Momme v. Sydow (2004). *Structural Bayesian Models of Conditionals*. In: K. Forbus, D. Gentner, T. Regier (Eds.). Proceedings of the Twenty-Sixth Annual Conference of the Cognitive Science Society. Mahwah, NJ: Erlbaum, 1411-1416 (CD: 1459-1464).

Structural Bayesian Models of Conditionals

Momme von Sydow (momme.von-sydow@bio.uni-goettingen.de)
Department of Psychology, Georg-August-Universität Göttingen, Abt. 1, Goßlerstr. 14
D-37073 Göttingen, Germany

Abstract

In the past decade the traditional falsificationist view of hypothesis-testing tasks, such as Wason's selection task, has become criticized from a Bayesian perspective. In this report a normative extension of Oaksford's and Chater's (1994, 1998) influential Bayesian theory is proposed, that not only takes quantitative but also qualitative (structural) knowledge into account. In an experiment it is shown that humans appear to be sensitive to both the quantitative and the qualitative preconditions of the proposed normative models.

Introduction

According to falsificationism only tests of hypotheses that may lead to a falsification are normatively justified (Popper, 1934/2002). In the psychology of thinking Wason's (1966) selection task (WST) has become the most studied single task to investigate the testing of hypothesis, typically a indicative conditional in the form of "if *p* then (always) *q*." In this task, four cards are presented. The visible front sides of these cards represent the logical cases *p, non-p, q, non-q*. It is known that one side of each card shows either a *p*- or *non-p*-case and the other side either a *q* or *non-q*-case. In order to test whether the hypothesis is true or false, participants should turn over those cards that are needed to test the hypothesis. To falsificationists, who have been predominant in psychology of reasoning for long, only the selection of a *p*-card *and* a *non-q*-card is correct.

Since over three decades studies have shown that humans do not act in a falsificatory manner (e.g., Johnson-Laird & Wason, 1970): most participants selected the *p*-card and the *q*-card and only 4% selected the 'correct' combination of a *p*- and a *non-q*-card. Since 96% gave wrong answers in this very basic logical task, this finding casts doubt on the rationality of the so-called *animal rationale/zoon echon logon* (Aristotle).

In psychology, theories have been developed which kept a falsificatory core, but which explained the selections by additional mechanisms (e. g. mental model theory). Also other theories flourished, which completely broke with the concept of normative rationality altogether (Cheng & Holyoak, 1985; Cosmides, 1989; Gigerenzer & Hug, 1992). In the last decade, however, probabilistic and Bayesian approaches to the WST have been also proposed (early proposals were e. g. Kirby, 1994; Oaksford & Chater, 1994; Evans & Over, 1996). The optimal data selection model of Oaksford and Chater (1994, 1996, 1998; Oaksford, Chater & Grainger, 1999) represents the most refined approach and has received most attention (e. g.: Evans & Over, 1996; Laming, 1996; Klauer, 1999; Oberauer, 2000; Osman et al., 2001). Hence, I am here going to focus on this approach.

Models

The model of hypothesis testing by Oaksford and Chater (1994, 1998), shown in Table 1, distinguishes a dependence sub-model M_D and an independence sub-model M_I, which represent the truth or falsity of the conditional. As in logics $P(p \wedge \neg q|M_D)$ is set as zero. Different to logics the other cells in this model are quantified. By comparing of M_D and M_I it can be seen (without the further modeling steps) that in such a model not only the falsificatory selections *p-/non-q*-card selections, but also *q*-card-selections may provide a certain *information gain*. However the *non-p*-card never becomes informative, since Oaksford & Chater (1994) set $P(p)$ and $P(non\text{-}p|q)$ to be equal in both sub-models, which in turn cause a flexible q-marginal probability.[1]

