Towards a Flexible Bayesian and Deontic Logic of Testing Descriptive and Prescriptive Rules

Explaining Content Effects in the Wason Selection Task

Dissertation

zur Erlangung des Doktorgrades der
Mathematisch-Naturwissenschaftlichen Fakultäten
der Georg-August-Universität Göttingen

vorgelegt von

Dr. Momme von Sydow (PhD)
aus Konstanz am Bodensee

Göttingen, 2006

D 7

Referent: Professor Dr. Michael R. Waldmann

Korreferentin: Professorin Dr. Margarete Boos

Tag der mündlichen Prüfung: 4. Mai 2006

„Cum deus calculat et cogitationem exercet, fit mundus“
(When God calculates and develops thought, he creates the world)
G. W. Leibniz, 1765 [1996, 25]

The Wason Selection Task is „probably the most intensive studied task in the psychology of reasoning […], which has raised more doubts over human rationality than any other psychological task“
M. Oaksford and N. Chater, 1998, 173, 174

The copyright of this thesis rests with the author.
(c) Momme v. Sydow, Göttingen, 2006

Abstract

Research on the Wason selection task (WST) has raised fundamental doubts about the rationality of human hypothesis testing and added to the development of both domain-specific and domain-general theories of reasoning. This work proposes a rational but domain-specific synthesis aimed at integrating converging lines of research in the WST debate. For this synthesis two realms are distinguished, that of testing descriptive rules (hypotheses) and that of testing prescriptive rules (prescripts). For both realms, accounts are proposed that have normative aspects, but also domain-specific aspects.

For the testing of descriptive hypotheses, a flexible Bayesian logic is developed, which is opposed to the falsificationist research program and builds on previous Bayesian accounts (on Oaksford and Chater, 1994, 2003, in particular). However, instead of advocating a universal Bayesian model a knowledge-based account is pursued which may explain the negative results of previous experiments. Additionally, the Bayesian treatment of conditionals is extended to other logical connectors.

For the testing of conditional prescripts, a flexible deontic logic is proposed which draws more fully on the deontic logic of prohibitions, obligations and permissions than previous accounts. Moreover, this is combined with a goal-based, but systematic, mechanism of cooperator and cheater detection.

Twelve experiments largely support the predictions of the proposed account. In two experiments the different strategies for testing descriptive or prescriptive rules were investigated. In seven experiments, positive support for the Bayesian account was obtained by actively introducing the preconditions of the models. Additionally, different Bayesian models of a conditional were distinguished and first evidence for a Bayesian logic of different connectors was obtained. In three experiments, the deontic logic of checking prescripts and its interaction with the goals of cheater or cooperator detection (including double foci) was investigated. The results cannot be explained by other current theories of the selection task, such as mental model theory, social contract theory, or pragmatic resoning theory, but do at least necessitate substantial extensions of these theories. In contrast, the results support the flexible Bayesian and deontic logic of testing descriptive and prescriptive rules.

Kurzreferat

Die Wasonsche Informationswahlaufgabe (WST) ließ fundamentale Zweifel an der Rationalität menschlichen Hypothesentestens aufkommen und inspirierte die Entwicklung sowohl domänenspezifischer als auch domänenübergreifender Theorien des Denkens. In dieser Arbeit wird ein Synthesevorschlag entwickelt, der auf der einen Seite domänenspezifisch ist, auf der anderen Seite aber einen rationalen Ansatz darstellt. Um Anomalien in der WST-Debatte erklären zu können, wird dabei zwischen dem Prüfen deskriptiver und präskriptiver Aussagen unterschieden.

Für das Prüfen deskriptiver Aussagen wird eine *flexible Bayessche Logik* entwickelt, die im Gegensatz zum falsifikationistischen Forschungsprogramm steht und auf vorangehende bayessche Ansätze (insbesondere auf Oaksford und Chater, 1994, 2003) aufbaut. Statt eines universellen Bayesschen Ansatzes wird ein wissensbasierter Ansatz vertreten, der die negativen Resultate früherer Forschung erklären könnte. Zudem wird der Bayessche Ansatz von Konditionalen auf andere logische Junktoren übertragen.

Für die Testung präskriptiver Konditionalaussagen wird eine *flexible deontische Logik* vorgeschlagen. Diese umfaßt eine deontischen Logik von Verpflichtungen, Erlaubnissen und auch Verboten, die mit einem zielabhängigen, aber systematischen, Mechanismus der *Cheater*- und *Cooperator-Detection* kombiniert wird.

In zwölf Experimenten konnten die meisten Vorhersagen des Ansatzes bestätigt werden. In zwei Experimenten wurde der Unterschied des Prüfens deskriptiver und präskriptiver Aussagen untersucht. In sieben Experimenten konnte der Bayessche Ansatz bestätigt werden, indem die Modellvoraussetzungen in der Instruktion absichtlich induziert wurden. Zudem wurden erste Evidenzen für unterschiedliche Effekte verschiedener Bayesianischer Modelle und für eine Bayessche Logik des Hypothesentestens vorgelegt. In drei Experimenten wurde die deontische Logik präskriptiver Aussagen und ihre Interaktion mit den Zielen *Cheater* oder *Cooperator Detection* (auch mit Doppelfokus) untersucht. Die Ergebnisse können nicht von anderen Theorien des WST (wie der Mental Model Theorie, der Social Contract Theory oder der Pragmatic Reasoning Schema Theorie) erklärt werden, sondern machen zumindest deren Erweiterung notwendig. Die Ergebnisse stützen den hier vertretenen Ansatz einer flexiblen bayesschen und deontischen Logik des Testens deskriptiver und präskriptiver Aussagen.

Table of Contents

0 Introduction

In the wake of logical atomism and falsificationism, psychology of hypothesis testing had been dominated by a rational and context-independent norm of reasoning based on predicate logic and falsificationism. Alternative approaches have been developed which have completely discarded a rational and systematic justification of reasoning and have instead advocated an adaptationist, modular and domain-specific understanding of rationality. As a contribution to a synthesis of these antagonistic positions, this work expounds and tests an account of rule testing, which is, on the one hand, domain-specific and knowledge-based, whilst, on the other, rational and systematic.

More specifically this work is concerned with the Wason Selection Task (WST), which has had a prominent impact on the rationality debate. The WST is one of the most widely investigated and theoretically influential tasks in the psychology of hypotheses testing. The task has been important both to the rise of domain-general as well as to the rise of domain-specific theories of the WST. The proposed rational but domain-specific synthesis aims to integrate converging lines of research in the WST debate, distinguishing between the realm of testing of descriptive rules (hypotheses) and the realm of testing of prescriptive rules (prescripts). For both fields the advocated proposals have normative and systematic aspects, but also domain-specific and knowledge-based aspects.

Part I introduces the WST and the main domain-general and domain-specific theories related to this task.

Part II is concerned with the testing of descriptive rules. A Bayesian theory of confirmation is advocated which breaks with falsificationism. The fundamental problem of induction and the paradox of the ravens are discussed. Based on the results of this discussion a knowledge-based Bayesian account of hypothesis testing is proposed. This advocated approach improves on and extends previous Bayesian accounts of the WST. Additionally, a more general Bayesian logic of the WST is proposed, extending both previous Bayesian and logical accounts of the WST alike.

Part III is concerned with the testing of prescriptive rules. A theory is proposed, which combines aspects of the logic of practical philosophy, deontic logic, and a goaldirected process of focusing. Deontic logic is normally understood as a general but realm-specific normative theory applicable only to prescriptive (deontic) sentences. This is combined with the claim that the goal of the tester, either aiming for cheater or cooperator detection, has a systematic influence on which cases are selected. It is claimed that the combination of these tenets can explain important findings in the WST and allows for novel predictions.

