.28	5.89	6.25	4.75	5.85	5.56
	significance (two-tailed test)	p = .008		p = .984		p = .004		p = .293	

table 2: Results for a posteriori analyses

Interpretation

Although subjects had experienced the logical structure of the problem in the experimental condition, in 36 % percent of all trials the positive testing instances p and q were selected. Not a single subject managed to select the correct cards in all four trials. This shows on the one hand a relative failure of the training session, on the other hand in 48 % of the trials conducted in the control group only positive testing was performed which is in line with Wason`s studies in which the equivalent portion was 46 % (reported in Wason and Shapiro). The difference between the proportions of trials with positive tests in the experimental and control conditon differed from chance expectancy (p < .05, chi-square test for comparing proportions from different samples, Fleiss, 1973). Altogether it can be stated that although there was an understanding for the task subjects failed to solve the problem normatively correctly.

The significant results regarding the second hypothesis support the interpretation of Davidsson and Wahlund which concludes that the positive strategy is an avoidance strategy. Empirical support for this interpretation, however, reinforces that subjects have a latent understanding despite their prescriptively wrong behavior.

Results referring to the third hypothesis and the a posteriori analyses also underline this contradiction between competence and performance. The assessments of the dependent variables utility and subjective probability for the correctness of the rule showed that subjects could differ between relevant and irrelevant instances and recognize the informative value of a falsification.

Conclusions

The title of this paper raised the question whether the "confirmation bias" (i. e. the positive test strategy) is a consequence of mental arithmetic. The data confirmed this supposition for the experimental group. Subjects were obviously driven by the motive to get results that were consistent with the statement they wanted to test. The question arises under which circumstances there might be a shift in such a way that they strive more for a normatively correct solution and less according to the hedonistic principle to find consistent results. Therefore the nomological Wason selection task paradigm has to be expanded in two aspects. Firstly, situational parameters (e. g. irreversibility of the decision, importance, decision maker`s responsibility, time constraints) should be varied, secondly personal parameters (e.g. knowledge, motivation) should be controlled. A suitable framework in this context might be the contingency model of Beach and Mitchell (Beach and Mitchell, 1978, Waller and Mitchell, 1984).

Another interesting issue concerns the causation of pseudo-certainty in this context. Possible explanations might be derivated from prospect theory (faulty weighting of probabilities, Tversky and Kahneman, 1986) or from regret theory (anticipation of regret, Loomes and Sugden, 1982).

In general, future studies on these topics are likely to encourage integration of studies in the two areas problem solving and decision making. Although the Wason Selection Task has become the most intensively researched problem in the history of the psychology of reasoning (Evans et al., 1993, p. 99) and strategies for decision making are but a subset of strategies in general (Christensen-Szalanski, 1978, p. 307) glances at reference lists show that very few articles of each other field are quoted. In the recent decade there seems to be a slight shifting away from this particularism (e. g. Evans et al., 1993). This is regarded to be a sound development. Reinforcing this trend is probably the essential benefit of future research in this domain. The work introduced in this paper shall contribute to this integration.

Literature:

Beach, L. R. and Mitchell, T. R. ‘A Contingency Model for the Selection of Decision Strategies’. *Academy of Management Review*, 3 (1978), 439 - 449.

Bortz, J., Lienert, G. A., and Boehnke, K. *Verteilungsfreie Methoden in der Biostatistik*, Berlin, Heidelberg, New York, London, Paris, Tokyo, Hong Kong, Barcelona: Springer, 1992.

Christensen-Szalanski, J. J. J. ‘Problem Solving Strategies: A Selection Mechanism, some Implications, and some Data’. Organizational Behavior and Human Performance, 22 (1978), 307-323.

Cosmides, L. ‘The Logic of Social Exchange: Has Natural Selection Shaped how Humans Reason? Studies with the Wason Selection Task’. *Cognition*, 31, (1989), 187 - 276.

Davidsson, P. and Wahlund, R. ‘A Note on the Failure to Use Negative Information’. *Journal of Economic Psychology*, 13 (1992), 343 -353.

Evans, J. St. B. T., Newstead, St. E., and Byrne, R. M. J. *Human Reasoning: The Psychology of Deduction*, Hove: Lawrence Earlbaum, 1993.

Fleiss, J. L. *Statistical Methods for Rates and Proportions*, New York, Lomdon, Sidney, Toronto: Wiley, 1973.

Gigerenzer, G. and Hug, K. ‘Domain Specific Reasoning: Social Contracts, Cheating and Perspective’. *Cognition*, 43 (1992), 127 - 171.

Jungermann; H. ‘The Two Camps on Rationality’ in Arkes, H. R. and Hammond, K. R. (eds.), *Judgment and Decision Making*, Cambridge: Cambridge University Press, 1986.

Kahneman, D. and Tversky, A. ‘Prospect Theory: An Analysis of Decision Under Risk’. *Econometrica*, 47 (1979), 263-291.

Klayman, J. and Ha, Y. ‘Confirmation, Disconfirmation and Information in Hypothesis Testing’. *Psychological Review*, 94 (1987), 211-228.

Liberman, N. and Klar, Y. ‘Hypothesis Testing in Wason’s Selection Task: Social Exchange Cheating Detection or Task Understanding’. *Cognition*, 58 (1996), 127-156.

Loomes, G. and Sugden, R. ‘Regret Theory: An Alternative Theory of Rational Choice Under Uncertainty’. *The Economic Journal*, 92 (1982), 805-824.

Roussas, G. G. *A Course in Mathematical Statistics*, San Diego: Academic Press, 1997

Thaler, R. H. ‘Mental Accounting and Consumer Choice’. *Marketing Science*, 4 (1985), 199-214.

Tversky, A. and Kahneman; D. ‘Rational Choice and the Framing of Decisions’ *Journal of Business*, 59 (1986), 251-277.

Waller, W. S. and Mitchell; T. R. ‘The Efffects of Context on the Selection of Decision Strategies for the Cost Variance Investigation Problem’ *Organizational Behavior and Human Performance*, 33 (1984), 397-413.

Wason, P. C. ‘Reasoning’ in Foss, B. M. (ed), *New Horizons in Psychology,* Harmondsworth, Middlesex: Penguin, 1966.

Wason, P. C. and Johnson-Laird, P. N. *Psychology of Reasoning: Structure and Content*, London: Batsford, 1972.

Wason, P. C. and Shapiro, D. ‘Natural and Contrived Experience in a Reasoning Task’. *Quarterly Journal of Experimental Psychology*, 23 (1971), 63-71.

****

The author thanks Andreas Goshöfer and Anja Kampmann for their assistance in collecting the data. Helpful comments by Uwe Meyer are gratefully acknowlegded.

The "Confirmation Bias": Does it Result from Mental Accounting? An Exploratory Analysis

The "Confirmation Bias": Does it Result from Mental Accounting? An Exploratory Analysis

prospect theory conditional reasoning

----

insert Exhibit 1 about here

----

insert Exhibit 2 about here

----

insert Exhibit 3 about here

Günter Molz is currently lecturer in the Department for Methodology at the Faculty of Psychology at the University of Giessen. He was previously research

fellow and assisstant teacher at the Technical Universities of Munich and Berlin. His main research interests concern rational and empirical epistemologies in psychological research.