This setting of parameters has been criticized by Laming (1996) as post hoc data model fitting, designed to preclude the prediction of (actually infrequent) *non-p*-selections. Laming argued that these assumptions could not be justified, since one may equally construct a model with different parameters that appears completely weird (Table 3). Oaksford & Chater (1996, p. 386) defended their setting of parameters: "Psychologically it reflects the finding that participants regard false antecedent instances (i.e., the not-p cases) as irrelevant to the truth or falsity of a conditional rule." (Cf. recently similar ideas by Over, in press, and by Evans, Headley and Over, 2003)

von Sydow (2002) argued at length against this view and in favor of a different approach. Oaksford's and Chater's above argument, for example, is against the spirit of their own approach, since by this argument also *non-q*-card selections could have been excluded a priori. Inspired by Laming's criticism, I discussed and empirically examined a model in which the resulting marginal probabilities $w(p_{res})$ and $w(q_{res})$ are set to be constant in both sub-models (Table 2). This model is actually long known from the philosophical literature on the raven paradox, but von Sydow has combined it with the further calculations of the refined model of Oaksford and Chater (1998), stressing that this model has necessary preconditions to be fulfilled by the empirical situation. At about the same time similar proposals were made (Hattori, 2002) and even Oaksford &

[1] Oaksford and Chater (1994, 1998) modified the probabilities used in Table 1 according to the following formula: $q:=[P(q)-P(p)\ P(M_D)]/[1-P(p)\ P(M_D)]$. Both models were analyzed, the pure model without the modification of $P(q)$ (Model 1) and the model with the modification. The predictions of both models are similar and in this paper I focus on the pure model (cf. table 4).

Table 1: Model of Oaksford and Chater (1994, 1998).
P(p) and P*(q|1-p)* are set to be the same in both sub-models (cf. footnote 1).
Notes for Table 1 to 3: The cells show probabilities for the Dependence Model M_D and the Independence Model M_I. Resulting marginal probabilities, $P(p_{res})$ and $P(q_{res})$, can differ from *P(p)* and *P(q)*. '*p*', '*q*' in italics abbreviates *P(p)*, P*(q)*.

M_D	q	non-q	
p	*p*	*0*	*p*
non-p	*(1-p)q*	*(1-p)(1-q)*	*1-p*
	q+p-pq	*(1-p)(1-q)*	*1*

M_I	q	non-q	
p	*pq*	*p(1-q)*	*p*
non-p	*(1-p)q*	*(1-p)(1-q)*	*1-p*
	q	*1-q*	*1*

Table 2: Model of von Sydow (2002), Oaksford and Wakefield (2003). *P(p)* and P*(q)* are set to be constant. (Cf. Table 1)

M_D	q	non-q	
p	*p*	*0*	*p*
non-p	*q-p*	*1-q*	*1-p*
	q	*1-q*	*1*

M_I	q	non-q	
p	*pq*	*p(1-q)*	*p*
non-p	*(1-p)q*	*(1-p)(1-q)*	*1-p*
	q	*1-q*	*1*

Table 3: Model of Laming (1996). *P(q)* and P*(p|q)* are set to be constant. (Cf. Table 1)

M_D	q	non-q	
p	*pq*	*0*	*pq*
non-p	*(1-p)q*	*1-q*	*1-pq*
	q	*1-q*	*1*

M_I	q	non-q	
p	*pq*	*p(1-q)*	*p*
non-p	*(1-p)q*	*(1-p)(1-q)*	*1-p*
	q	*1-q*	*1*

Wakefield (2003) in a revision turned to this model.[2] However, in these other proposals it was never stressed that in principle *all* models could be normatively justified (also the original one of Oaksford & Chater, 1994). In contrast, von Sydow argued that all could be justified, provided that their (implicit) preconditions hold in the experimental situation. In three experiments von Sydow (2002) ensured that the preconditions of the model from Table 2 were fulfilled, by ensuring fixed marginal probabilities. The results showed the predicted increase of *non-p*- and *non-q*-selections in high base rate conditions.