The advocated theories on descriptive and prescriptive WSTs are interpreted as syntheses of previous accounts. The novel predictions of the expounded theories, including, for instance, experiments on the proposed different test strategies either for descriptive or prescriptive WSTs have been tested in various experiments. Most experiments have corroborated the advocated synthesis.

In conclusion, a novel account of the testing of descriptive and prescriptive rules is elaborated in this work, covering important aspects of the WST debate. Whereas earlier psychological theories on the WST advocated an antagonism between universal-rational and domain-specific irrational accounts, the account supported here is rational but domain-specific and knowledge-based. Since the WST debate has been intimately connected to the more general rationality debate, this work is not only intended as a contribution to the WST debate but also as one to the larger rationality debate.

Part I Wason Selection Task and Theories of Hypothesis Testing

For the last four decades the Wason selection task, introduced by Peter Wason (1966, 1968), has presumably been the most studied and seminal task in the psychology of reasoning. The task has been called "one of the most extensively used paradigms in human experimental psychology" (Manktelow & Over 1990, 153), the "most investigated logical reasoning problem in the psychological literature" (O'Brien, 1995, 189) or "probably the most intensive studied task in the psychology of reasoning" (Oaksford & Chater, 1998, 173). The task became notorious for raising doubts on the rational behaviour of humans in testing hypotheses. As a result, the WST played an important role for the development of domain-general and domain-specific psychological theories of rationality.

Before introducing, elaborating and testing my own proposals in Part II and Part III of this thesis, the current section introduces the WST and the main theories in the field. It consists of two chapters: Chapter 1 explains the logic of the WST and its falsificationist standard solution. Chapter 2 provides an overview of the WST debate and the most important theories in this field.

1 The Wason Selection Task and the Falsificationist Logic of Hypothesis Testing

1.1 The Structure of the Wason Selection Task

The Wason selection task (WST) is a hypothesis-testing task, which has been closely linked to the investigation of reasoning and logic. In the WST, the truth or falsity of a given hypothesis (the theoretical world) is to be tested against an empirical world made up of four cards (the empirical world).

The hypothesis is normally a simple logical sentence made out of two atomic propositions (*p* and *q*) and a logical (dyadic) connector linking these propositions (cf. Table 1 later). Typically, the WST has been concerned with a conditional of the form "if *p* then (always) *q*". This conditional can be a thematic sentence, like "if a bird is a raven then it is black”, or an abstract sentence, like “if a card has an ‘A’ on one side, then it has a ‘2’ on the other side”. The earliest WSTs were conducted with the latter abstract letter-number hypotheses (Wason, 1966, 145-147, 1968; Wason & Shapiro 1971; cf. Johnson-Laird & Wason, 1972).

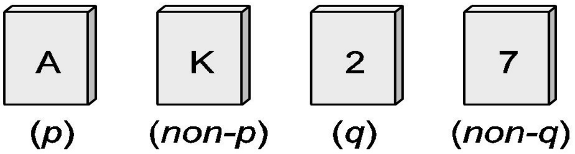


Figure 1. An example for the used four cards in Wason’s original letter-number selection task

The visible front sides of the four cards represent examples for all logical categories mentioned in the conditional: *p, non-p, q, and non-q*. In a letter-number WST these cases are, for instance, ‘A’, ‘K’, ‘2’, and ‘7’ (cf. Figure 1). As it is known in a letter-number WST that on one side of each card there are letters and on the other numbers, it is generally known in all WSTs that on one side of each card there is a *p* or a *non-p,* and on the other side a *q* or a *non-q*.

The task of the participants is to select those card(s), which they would turn over in order to test the truth or falsity of the rule.

Before the falsificationist solution to the task is considered, the WST is briefly discussed in relation to other tasks in order to understand its place in the debate on reasoning:

The WST and related tasks. The WST has been developed in the context of the debate on reasoning and the WST involves drawing conclusions. Nonetheless, the WST obviously differs from propositional or syllogistic reasoning tasks (for an overview see, e.g., Hussy, 1986; Eysenck & Kean, 1995; Waldmann & von Sydow, in press). In these ‘conclusion drawing tasks’ the premises, like “if a bird is a raven then it is black” (if *p* then *q*) and “this is a raven” (*p*), have to be assumed to be valid and the task is to draw a conclusion from these premises; here it follows by *Modus ponens,* $(((p \rightarrow q) \wedge p) \Rightarrow q)$, that the bird is black (*q*). In contrast, in the WST normally the truth or falsity of a general sentence is exactly what is in question.

The WST also needs to be delineated from other types of hypothesis testing tasks. Another prominent hypothesis-testing task is the 2-4-6 task, which has also been proposed by Wason (1960). In this task participants actively have to formulate a hypothesis and test possible hypotheses successively against data (cf., e.g., Klayman & Ha, 1987; Jonas, Schulz-Hardt, & Frey, 2001). Thus, the 2-4-6 task is a *hypothesis-identification* task. In contrast, the WST is a *hypothesis-testing* task, in which participants test the truth or falsity of a *given* hypothesis, presented by the experimenter.

Finally, the WST can also be distinguished from relatively similar *truth table tasks*, in which single evidences or, more often, a whole set of evidences, like black ravens, white swans, black shoes etc. are presented to the participants, and in which the participants have to come up with a hypothesis (productive truth table task) or in which they have to answer whether a given evidence is coherent with a hypothesis (evaluative truth table task) (e.g., Wason, 1968, 278; van Duyne, 1976; Gebauer & Laming, 1997; McKenzie & Mikkelsen, 2000; Evans, Handley, Over, 2003; Barres & Johnson-Laird, 2003; cf. also Kao & Wassermann, 1993; White, 2000). In contrast, subjects in a WST never actually get evidences, but they are asked which card they hypothetically would turn over in order to check the hypothesis. Hence, the WST investigates the active selection of information, not the passive evaluation of given evidence.

1.2 The Traditional Solution for the WST: Formal Logic and Falsificationism

The traditional normative solution of the WST is based on the logical interpretation of the hypothesis combined with a falsificationist norm of hypothesis testing.

The Logical Interpretation of Hypotheses

The customary conditional hypothesis in the WST, ‘if *p* then *q*’, has traditionally been interpreted in the wake of logicism as a material implication of formal logic. According to modern propositional logic (G. Frege, 1879; L. Wittgenstein, 1922; A. N. Whitehead & B. Russell, 1925) logical connectors can be defined by truth tables (for an introduction see, e.g., F. von Kutschera, A. Breitkopf, 1992). Some truth tables of basic connectors of two-valued propositional logic are shown in Table 1. According to these definitions, the truth or falsity of a sentence is solely determined by the logical form of the sentence – the logical connector(s) – and the truth or falsity of the elementary propositions *p* and *q*. A sentence based on an if-then connector, $p \rightarrow q$ (subjunction), is false only if *p* is true and *q* is false at the same time; it is true only if the antecedent *p* is false or the consequent *q* is true (Table 1a).

Table 1a, b, c, d

Truth Table Definitions of Four Dyadic Connectives of Formal Propositional Logic

p, q combinations		(a) Subjunction, implication, conditional	(b) Equivalence, biconditional	(c) Conjunction	(d) Adjunction, inclusive disjunction
p	*q*	*If p then q*	*Only if p then q*	*p and q*	*p or q*
		$p \rightarrow q$	$p \leftrightarrow q$	$p \wedge q$	$p \vee q$
True	True	True	True	True	True
True	False	False	False	False	True
False	True	True	False	False	True
False	True	True	True	False	False

The four combinations of True and False in the two columns for *p* and *q* on the left exclusively and exhaustively represent the four logically possible states of the world. 16 binary connectors are possible. In the four columns a, b, c, d on the right four basic connectives are defined.

Formal two-valued propositional logic is not limited to implications (subjunctions) but provides a system of theoretically 16 connectors (cf. Chapter 7). Only the most important four connectors are defined in Table 1. In the case of the adjunction (inclusive disjunction), for instance, a sentence ‘*p* or *q*’ is true if *p* is true, or *q* is true or both are true; it is false only if *p and q* are false (Table 1d). The other examples of connectives show that hypothesis interpretation based on formal logic is a more general conception, not only relevant for conditionals.

Additionally, one monadic connective, the negation, symbolized by '¬', will be used, applicable even to a single atomic sentence. The negation changes the truth function from true to false, and from false to true.

Only few WSTs were concerned not with conditionals, but with other connectives, such as for instance adjunctions (e.g., Wason & Johnson-Laird, 1969). Subsequently only conditional hypotheses will be considered, only later on other connectives will be discussed as well (see Chapter 7).

The above logical interpretation of a conditional as subjunction is not applicable if the hypothesis is probabilistic or if it is reinterpreted as a biconditional 'if and only if' statement. To exclude the former meaning in a WST, one can formulate the if-then sentence explicitly in a deterministic way: 'if *p* then *always q*'. To exclude the latter meaning one needs to use appropriate material (or, as done here, consider frequencies which exclude a biconditional interpretation).

The Falsificationist Norm of Testing If-Then Hypotheses

(a) K. R. Popper's (1902-1994) falsificationism provides a universal norm for rational hypothesis testing (Popper, 1934/1994/2002, 1972, 1974, 1977, 1996). This norm is mainly based on two interconnected arguments, the asymmetry of falsification and verification, and a particular interpretation of Hume's problem of induction. Here I only sketch these arguments (see pp. 26 f. for more details).

Firstly, there is an asymmetry between falsification and verification. A universal if-then proposition can logically never be verified – a disconfirming case may always appear in the future. In contrast, the truth of a deterministic conditional can be logically falsified by a single case only.

Secondly, the Humean problem of induction sheds doubt on any conception of confirmation. A repeated confirmatory outcome, like the repeated rise of the sun, mathematically allows to infer on a high probability of further confirming outcomes (the sun will probably rise tomorrow) only if the constancy of nature is assumed. But this seems to beg the question. Based on this problem Popper argued that we have to replace the conception of confirmation by a completely negative falsificationist methodology. Later we will discuss Hume's problem of induction in depth, showing that also falsificationism falls pray to this problem and that another solution is needed see pp. 23 f.).

For these arguments, falsificationism postulates that rational hypothesis testing can never be a search for confirmations but needs to be a search for falsifications only.

(b) Falsificationism, if applied to the WST, demands that one should invariably test exactly all those cards which may reveal falsifying instances.

From the beginning of the WST research tradition, falsificationism has been the yardstick with which to evaluate the correctness of answers in the WSTs (Wason, 1966; Johnson-Laird & Wason, 1970a, 138; cf. Johnson-Laird & Wason, 1977, and today, e. g., Stanowich & West, 2000). For instance, Wason (1968, 273) argued, "a valid inference depends crucially upon the possibility of meeting the falsifying contingency", and Johnson-Laird and Wason (1970a, 136) explicitly referred to Popper's methodology (1959, cf.: 1934) as the norm for correct answers in the WST. Falsification for a long time became the undisputed criterion by which to judge the adequacy of selections in the WST. Van Dyne (1976, 85) consequently even proposed calling the task 'conditional falsification problem'.

Table 2
Truth Table of a Material Implication (Subjunction)

p	q	$p \rightarrow q$
True	True	True
True	False	False
False	True	True
False	True	True

Note: The falsification case is darkened.

As we have seen the conditional hypothesis "if *p* then (always) *q*" is only false, if *p* is true and *q* is false at the same time (cf. Table 2). Therefore, falsificationism demands for all WSTs with a conditional hypothesis that participants select exactly the *p* and the *non-q* card, since these two cards can each lead to falsificatory evidence. This normative demand of falsificationism is independent of the content of the conditional. For instance, in the case of the mentioned letter-number WST, with the hypothesis "if a card has an 'A' on one side, then it has a '2' on the other side", falsificationism demands the selection of the 'A' card and the '7' card (cf. Figure 1).

1.3 Early Results and Doubts Over Human Rationality – The Two Anomalies of the WST

From the beginning of the WST research, the empirical results have shown that most participants in standard WSTs strongly deviate from the norm of falsificationism. As a result, the WST has raised more doubts about human rationality than presumably any other psychological task. For our later discussion, I want to distinguish two anomalies for the logical-falsificationist research program: firstly, the predominance of confirmatory *p* and *q* selections in many WSTs, and, secondly, content effects.

The First Anomaly: The Predominance of Confirmatory Selection Patterns

The application of the logical-falsificationist norm to the WST led to devastating results for human rationality. Even the first WST conducted showed that people are not able to adhere to the rational standard of logic combined with falsificationism. Wason (1966, 146-147) found in his first letter-number WST an illogical 'verification bias', which he had found earlier also in the 2-4-6-task (Wason, 1960).

Johnson-Laird and Wason (1970a) summarised the results of earlier WSTs they had conducted (with letter-number material). If one also interprets incomplete *p* selections to be false, the rate of false selections sums up to 96 % (Figure 2)!

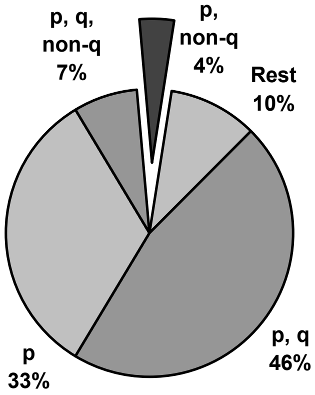


Figure 2. Pie chart of the percentage of selection patterns in Johnson-Laird and Wason (1970a).

Early 'therapy' experiments involved procedures to help subjects to understand where they were going wrong in order to elicit a correct falsificatory *p* and *non-q* selection strategy (Wason 1968, 1969; Wason & Golding, 1974; cf. Cheng, Holyoak, Nisbett, & Oliver, 1986). Although strong manipulations in so-called 'therapy WSTs' in some cases improved the performance of the participants, it turned out to be quite difficult to educate subjects to become falsificationists. Normally the majority did not adopt a falsificatory test strategy. Such results fostered doubts about human rationality raised by the earlier WSTs.

Research has shown that people encounter extraordinary difficulty in finding the solution to the WST dictated by propositional logic and falsificationism. Most subjects, at least in standard letter-number WSTs, have adopted a confirmatory *p* and *q* test strategy. Nonetheless, falsificationism remained completely unchallenged up to the mid 1980s – and until today falsificationism has remained the dominant normative theory of the WST.

The Second Anomaly: Content Effects

Content effects constitute another class of phenomena, which are problematic for a logical-falsificationist understanding of the WST.