The aim of this present paper is to directly investigate whether humans are actually sensitive to these *different* structural preconditions. Therefore the original model of Oaksford and Chater (1994, 1998)[1], the model of von Sydow (2002) and also the model of Laming (1996) were modeled along the same lines. In regard to further steps of modeling (Bayes' Theorem, Wiener-Shannon-Information and the resulting *expected information gain* measure) I completely followed Oaksford and Chater (1998).[3]

Here only an extract of the modeling results can be presented (see Table 4). *Expected information gain* (*EIg*) values are shown for the different models for the parameter values used in the experiment (low base rate: *P(p)=.10, P(q)=.20, P(H_D)=.50*, high base rate *P(p)=.80, P(q)=.90, P(H_D)=.50*). Additionally also the normative predictions for the estimates of the marginal probabilities are shown. (The predictions are also mentioned in the results section.)

Table 4: Expected information gain and standardized expected information gain (with an error parameter) for card selections in different structural models (for low, *.10→.20,* and high base rates, *.80→90).* Resulting marginal probabilities $P(p_{res}|M_D)$, $P(q_{res}|M_D)$, $P(p_{res}|M_I)$, $P(q_{res}|M_I)$.[1]

EIg	von Sydow-Model				Oaksford-Chater-Model 1			
SEIg	p	¬p	q	¬q	p	¬p	q	¬q
low	.61	.01	.15	.05	.61	.00	.07	.05
	.58	.09	.20	.20	.63	.09	.15	.13
M_D	$P(p_{res})=.10$		$P(q_{res})=.20$		$P(p_{res})=.10$		$P(q_{res})=.28$	
M_I	$P(p_{res})=.10$		$P(q_{res})=.20$		$P(p_{res})=.10$		$P(q_{res})=.20$	
high	.05	.15	.01	.61	.05	.00	.00	.61
	.12	.20	.09	.58	.14	.09	.09	.67
M_D	$P(p_{res})=.80$		$P(q_{res})=.90$		$P(p_{res})=.80$		$P(q_{res})=.98$	
M_I	$P(p_{res})=.80$		$P(q_{res})=.90$		$P(p_{res})=.80$		$P(q_{res})=.90$	
EIg	Oaksford-Chater-Model2				Laming-Model			
SEIg	p	¬p	q	¬q	p	¬p	q	¬q
low	.67	.00	.10	.05	.61	.00	.00	.05
	.63	.08	.16	.12	.67	.09	.09	.16
M_D	$P(p_{res})=.10$		$P(q_{res})=.24$		$P(p_{res})=.02$		$P(q_{res})=.20$	
M_I	$P(p_{res})=.10$		$P(q_{res})=.16$		$P(p_{res})=.10$		$P(q_{res})=.20$	
high	.09	.00	.00	.61	.05	.07	.00	.61
	.17	.09	.09	.65	.13	.15	.09	.63
M_D	$P(p_{res})=.80$		$P(q_{res})=.97$		$P(p_{res})=.72$		$P(q_{res})=.90$	
M_I	$P(p_{res})=.80$		$P(q_{res})=.83$		$P(p_{res})=.80$		$P(q_{res})=.90$	

[2] Oaksford, Chater & Larkin (2000) had distinguished a similar model of reasoning from their model of hypothesis testing. The revision has been announced – without any reasons and without own data – in an overview article (Oaksford & Chater 2001, p. 353), which can not count as a full revision of their model.

[3] For alternative proposals cf. Laming (1996), Klauer (1999), and Chater & Oaksford (1999).

Method

Design and Participants The experiment had a 2 (low versus high base rate condition) × 3 (the three structural models) between-subjects design.

Seventy-two participants from the University of Göttingen took part in the experiment. The participants were randomly assigned to the six experimental conditions.

Materials and Procedure. Each participant was presented with what I call a 'Many Cards Selection Task' (MST) with many depicted cards (instead of four cards in a WST) in a paper and pencil version.

In all conditions the same cover story was used. Participants were asked to suppose that they were physicians at a university hospital. Their task was to find out whether the following hypothesis was true or false: "If a patient is infected by the Virus Adenophage (A), then he always shows the symptom Thoraxpneu (●)" This hypothesis was set in bold print. In order to set the parameter $P(M_D)$ in all models *to 0.5* the participants were told that it is equally likely, that the hypothesis is true or that there is no correlation between the virus and the symptom at all. The participants were told, that the head nurse is in charge of all the patient files, in the form of 100 patient cards. Each patient card on the front side provides information about tested viruses and on the backside information about symptoms.