It was shown in the 1970s that *thematic* WSTs, opposed to abstract letter-number WSTs, can facilitate a logical solution of the task (Wason & Shapiro, 1971; Johnson-Laird, Legrenzi, & Legrenzi, 1972). First it seemed that 'realistic', 'thematic' or 'concrete' content generally improves the performance in WSTs (thematic content hypothesis). There was a debate on whether experience with the rule or availability of knowledge in a domain was necessary to elicit the falsificationist solution.[1]

Later it was shown that certain thematic rules, such as "If I eat haddock, I drink gin," do not enhance the performance (Manktelow & Evans, 1979). On the other hand most social rules, such as "If a person is drinking beer, then the person must be of full age", enhance the performance (cf. Johnson-Laird, Legrenzi & Legrenzi, 1972; Griggs & Cox, 1982; but cf. also Wason & Shapiro 1971).

Facilitating or aggravating content effects (the latter will be outlined when introducing domain-specific theories) constitute the second central anomaly of the WST – at least from a purely falsificationist and logical perspective. They show that the testing of hypotheses is not purely formal and independent of the content, as demanded by formal logic and falsificationism.

[1] I refrain from discussing availability theories, like the memory cueing hypothesis (Cox & Griggs, 1982; Griggs & Cox, 1982). Today they are mostly regarded to be refuted by results of the later domain-specific theories (Cheng & Holyoak, 1985; Cosmides, 1989; Gigerenzer, 1992). However, I think that some results for instance of Cox & Griggs (1982) would be perfectly compatible with a recent dual source theory distinguishing form and content competence (cf., e.g., Beller & Spada, 2003). A review on the availability theories is given by Griggs, 1983 (cf. Jackson & Griggs, 1990).

2 The Main Psychological Theories of the WST: A Normative or a Domain-specific Approach?

The deviations from the norms of logic and falsificationism contributed to the formulation of psychological theories of the WST. The psychological accounts of the WST, which will be outlined here, fall into two classes:

(a) Domain-general psychological approaches of the WST on the one hand maintain the logical and falsificationist norm of hypothesis testing, but on the other hand postulate more simple psychological mechanisms, to account for deviations found. As a psychological mechanism either an incomplete set of reasoning rules of a ‘natural logic’ (mental logic theory) or incomplete representations of the task (mental model theory) have been postulated as being responsible for reasoning biases found in the WST and other tasks.

(b) In contrast, domain-specific approaches completely abandoned logic and falsificationism as a rational and general basis of hypothesis testing. Instead, domain-specific approaches have postulated that reasoning schemas (pragmatic reasoning schema theory) or evolutionary and modular Darwinian algorithms (social contract theory) are needed to explain content effects. The postulated learned schemas or evolved modules need not refer to any rational norm of reasoning and may even be opposed to formal logic.

Here only a sketch of the main theories of the WST will be given in order to provide an introduction to the WST debate. Other theories will be considered later (see the introductions and discussions of Part II and Part III).

2.1 Mental Logic Theory

Proponents of mental logic theory (ML theory, Braine, 1978; Rips, 1994; O’Brien, 1995) continue to advocate universal and abstract laws of thinking. According to ML theory, the found empirical deviations from logic are mainly due to a mentally given ‘natural logic’ which is assumed to have a more limited basis than formal logic (Braine, 1978, p. 18). Natural logic is not based on truth tables but on inference rules. This approach goes back to Gentzen’s (1935) system of natural deduction.

According to ML theory, a conditional sentence for instance is normally interpreted as a *Modus ponens* rule of inference. The number of mental inference rules is lower than of logically valid inferences, but the mental rules are never illogical. For instance, ML theory postulates that there is a mental *Modus ponens* rule but no mental *Modus tollens* rule. Hence, common sense reasoning that would also need to use the *Modus tollens* needs to be implemented by the use of a number of other inference rules. This explains why the *Modus tollens* is more difficult than the *Modus ponens*. Nonetheless, natural logic and formal propositional logic would come to the same results if the difficulties played no role: “natural and standard propositional logic are the same system on different foundations“.

Rips (1994, 179 f.) explains the standard errors in the WST by incomplete representations of reasoning rules. Because a *Modus tollens* rule is psychologically not available, it is difficult to see that *non-q* cases may lead to a falsification of the if-then clause. (Additionally, Rips assumes that subjects may have problems to represent the backsides of the cards.)

O’Brien (1995) showed in detail that, based on the postulated natural logic, many steps are needed to prove that only a *p* and *non-q* card pattern is a correct solution. He concludes that the WST should not count as a proper field of application for ML theory. I cannot discuss here whether this argument may have been an immunisation strategy, since ML theory indeed has problems to account for many of the findings in the WST debate. In any case, later we will discuss ML theory only briefly.

2.2 Mental Model Theory

Mental model theory (MM theory) likewise maintains the logical-falsificationist norm of hypothesis testing, but explains the found deviations not by incomplete inference rules but by incomplete representations.

MM theory was founded by Johnson-Laird (1983). In this work, I discuss the updated MM theory by Johnson-Laird & Byrne (1991, 1992, 1995, and 2002)[2]. Although MM theory was also built on findings for other tasks, particularly propositional reasoning tasks, MM theory was explicitly formulated to account also for the WST.

According to MM theory, a representation of the tested hypothesis is constructed, consisting of so-called ‘mental models’. This construction is based on the interpretation of tested sentence, context or task description and hence on syntactic,

[2] In the Discussion of Part III the modified MM theory of deontic syllogistic reasoning by Bucciarelli and Johnson-Laird (2005) will also be applied to the WST.

semantic and pragmatic knowledge. The core of the MM theory of the WST gives an account of how conditionals and other connectors are normally represented. Mental models represent possible states of affairs, which are also the logical cases in a truth table (cf. Table 1).

However, in contrast to formal logic, MM theory can explain deviations from logic by postulating that the possibilities of a truth table are mentally represented in an incomplete way. The logical representations are in two respects incomplete:

Firstly, MM theory generally postulates a preferred representation of true states of affairs. This has been called the principle of truth and has been postulated for all logical connectors (Johnson-Laird & Byrne, 2002). False cases are normally represented only with some effort.

Secondly, MM theory made additional specific proposals with regard to conditionals, also to account for anomalies in the WST (see 1.3). MM theory supposes that the true cases of a conditional too are normally not represented completely. The assumed incomplete representation of a standard conditional (Johnson-Laird & Byrne, 1991, 79) is shown in Table 3. Only *p* and *q* cases are represented, together just with a mental footnote that *p* is represented completely and with an ellipsis '…' indicating the possibility of further models.[3]

Table 3
Incomplete and Complete Mental Models of an Implication "if p *then* q*"*

(a) Incomplete Mental Model	(b) Complete Mental Model With Explicitly Represented False Case
[p] *q*	*p q*
…	*p non-q* (mental footnote: false case)
	non-p q
	non-p non-q

Note. The brackets in the incomplete mental model stand for a mental footnote that the *p* cases are represented completely.

If subjects represent the tested conditional in the above incomplete way (Table 3a), they should not select the correct *p* and *non-q* pattern, but *p* cases only. According to MM theory, correct *p* and *non-q* selections are only elicited if the representation is 'fleshed out' including an explicit representation of the false case:

[3] The formulation of the incomplete mental model of a conditional is similar to an earlier idea of Wason (1966, 146; cf. 1968, 274): „Subjects assume implicitly that a conditional statement has not two truth values, but three: true, false and 'irrelevant'. Vowels with even numbers verify, vowels with odd numbers falsify and consonants with any number are irrelevant."

> "In short, the model theory predicts that people will select the card falsifying the consequent whenever the models are fleshed out with explicit representations of that card" (1991, 80). Any "experimental manipulation that lead reasoners to flesh out their models of conditional, and, in particular, to construct an explicit model of an appropriate counterexample, should enhance performance in the selection task“ (Johnson-Laird & Byrne, 1995, 346).

To account also for the found *p* & *q* selections, Johnson-Laird and Byrne (1991, 80) argued that conditionals may also be interpreted as statements of equivalence (Table 1b, p. 5). For this case they introduced an incomplete model for an equivalence, which is similar to that of the implication, only with the additional mental footnote that *q* is represented completely as well (cf. Johnson-Laird & Byrne, 1992, 1995; more generally on deontic WSTs see the Discussion of Part III).

2.3 Pragmatic Reasoning Schema Theory

Cheng and Holyoak (1985, cf. 1989; Holyoak & Cheng, 1995a, 1995b) were the first to abandon normative logic and falsificationism as the psychological basis for testing rules in the WST. Instead, they proposed a domain-specific pragmatic reasoning schema theory (PRS theory). According to PRS theory, content effects were due to specific reasoning schemas, which are linked to goals and which are based on abstractions of recurring experiences in society. It is argued that these schemas do not always enhance logical selection patterns for conditionals; they may in principle also trigger illogical ones (Cheng & Holyoak, 1985, 397; see Holyoak & Cheng, 1995b).

For the social realm, they explicitly proposed specific production rules for a permission schema and an obligation schema (Holyoak & Cheng 1985; Cheng, Holyoak, Nisbett, & Oliver, 1986; cf. Politzer & Nguyen-Xuan, 1992; Holyoak & Cheng, 1995) and provided first empirical support for these two schemas. Both schemas are characterised by four production rules. The permission schema[4] has been defined by the following four production rules (P1-P4, Cheng & Holyoak, 1985, 397):

P1: If the action is to be taken, then the precondition must be satisfied.
P2: If the action is not to be taken, then the precondition need not be satisfied.
P3: If the precondition is satisfied, then the action may be taken.
P4: If the precondition is not satisfied, then the action must not be taken.

[4] There has been confusion whether the original permission schema of Cheng & Holyoak (1985) should not better be called obligation schema (Manktelow & Over, 1990, cf. Oaksford & Chater, 1998a, 206).

Later PRS theory provided also a more explicit formulation of an obligation schema, which consists of the rules O1-O4 (Cheng, Holyoak, Nisbett & Oliver, 1986; cf. Politzer & Nguyen-Xuan, 1992; Holyoak & Cheng, 1995a, cf. also 1995b):

O1: If the precondition is satisfied, then the action must be taken.

O2: If the precondition is not satisfied, then the action need not be taken.

O3: If the action is to be taken, then the precondition may have been satisfied.

O4: If the action is not to be taken, then the precondition must not have been satisfied.

Cheng and Holyoak (1985) have additionally mentioned a schema for causality and for covariance that should both lead to *p* and *q* selections in the WST. However, they have not elaborated this aspect of their PRS theory of the WST (however, cf. e.g. Cheng, 1997).

Empirically, Cheng and Holyoak (1985) showed that an activation of such a schema, without direct previous experience, is enough to facilitate the performance in the WST, resulting in clear-cut *p* and *non-q* patterns. The results of Cheng and Holyoak (1985; cf. Gigerenzer, 1992) contested the availability theories, (e.g., Cox & Griggs, 1982; cf. p. 9, footnote 1) which were taken to have postulated that facilitation in the WST, is caused by direct former experience with the rule. In contrast, Cheng and Holyoak elicited facilitation effects by activating a schema in a situation in which subjects had no previous experience of the rule. Although Cheng and Holyoak (1985, 397) have mentioned that illogical results may also be elicited by these schemas, only after Cosmides (1989) presented her illogical approach, Politzer & Nguyen-Xuan (1992) showed that pragmatic schemas can also elicit illogical *non-p* and *q* selection patterns.

In other articles, pragmatic reasoning schema theory has been vindicated against both mental model theory and evolutionary social contract theory (Cosmides, 1989; Cheng & Holyoak, 1989; Jackson & Griggs, 1990; Kroger, Cheng, & Holyoak, 1993). (Cf. also the general discussion of Part III.)

2.4 Social Contract Theory

The domain-specific social contract theory proposed by Cosmides and colleagues was even more pronounced in abandoning any normative logical basis for reasoning in the

WST (Cosmides, 1989; Cosmides & Tooby, 1992; cf. Cummins, 2000; Fiddick, Cosmides, & Tooby, 2000). Instead Cosmides (1989; Cosmides & Tooby, 1992) based her social contract theory (SC theory) on evolutionary considerations (particularly on Trivers, 1971; Dawkins, 1982; Axelrod & Hamilton, 1981) and postulates that the clear-cut 'correct' *p* and *non-q* selection patterns observed so far, are not based on a facilitation of logic, but on an activation of a cheater detection module, which is understood as a specific evolutionary adaptation.

Cosmides and Tooby (1992) argued that there was a strong selection pressure on individuals in Pleistocene groups of hunters and gatherers to develop a cheater detection module, since only then these individuals can sustain cooperative reciprocal exchange in their own interest without being exploited. Cosmides and collegues have discarded selection for the good of the group, but have stressed that the evolution of a cheater detection mechanism nonetheless would have been strongly advantageous on the level of self-interested individuals.

Cosmides (1989, 200) claimed straightforwardly that no thematic rule that is not a social contract and which does not involve cheater detection has ever produced a content effect that is both robust and replicable. She argued that the activation of the postulated 'Darwinian algorithm' of cheater detection in a social cost-benefit context and not the facilitation of logic has caused the found clear-cut selection patterns.

Empirically, Cosmides in WSTs with standard social contracts (e.g., "If I give you $20, you give me your watch") provided evidence for clear *p* & *non-q* selection patterns – still coherent with logic. But additionally Cosmides showed for switched social contracts ("If you give me your watch, I'll give you $20") opposed illogical *non-p* & *q* patterns. Cosmides followed from this that participants do not select logically correct patterns but cheater detection patterns that are in their own interest. Consequently, the results for switched social contracts were taken as evidence against an approach based on formal logic (but cf. e.g., Johnson-Laird & Byrne, 1992).

Additionally, Cosmides' experiments corroborated her prediction that the goal of altruist detection does not elicit clear-cut selection patterns. (She postulated on evolutionary grounds that there was no selection pressure to develop an altruist detection module, since there has never been – according to Cosmides and particular evolutionary approaches – any evolutionary stable altruism. Cf. Cosmides & Tooby, 1992, pp. 193-197.)

The evolutionary approach of Cosmides has not remained unchallenged. Critics have pointed out that, for example, her assumption that cheater detection is based on an innate and specialised module can neither be warranted by evolutionary theory nor by the empirical findings (Cheng & Holyoak, 1989; Pollard, 1990; Lloyd, 1999; Sperber & Girotto, 2002, 2003; cf. v. Sydow, 2001; for more details cf. General Discussion of Part III).