The cards were then shown to the participants. First the head nurse laid out the front sides of the cards, showing whether a patient had the specific virus (A) or not (-). Then she quickly takes up the cards. Thereby the cards are completely mixed (bold print). Secondly she then laid out the backsides of the cards, showing whether a patient has shown the specific symptom (●) or not (O).

Depending on the experimental condition it varies which cards are shown. The proportion of cards *p*- versus *non-p*-cards and *q*- versus *non-q*-cards resulted from how the parameters were set (low base rate: $P(p)=0.1$, $P(q)=0.2$, high base rate condition $P(p)=0.8$, $P(q)=0.9$).

In the structural condition with constant marginal probabilities (von Sydow; 2002) $P(p|H_D)=P(p|H_I)$ and $P(q|H_D)=P(q|H_I)$ were induced by showing *all* fronts and backs of the cards (after mixing them in between). For the Oaksford and Chater (1994)-model, with $P(p|H_D)=P(p|H_I)$ and $P(q|non\text{-}p \mid H_D)=P(q|non\text{-}p| H_I)$, also all cards were first shown with the virus-side facing upwards ($P(p|H_x)$). But after mixing, the symptom-sides only of those patients were shown who had no virus ($P(q|non\text{-}p \mid H_x)$). Thereby I directly provided information on $P(\neg p \wedge q)$ and $P(\neg p \wedge \neg q)$, which should remain constant in this model. No direct information was provided of the *q-/non-p*-marginal probabilities, which are not constant in this model. Similarly, in the Laming (1996)-condition, with $P(q|H_D)=P(q|H_I)$ and $P(p|q \mid H_D)=P(p|q \mid H_I)$, all cards were first shown now with the symptom-side visible. After mixing, the virus-sides of the cards only of those patients were shown, who have had the specific symptom. Thereby I directly provided information on $P(p \wedge q)$ and $P(\neg p \wedge q)$, which are constant in that model and no direct information on the *p*- and *non-p*-marginal probabilities.

All participants were then instructed that the head nurse was not willing to turn over many cards separately. She would only allow *one* card to be turned over on its own. Participants were asked, what card they would select to test their hypothesis. Firstly the participants should suppose the head nurse had put two patient cards in front of them, one of a patient with the virus (A) and one card of a patient without the virus (-) (*p*-card, *non-p*-card).[4] Secondly they should instead suppose a situation in which two patient cards were placed before them, one of a patient with the symptom (●), one of a patient without the symptom (O) (*q*-card, *non-q*-card). In both cases they had to choose which card they would turn over.

Finally, four questions were used (in a frequency format), to survey the participant's estimation of the marginal probabilities resulting in each model, that is: $P(p_{res}|H_D)$, $P(q_{res}|H_D)$, $P(p_{res}|H_I)$, $P(q_{res}|H_I)$. The participants were asked how many of all 100 patients would have the Virus A and how many of all 100 patients would have Symptom T, assuming that the hypothesis is true or false.

Results and Discussion

First the card selections are described, then the estimations of the marginal probabilities.

Card selections

Table 5: Percentages and number of selections of the *p*- and *non-p*-cards and *q*- and *non-q*-cards. (N=72)

	Structural Models					
	Sydow		Oaksford		Laming	
	low	high	low	high	low	high
p	92%, 11	25% 3	83% 10	83% 10	83% 10	58% 7
non-p	8% 1	75% 9	17% 2	17% 2	17% 2	42% 5
q	75% 9	25% 3	58% 7	17% 2	42% 5	45% 5
non-q	25% 3	75% 9	42% 5	83% 10	58% 7	55% 6

[4] The original WST with its four cards may be interpreted as a sequential task, in which the first selection may influence the second, or in which even a planned second selection may influence the first. This would not be modeled by the general approach of Oaksford & Chater (1994). Such effects are minimized by this forced choice design. (Cf. also Klauer 1999.)

Moreover, this design is a severe test of the predicted increase of *non-q*- and *non-p*-card selections: not only the relevance of these cards, but their relative predominance is tested against the normally common *p*-card and *q*-card selections.

von Sydow (2002)-Model For this model a rise in the proportion of *non-q*-selections and *non-p*-selection was predicted for the high base rate condition. The descriptive results are shown in Table 5 and visualized in Figure 1.

Both differences were statistically significant, the *q-/non-q*-effect (Pearson: χ^2 *(1, n=24)=6.0*, one-tailed, *df=1; p<.01*) as well as the *p-/non-p*-effect (Pearson: χ^2*(1, n=24) =10.9*, one-tailed, *p<.001*). The parameters in this experiment were chosen that for this model *EIg(non-q|high)= EIg(q|low)* and *EIg(non-p|high)= EIg(p| low)*. The results of the *q-/non-q*-effect are indeed perfectly symmetrical, the *p-/non-p*-effect descriptively only shows a small *p*-bias. *Within* the high base rate condition more *non-p* than *p* and more *non-q* than *q*-selections were predicted. These cards even became predominant in a statistically significant way (both: χ^2 *(1, n=12)=3.0*, one-tailed, *p<.05*).

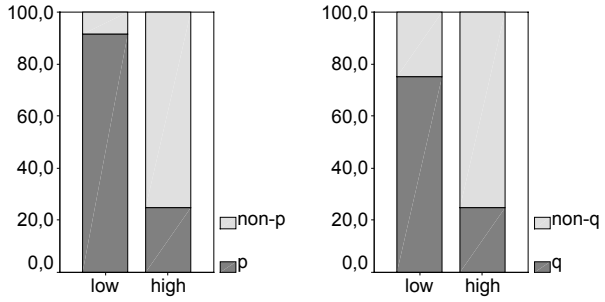

Figure 1: von Sydow-Model: (a) Proportion of *p-/non-p-*card selections and (b) proportion of *q-/non-q*-card selections in high and low base rate conditions.

Oaksford and Chater (1994)-Model This model similarly predicts an increase of *non-q-card* selections in the high base rate condition, but it does not predict an increase of *non-p*-card selections. The results are visualized in Figure 2.

For the *p-/non-p*-cards there was indeed no difference between the low and high base rate condition (Fisher-Yates test (*1, n=24*, one-tailed): *p=0.70*). As also hypothesized, the frequency of *non-q*-card selections was significantly higher in the high base rate condition than in the low base rate condition (Fisher-Yates test (*1, n=24,* one-tailed): *p<.05*). Even the perhaps surprising high rate of *non-q*-card selections in the low base rate condition appears reasonable with regard to the *EIG* and *SEIg* values (cf. Table 4).

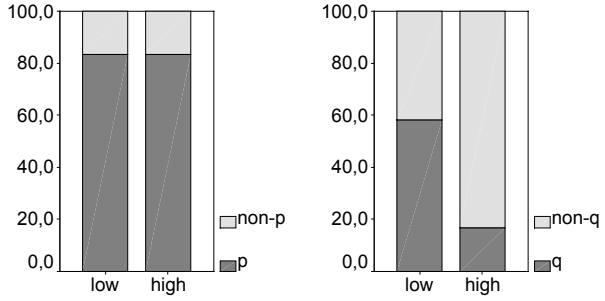

Figure 2: Oaksford-Chater-Model: Proportion of selections.

Laming (1996)-Model Although Laming's proposal was originally only thought as an absurd example, it was modeled and the prediction of a constantly high *non-q*-card selection and a *p-/non-p*-effect was derived.

As expected, no *q-/non-q*-effect was found (Fisher-Yates test (*1, n=23,* one-tailed): *p=.58*). But in difference to the predictions the *p-/non-p*-effect was not significant (Fisher-Yates test (*1, n=24,* one-tailed): *p=.18*). However, even here the results descriptively point in the predicted direction and in the high base rate condition over 40% preferred a *non-p*-card (figure 4).