Gigerenzer and Hug (1992) have substantially extended and modified social contract theory. Firstly, they showed that the perspectives into which participants can be cued, and the corresponding individual interests, strongly influence the found selection patterns (cf. earlier: Manktelow & Over, 1991). It is predicted that participants, again based on their individualist interests and perspectives, only tested those cases, by which cheating of the other party could be detected. The adaptationist interpretation of perspective effects was an important contribution inspiring the research program of Gerd Gigerenzer's research group at the MPI in Berlin, which has explicitly abandoned a normative theory of rationality and instead advocated a bounded rationality approach, based on specific adaptations (e.g., Gigerenzer & Goldstein, 1996; Gigerenzer & Selten, 2001). The human mind is seen as a "toolbox" equipped with specific modules adapted to solve specific problems (cf. Gigerenzer, Todd, & the ABC Research Group, 1999). Perspective effects became the topic of a vigorous debate, in which many authors in fact have interpreted perspective effects from a different perspective than Gigerenzer and Hug: Johnson-Laird and Byrne (1991, 78-79), Manktelow and Over (1991), Johnson-Laird and Byrne (1992), Politzer and Nguyen-Xuan (1992), Johnson-Laird and Byrne (1995), Holyoak and Cheng (1995a, 1995b), Liberman and Klar (1996), Dove (1996), Erdfelder and Dove (1997), Fairly, Manktelow, and Over (1999), Almor and Sloman (2000), Staller, Sloman and Ben-Zeev (2000), and, finally, Beller and Spada (2003).

Secondly, Gigerenzer and Hug (1992) dissociated the concepts of a social contract and of cheater detection. They could show that social contract rules alone are not sufficient for obtaining Cosmides' results. Therefore, they concluded "the crucial issue about social contracts is the cheating option" (p. 165; cf. also Cosmides & Tooby, 1992).

Some authors, however, interpret the phenomena discussed by the 'cheater detection approach' not in a domain-specific, but a domain-general way (e.g., Almor

& Sloman, 1996, 2000; Fairley, Manktelow, & Over, 1999; Johnson-Laird & Byrne, 1992; Liberman & Klar, 1996; Staller, Sloman, & Ben-Zeev, 2000; Manktelow & Fairley, 2000; Sperber, Cara, and Girotto, 1995; Sperber, 2002, 2003). Fiddick, Cosmides, and Tooby (2000) and Fiddick (2004) in turn have defended a domain-specific approach to reasoning, not based on a universal rational norm, but on specific and modular adaptations (cf. similarly: Hiraishi & Hasegawa, 2001).

2.5 Relevance Theory

One of the contributions to the WST debate, which was partly formulated in reaction to SC theory, is *relevance theory* (Sperber, Cara, and Girotto, 1995; Sperber, & Girotto, 2002, 2003; cf. Fiddick, 2004). Here it should be briefly introduced, particularly since I am going to make use of one aspect of this theory in Part III. Sperber et al. (1995) have proposed that relevance theory (Sperber & Wilson, 1986) can account for attentional changes in the WST. In a domain-general proposal, the WST is interpreted not as a deductive task or a hypothesis-testing task, but as a task to check potentially relevant evidence (cf. Evans, 1994). The findings of Cosmides are claimed to be due to changes in the relevance of cases. However, relevance, of course, needs to be defined, since also SC theory, a Bayesian approach (cf. Part II), or a falsificationist approach all may be interpreted as making claims about relevance. According to Sperber et al. (1995), cards lead to unreflective intuitions of relevance, if they are connected with a high 'cognitive effect' and a low 'cognitive effort'. In my view, this is still close to a tautology, as long as 'cognitive effect' and 'cognitive effort' are not defined more precisely (cf. Discussion Part III, pp. 281 f.). However, Sperber et al. (1995) provided interesting results, which showed that a variation of 'relevance' changes the selections in different tasks. Nevertheless, Oaksford and Chater (1995a) reinterpreted those findings and suggested that their own Bayesian approach (Oaksford and Chater, 1994) provides a more formal measure of relevance. I agree with Oaksford and Chater on this matter. Moreover, Oaksford et al. pointed out that relevance theory, formulated in terms of cognitive effort and effect, has no immediate application to purely probabilistic manipulations as used in their descriptive WSTs (e.g., Oaksford, Chater, Grainger, Larkin, 1997; see Part II, pp. 71, and, particularly, pp. 177 f. for details). More recently Sperber and Girotto (2002, 2003) have provided evidence, which is connected with what I will call 'focus effects', without linking this phenomena to deontic logic (in Part III; see particularly the discussion of relevance theory, p. 281).

In regard to focus effects, I will build on the work of Sperber and colleagues. To focus on particular cases is, in my view, a standard way of checking prescriptive tasks. In these tasks we do not aim to test the truth or falsity of a hypothesis, but, for instance in the role of a police officer, aim to find and punish only those who deviate from a rule. However, I will distinguish focus effects from normal hypothesis testing effects in the Bayesian sense. Moreover, in Part III focus effects will be combined with a deontic logic of testing prescriptive conditionals.

2.6 A Remark on the Research Agenda of Part II and Part III

In Part I, the WST and the main domain-general and domain-specific theories of the WST were introduced. In the following two main parts, my own approach and its more specific background will be layed out. Part II is concerned with the standard testing of hypotheses (descriptive rules), whereas Part III is concerned with the standard testing of prescripts (prescriptive rules). The dichotomy of descriptive and prescriptive rules will be discussed and directly tested (see Section 9.1, Chapter 10). In Part II a Bayesian logic of hypothesis testing is proposed; in Part III a deontic logic of testing prescripts is advocated. In both parts, the proposed positions extend previous accounts. Both combine normative aspects with the flexible use of additional knowledge. In Part II it is advocated that the construction of the Bayesian models depend on the given situation. In Part III, deontic logic is combined with an additional goal dependent focus of cheater and cooperator detection. It will be argued that the advocated accounts on the WST in the descriptive and the prescriptive realms steer a middle course between traditional domain-general and domain-specific approaches. In both main parts, several experiments will be presented to test the advocated accounts.

Part II Towards a Flexible Bayesian Logic of Testing Descriptive Rules

Part II is concerned with the testing of descriptive rules in WSTs. As an alternative to a falsificationist position or to giving up a normative approach altogether, I advocate a Bayesian account of hypothesis testing. More specifically, building on previous Bayesian accounts and I will here advocate a more refined knowledge-based Bayesian account. A knowledge-based Bayesian account can resolve fundamental philosophical problems connected with the problem of induction and the WST debate. According to this position, the Bayesian model of the testing conditionals depends on additional assumptions and knowledge about the situation in place. Since former Bayesian approaches did not achieve good empirical support, it is hoped that the current knowledge-based approach may improve this situation. Additionally, the Bayesian account of the WST, which has been limited to conditionals, will be extended to other connectors.

Outline of Part II. In Chapter 3 a knowledge-based Bayesian account will be developed as a solution to two philosophical problems of induction: Hume's fundamental problem of induction and Hempel's paradox of the ravens. The standard falsificationist norm of hypothesis testing has explicitly been based on these two problems of induction and any normative alternative to falsificationism needs to solve these problems before becoming applicable as a rational standard to a specific task like the WST. Although the Bayesian philosophy of science has proposed a solution to the paradox of the ravens, it as far as I know has not provided a solution to Hume's more fundamental problem of induction. It will be argued that not only naïve inductionism but also falsificationism falls pray to Hume's problem, and that only a knowledge-based account may provide a way out of the abyss of irrationalism and scepticism. In respect to the paradox of the ravens, it is shown that a knowledge-based extension of the standard Bayesian solution can provide not only one but two (or even several) resolutions of the ravens paradox.

In Chapter 4 the basic idea of the Bayesian approach to the WST, which has first been fully developed by Oaksford and Chater (1994, 1998), will be presented. Here also criticism of this universal Bayesian model of testing a conditional will be reviewed, particularly the criticism of Laming (1996).

In Chapter 5 the knowledge-based modification of the Bayesian account of the WST is proposed. Firstly, an alternative model of the WST is outlined, which was first investigated by von Sydow (2002, cf. Oaksford & Wakefield, 2003), and which makes the problematic prediction of *p* versus *non-p* frequency effects. To establish the preconditions for this model a many cards selection task (MST) is introduced. Experiment 1 is for the first time concerned with MSTs with material from the paradox of the ravens. Consistent with the advocated knowledge-based account the preconditions for the Bayesian model used are explicitly introduced. The experiment tests whether under these circumstances the Bayesian resolution of the paradox of the ravens can be corroborated empirically. In Experiment 2 an MST with the traditional letter-number material is conducted. Using the more adequate MSTs, it is aimed to achieve better results than found in former experiments by other authors.

In Chapter 6 it is argued that not only the model tested in Chapter 5 can be supported empirically, if its preconditions are given. In Experiment 3, three alternative models are tested, again securing their preconditions. This provides a direct test of the proposed knowledge-based approach. Further research in this field is needed and it is discussed that even more models can be imagined. A Bayesian model is proposed which even makes equivalent predictions to the traditional norm of logic combined with falsificationism.

In Chapter 7 the Bayesian account of the WST, which has been limited to the test of conditionals, is here for the first time extended to other connectors. This can be seen as a first step towards a Bayesian logic of hypothesis testing, which transcends both, traditional Bayesian approaches and approaches only based on propositional or predicate logic. The predictions of this *Bayesian* logic of hypothesis are tested in a number of MSTs.

In Chapter 8 the results of Part II are discussed, particularly with respect to the alternative theories of the WST.

Editorial note on Part II: In the Chapters 3, 4, and 5 I have developed and extended thoughts that are in part directly based on the work of von Sydow (2002). Together with a previous experiment, Experiment 1 has been presented on a conference (von Sydow, 2004b). Parts of Chapter 7 including Experiment 3 have been presented on three conferences (von Sydow, 2004a, 2004c, von Sydow, Hagmayer, Meder, & Waldmann, 2005).

3 Philosophical Considerations: The Problem of Induction and a Knowledge-Based Bayesian Account

The vigorous empirical debate on hypothesis testing in the WST in important aspects epitomizes the modern psychological version of the old philosophical debate on induction. Moreover, here it will be shown that from the philosophical debate lessons can be learned for the WST debate on induction as well.

In this chapter we will first discuss the fundamental problem of induction (Aristotle, 350 BC; Hume, 1739) and then the more specific problem of induction which has been called Raven's Paradox (Hempel, 1945).[5]

For the WST as a task of hypothesis testing the philosophical problems of induction are of importance for two main reasons. Firstly, the falsificationist norm of hypothesis testing, predominant in the WST debate, is philosophically based on these problems. Anyone who criticises the norm of falsificationism in hypothesis testing also needs to address the problem of induction, in order to allow for an alternative norm. Secondly, anyone who positively supports a rational concept of confirmation, like a normative Bayesian account of hypothesis testing, needs to address these problems of induction in the first place. Nonetheless, the Bayesian account of confirmation, which I will persecute in this work, normally simply ignores the first problem of induction, Hume's problem. Moreover, the standard Bayesian resolution of second problem of induction, the raven's paradox, in my view remains inconclusive. Hence, here both problems of induction will be addressed.

In the section on the fundamental problem of induction, I will first outline the Humean problem itself. Then we come to Popper's falsificationist negative solution of that problem. I will then argue that, if Humean scepticism is taken seriously, falsificationism itself falls pray to Hume's problem. A resulting sceptical conclusion would actually continue the old tradition of scepticism about the *mundus sensiblis*, the empirical world, present up to today: "What is empirical is not certain. What is certain is not empirical" (Einstein). Although I will support the claim that confirmatory and disconfirmatory empirical knowledge is always fallible, I am going to argue that this

[5] In the second half of the 20th century another specific problem of induction has been discussed, Goodman's (1955/1963) grue paradox. Since that paradox is less directly related to the current work, I here only refer to a Bayesian resolution of this paradox (Rosenkrantz, 1982).

needs not render any justification of hypotheses testing irrational. A knowledge-based (and only partly self-referential) inductive justification of inductive inferences will be proposed as a possible solution to the fundamental problem of induction. A ‘synductive’ understanding of empirical knowledge acquisition is advocated, which steers a middle course between a bottom-up and a top-down approach. This synductive approach to induction will also be pursued later in the WST debate.

In the second section of this chapter, the raven’s paradox will be discussed. Firstly, the paradox will be presented, likewise arising from any naïve concept of induction. Secondly, it will be argued that falsificationism, which has also been claimed to provide a resolution of this paradox, itself falls pray to this paradox. Thirdly, the standard Bayesian solution of the paradox is described, which has some synductive aspects. Finally, after pointing out some problems of the standard Bayesian resolution of the paradox, an even more synductive or knowledge-based Bayesian resolution of this paradox will be advocated.

The reader who is mainly interested in the psychological WST debate may skip this Chapter, but basically the discussion of these problems of induction is indispensable if one aims to favour a Bayesian norm of hypothesis over a falsificationist norm, providing us with the philosophical basis for the later knowledge-based modification of Oaksford and Chater’s (1994, 1998, 2003) standard Bayesian theory of the WST.

3.1 The Fundamental Problem of Induction

Is the common-sense belief that the sun will rise tomorrow justifiable at least to some probabilistic degree, and is this belief more justified than the opposed assumption that the sun will not rise tomorrow? Hume’s formulation of the fundamental problem of induction made it apparent that all such inferences from known facts to general theories (and then again to specific predictions) may be no more than arbitrary and irrational habits. Before discussing this ‘scandal of modern philosophy’ and some possibilities for escaping this problem, we first explicate the notion of ‘induction’.

The Notion of Induction

‘Induction’ (lat.: *inductio*) is derived from Cicero’s translation of the Aristotelian term ‘™pagwg»’ (*epagoge*). Despite changes in the meaning of epagoge or induction, throughout the history of thought these concepts were always used to describe the (rational) inference from the particular to the general (for a historical overview, see Lumer, 1990, 660 f.).

Beside this characterisation of induction as an inference from the particular to the general, induction is often characterised in opposition to deduction as a generalisation which is insecure (excluding what has been called exhaustive or mathematical induction[6]). Consistent with this usage, we are here only concerned with insecure empirical induction. *Empirical* induction is concerned with the derivation, confirmation or disconfirmation of general hypotheses or theories based directly on perceptions, sense data or empirical protocol sentences.

Philosophically the notion of induction has a connotation of pure empiricism although moderate rationalists, of course, have also used induction. For our discussion of the problems of induction and my later proposal of the term ‘synduction’, it is helpful to spell out this connotation in more detail. In the wake of empiricism and logical positivism induction has often been seen, firstly, as exclusively a bottom-up process leading from the empirical to the theoretical. Secondly, it has been seen as what one may call an ‘external-to-internal-process’ leading from external facts to internal theory. Extreme empiricism can be understood as the position that the mind is originally a blank tablet (*tabula rasa*) and that any knowledge it possesses is imprinted on it directly by the senses. According to empiricism, theories are to be tested against the reality itself and this testing should ideally not be affected by any prior knowledge. In this perspective, induction has often been seen as a neutral and passive process based on ‘given’ sense data. Knowledge should be acquired, but any use of knowledge to mould (or improve) the inductive process of acquiring new knowledge is seen generally as a distortion of a neutral, dispassionate and context free process of passively receiving information from the senses.