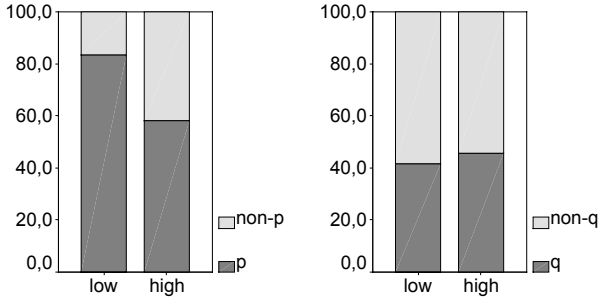

Figure 3: Laming-Model: Proportion of selections.

In summary, the card selections were clearly confirmative for both the von Sydow-model and the Oaksford-Chater-model, and they at least pointed in the predicted direction for the Laming-model.

Estimates of marginal probabilities Participants' estimates of the resulting marginal probabilities was a second depended variable to assess whether the participants fully understood the implications of the induced models.

Only an abridged analysis of these data can be given here. In Table 5 the means and modes of the subjective estimates of the marginal probabilities $P(p_{res})$ and $P(q_{res})$ are shown, if the participants had to assume the rule to be true or false.

An analysis of the data shows that the means are not the appropriate measures to assess the differences between the conditions, since in some cases *two* types of answers clearly predominated. Hence Table 5 also shows the modes (two modes are shown when both have the same frequency or when their frequency differed only by one). It was tested whether the number of cases represented by each mode of the estimations (or by the two modes) is predominant relatively to all other cases not matching that mode(s). This was tested for significance with a Chi^2-test *(df=1, one-tailed, 12≥N≥9).* (For Results cf. Table 5.)

In the von Sydow-Model the modes were all normative. Each mode had a frequency of over 70%. The χ^2-tests showed, that the number of estimations matching the modes was in all but one case significantly higher than all other estimations taken together (Table 5).

In the Oaksford-Chater-Model and in the Laming-Model a relevant number, but not all, of estimations confirmed the predictions. But it will be shown that the deviations also showed an interesting inner consistence.

Table 5: Estimates of the resulting marginal probabilities given the truth (M_D) or falsity (M_I) of the hypothesis. For each model the following values are shown: normative answers for $P(p_{res})$, then for $P(q_{res})$, means of these answers, modes. A mode (or two taken together) got an asterisk (*), if their predominance was also statistically significant. (They also always united over 75% of the answers.)

$P(p_{res})$ $P(q_{res})$		Sydow			Oaksford			Laming		
		Normative	Mean	Mode	Normative	Mean	Mode	Normative	Mean	Mode
Low base rate	M_D	10 20	10 18	10* 20*	10 28	10 19	10* 28;10*	2 20	13 17	2; 20* 20*
	M_I	10 20	10 19	10* 20*	10 20	10 18	10* 18*	10 20	5 19	2* 20*
High base rate	M_D	80 90	73 79	80* 90*	80 98	77 81	80* 98,80*	72 90	76 84	72;90* 90*
	M_I	80 90	76 85	80[5] 90*	80 90	73 58	80* 90; 50	80 90	67 77	72 90*

In both models the results clearly and significantly confirmed the predictions in regard to the constant marginal probabilities, that is in regard to $P(p_{res})$ in the Oaksford-Chater-Model and in regard to $P(q_{res})$ in the Laming-Model). In each of these four cases there was only one mode, which in number outweighed all other predictions significantly. Also as predicted, a change of modes (between M_D and M_I) was found in the Oaksford-Chater-Model in regard to $P(q_{res})$, and conversely in the Laming-Model in regard to $P(p_{res})$. But opposed to the predictions in both models two modes were found, given the hypothesis is assumed to be true. The two modes taken together significantly outweighed all other predictions in all four cases. In all these cases – independent of a high or a low base rate – one of the two modes exactly was the predicted one. The other mode in all cases was consistent with an equivalence interpretation of the hypothesis. In the Oaksford-Chater-Model this second mode of $P(q_{res}|M_D)$ matched the correct estimations of $P(p_{res}|M_D)$. Conversely in the Laming-Model the second mode of $P(p_{res}|M_D)$ exactly matched $P(q_{res}|M_D)$. Hence in both models one set of answers exactly shows the expected changes between M_D and M_I. Another set of answers is consistent with an interpretation of the hypothesis not as implication, but as equivalence. (Based on the low *N* no further analysis of this additional effect was possible.)

In summary, also the results for the estimations show that participants distinguished the tested models. The results for the von Sydow-model were unambiguously positive. In the Oaksford-Chater-Model and the Laming-model a substantial number of answers fully confirmed the predictions. In these models a second group, however, was consistent with an interpretation of the rule as equivalence. Interestingly, the ambiguity of the interpretation of the hypothesis appears to be a function of the induced model.

[5] Also here 70 % (of 11 answers) matched the mode.

General Discussion and Conclusion

The empirical results provide evidence that humans are sensitive both to the structural as well as to the quantitative aspects of the tested Bayesian models.

The card selections largely confirmed the predicted differential effects of structural models and of the card frequencies. Estimates of the resulting marginal probability provide evidence that at least a substantial part of the participants also understood these implications of the models.

Implications for Non-Bayesian Approaches

Approaches that are normatively based on basic formal logics (excluding e. g. fuzzy logics) and its falsificationist interpretation have clear normative predictions in all conditions of the experiment. In each and every case one equally ought to select the *p*- and the *non-q*-card, since these are the only cards by which a (conclusive) falsification could be achieved. The main traditional psychological theories of conditionals, the mental logics theory and the mental model theory (cf. Johnson-Laird & Byrne, 2002) are normatively still tightly linked to the falsificationist research program. But also with their additional psychological assumptions these theories cannot explain the particular pattern of probabilistic results found in this experiment.

Likewise the psychological theories which even break with any concept of normativity, such as the original pragmatic reasoning theory (Cheng & Holyoak, 1985) or the evolutionary social contract theory (Cosmides, 1989; Gigerenzer & Hug, 1992) cannot explain the fit of data to these normative models of reasoning.

The normative models as well as the empirical results of the experiment at least show the incompleteness of all these theories. This has to be said in such a cautious manner, since one has to concede that Bayesian models have not yet explained all the effects predicted by all these quite different theories either.

Implications for Bayesian Approaches

On the one hand the results of the present work show that the discussed Bayesian approaches of hypothesis testing (of single conditionals) need to be extended by a structural component, which determines what parameters are constant in that models.[6] On the other hand this extension (norma-

[6] The structural component proposed in this paper may be regarded as the *microstructure* of a conditional, which perhaps complements the effects of *macrostructure* already discussed in the context of causal Bayes-nets (cf. Waldmann & Hagmayer, 2001)

tively as well as empirically) strongly confirms the general approach of exactly these extended Bayesian models.

Working with a MST and by clearly fixing the preconditions, the results, not only of the model von Sydow (2002), Hattori (2002) and Oaksford & Wakefield (2003) *but also* the original model of Oaksford & Chater (1994, 1998, similarly now Over, in press, Evans et al., 2003) could be supported. (The evaluation of the model of Laming remained ambivalent.) Moreover the objection of Laming (1996) that the assumptions of the discussed basic models are licentious, which in principle affects all models, has been ruled out by introducing experimentally exactly the preconditions of these models.

But these largely confirmative results also show the necessity to extend Bayesian models discussed by the structural aspect examined. From this it results that it is false, both normatively and empirically, to assume that only *one* universal Bayesian model could and should fit all data. Also those authors who have adopted a probabilistic or Bayesian account, mostly still seek a universal model for hypothesis testing or reasoning with conditionals (e. g. Oaksford & Chater, 1994, 1998; also Oaksford & Wakefield, 2003; and even Evans et al., 2003 and Over, in press). Instead the results of my experiment show that additional hidden preconditions need to be taken into account. In this regard I do follow early writings of Evans & Over (1996), which stressed that there is no universal technical measure of uncertainty reduction. On the other hand, in my opinion, only the more sophisticated models in the tradition of Oaksford & Chater do allow a detailed investigation of the phenomena in question. This paper could be seen as contribution towards a synthesis of these positions.

On the larger scale such a synthesis would sustain normative necessity, as the logicistic research program also has done. Nevertheless it allows for a plurality of preconditions, which has been stressed by domain specific accounts. Whether *domain-specific normative Bayesian models* may serve as a more general research program can only be found out by further theoretical analysis and empirical investigation.

Acknowledgments

I am grateful to York Hagmayer, Michael Waldmann and Björn Meder for useful comments on the experiment and on earlier versions of this paper. I also would like to thank Nick Chater and two anonymous reviewers for their helpful comments.

References

Cheng, P. W., & Holyoak, K. J. (1985). Pragmatic reasoning schemas. *Cognitive Psychology, 17,* 391-416.

Cosmides, L. (1989). The logic of social exchange, *Cognition*, 31, 187-276.

Evans, J. St. B.T., Handley, S.H., & Over, D.E. (2003). *Conditionals and conditional probability*. *Journal of Experimental Psychology: Leaning, Memory, and Cognition*, *29*, 321-335.

Evans, J. St. B. T., & Over, D. E. (1996). Rationality in the selection task: Epistemic utility versus uncertainty reduction. *Psychological Review, 103*, 356-363.

Gigerenzer, G., & Hug, K. (1992). Domain-specific reasoning: Social contracts, cheating, and perspective change. *Cognition, 43*, 127-171.

Hattori, M. (2002). A quantitative model of optimal data selection in Wason's selection task. *The Quarterly Journal of Experimental Psychology, 55 A (4)*, 1241-1272.

Johnson-Laird, P.- N. & Byrne, R. (2002). Conditionals : A theory of meaning, pragmatics, and inference. *Psychological Review, 109 (4)*, 646-678.

Kirby, K. N. (1994). Probabilities and utilities of fictional outcomes in Wason's four-card selection task. *Cognition, 51*, 1-28.

Klauer, K. Ch. (1999). On the normative justification for information gain in Wason's selection task. *Psychological Review, 106, 1*, 215-222.

Oaksford, M., & Chater, N. (1994). A rational analysis of the selection task as optimal data selection. *Psychological Review, 101*, 608-631.

Oaksford, M., & Chater, N. (1996). Rational explanation of the selection task. *Psychological Review, 103*, 381-391.

Oaksford, M., & Chater, N. (1998). *Rationality in an uncertain world.* Psychology Press, Hove (GB).

Oaksford, M., & Chater, N. (2001). The probabilistic approach to human reasoning. *Trends in Cognitive Science, 5 (8)*, 349-357.

Oaksford, M., & Chater, N., Grainger, B. (1999). Probabilistic effects in data selection. *Thinking and Reasoning, 5 (3),* 193-243.

Oaksford, M., & Chater, N., Larkin, J. (2000). Probabilities and polarity biases in conditional inference. *Journal of Experimental Psychology. Leaning, Memory, and Cognition, 26, 4*, 883-899.

Oaksford, M., & Wakefield, M. (2003). Data selection and natural sampling, Probabilities Do Matter. *Memory and Cognition, 31,* 143-153.

Oberauer, K., Wilhelm, O., & Diaz, R.-R. (1999). Bayesian rationality for the Wason selection task? *Thinking and Reasoning. 5 (2),* 115-144.

Over, D. E. (in press). Naïve probability and its model theory. In V. Girotto & P.-N. Johnson-Laird (Eds.) *The Shape of Reason.* Hove: Psychology Press.

Osman, M., & Laming, D. (2001). Misinterpretation of conditional statements in Wason's selection task. *Psychological Research, 65,* 128-144.

Popper, K. R. (1934/2002) *Logik der Forschung.* Mohr Siebeck: Tübingen.

Sydow, Momme von (2002). *Probabilistisches Prüfen von wenn-dann-Hypothesen.* Diplomarbeit (MA-thesis), Department of Psychology, Universität Bonn.

Wason, P. C. (1966). Reasoning, 135-151. In B. M. Foss (Ed..), *New Horizons in Psychology*. Harmondsworth: Pinguin.

Waldmann, M. R., & Hagmayer, Y. (2001). Estimating causal strength: The role of structural knowledge and processing effort. *Cognition, 82,* 27-58.