Hume’s Fundamental Problem of Induction

Tentatively, we can characterize the fundamental problem of induction by the argument that any general hypothesis, theory or law, which transcends past experience and makes future predictions, can never be logically derived from observations

[6] Historically, an insecure generalisation has been called *induction probabilis* (Albertus Magnus), opposed to the complete and necessary generalisation called *inductio perfecta*. Today *inductio perfecta* is often seen as a particular case of deduction, a secure inference based on the exhaustive observation of all cases or on mathematical (complete) induction.

– exactly because it transcends past experience. Therefore, any justification of an inductive or generalising inference may appear problematic.

The fundamental problem of induction is sometimes attributed to the 18th century philosopher David Hume. However, the problem is as old as the reflection on the notion of induction itself. The problem of induction took centre stage in many classical treatments of epistemology, methodology and metaphysics, starting with Aristotle's *Organon* (about 350 BC, e.g., *Topica*, I, pp. 8, 103, *Analytica Posteriora,* I, pp. 18, 31) and it provided a recurrent central theme in the history of Western philosophy.[7]

However, the most influential formulation of the problem of induction in modern philosophy is indeed found in David Hume's *A Treatise of Human Nature* (1739, Book I, Part III). The original goal of the empiricist Hume was to show that knowledge about causal laws – which are understood as necessary and deterministic connections – could be rationally based on empirical grounds of sensory experiences. Nonetheless, his problem of induction made him prominent as a gravedigger of a rational justification of empiricism.

According to Hume, any universal causal law of nature can only be based on the constant conjunction, in all past impressions. But according to Hume we can never justify the assumption of a resulting necessary connection of ideas: "From the mere repetition of any past impression, even to infinity, there never will arise any new original idea, such as that of a necessary connexion; and the number of impressions has in this case no more effect than if we confin'd ourselves to one only" (Part III, Section 6, p. 88).

Hume's original statement of the problem of induction actually affects not only the induction of a general law but also the induction of *each* causal observation connecting an entity A and an entity B, since each such observation makes an inference that these entities and no other entities were causally connected. To assess whether an observed entity A really caused entity B, we, for instance, need to know that the effect has not coincidentally been caused by some hidden cause. No causal relation can directly be observed. We only had impressions of object A being located left or right from B, but we never had a direct impression of A causing B. To attribute that there was a causal connection between cause and effect we need to know which entities are candidates for such a connection in the first place.

[7] Also, for instance, pre-Humean rationalism depended on the fundamental problem of induction. Leibniz (c. 1704) wrote: "The senses, although they are necessary for all our actual knowledge, are not sufficient to give us the whole of it, since the senses never give anything but instances, that is to say particular or individual truths. Now all the instances which confirm a general truth, however numerous they may be, are not sufficient to establish the universal necessity of this same truth, for it does not follow that what happened before will happen in the same way again. […] From which it appears that necessary truths, such as we find in pure mathematics, and particularly in arithmetic and geometry, must have principles whose proof does not depend on instances, nor consequently on the testimony of the senses, although without the senses it would never have occurred to us to think of them […]" (Preface, pp. 150-151).

But to limit our observations to particular causally relevant cases begs the question what is causally relevant (Part III, Section 3, cf. Section 6) and leads back to the problem of generalisation.

To solve the problem of generalisation one would have to beg the question by illegitimately assuming a constancy of nature: “there can be no demonstrative arguments to prove that those instances, of which we have had no experience, resemble those, of which we have had experiences. We can at least conceive a change in the course of nature […]” (Part III, Section VI, p. 89, cf. p. 91).

Hume considered the objection that a necessary (deterministic) conjunction of objects might be a too strict criterion for induction. Consequently, Hume also discussed probabilistic cause and effect relations and confirming or disconfirming evidence (Part III, Section 12, cf. Section 6 and 11). Hume for probabilistic laws likewise stressed that any calculation of probabilities is “founded on the transferring of past to future” (p. 137), and that “the supposition, that the future resembles the past, is not founded on arguments of any kind, but is deriv’d entirely from habit […]” (p. 134; see also p. 139).

Finally, Hume argued that any inductive justification of induction has to be excluded because this would necessarily be circular (Part III, Section 6, pp. 90-91; we will come back to this later).

Based on this fundamental problem the empiricist Hume was led to take a sceptical stance on empirical knowledge. Although induction is ubiquitously used in human reasoning and science, Hume regards induction to be nothing but an irrational useful habit, which in his view can neither be justified nor overcome.

Hence, Hume’s position was alleged to promote scepticism and irrationalism. For instance, Russell wrote (1991/1961, 645) that Hume’s philosophy represents “the bankruptcy of eighteenth-century reasonableness”. If there would be no resolution of the problem of induction, “there is no intellectual difference between sanity and insanity. The lunatic who believes that he is a poached egg is to be condemned solely on the ground that he is in a minority […]” (646). Hume’s formulation of the fundamental problem of induction had an influence on modern philosophy which cannot be overestimated. Even Kant (1781/87), who did not aim to abandon the enlightenment project, wrote that Hume’s scepticism awoke him from his ‘dogmatic slumbers’.

However, Kant, Mill, Cassirer, Russell, Carnap, Reichenbach and Popper all developed positions to overcome this problem and to justify induction. Since I cannot provide a full history of the fundamental problem of induction here, I will directly discuss Popper's falsificationism, which – as we have seen – has also been crucial for the psychology of hypothesis testing.

Falsificationism as Proposal to Solve the Problem of Induction

Popper's (1934, 1972, 1974, 1996) falsificationist methodology directly builds on Hume's fundamental problem of induction (e.g., Popper, 1934, Chapter 1, cf. Chapter 10; 1972, Chapter 1 and 2). Hume showed that induction is not rationally justified, but he still regarded it as practically indispensable. Popper follows Hume in the assessment that induction is irrational, but he consistently discards induction and replaces it by a negative methodology exclusively based on falsification (1934, 14, 226; 1972, Chapter 1; cf. Quine, 1974).

Popper reformulated Hume's problem of induction based on formal logic (1972, 7, 12, 20; 1934, 45). Popper connects Hume's problem of induction to the argument that there is a logical asymmetry of verification (total confirmation) and falsification (total disconfirmation) of single universal statements (hypothesis, laws or theories). A universal hypothesis can be falsified by observations, but it can never be verified. Universal deterministic hypothesis H can logically be falsified by one single observed counterexample ($\neg O$), which by Modus tollens proves the hypothesis false: $H \rightarrow O \wedge \neg O \Rightarrow \neg H$. In contrast, even a large number of 'confirming' observations never allows a verification of a hypothesis. It always remains possible that future evidence will falsify a highly 'confirmed' hypothesis. Popper argued that the only methodologically correct way out of the problem of induction is to abandon induction. According to Popper, one should only be concerned with the falsification of theories, never with their confirmation.

Popper, in *Objective Knowledge, an Evolutionary Approach* (1972) additionally links his falsificationism with a Darwinian metaphysics. On his account any process of knowledge acquisition from amoeba to Einstein is always concerned with blind conjectures and external refutations, similar to mutation and selection in biology and trial and error in psychology of learning (for a detailed discussion of this interesting position see v. Sydow, 2001).

The resulting falsificationist methodology of hypothesis testing became the normative standard in the psychology of reasoning, particularly in the research on the WST (cf. Chapter 1, 2).

In the next section, falsificationism will be criticised as not resolving Hume's actual fundamental problem but as falling pray to the problem itself. Subsequently, and as a way out, I try to expose an alternative knowledge-based solution of the problem of induction.

Falsificationism Falsified? – The Fundamental Problem of Falsification

The negative 'solution' of the problem of induction by falsificationism, which completely discards the concept of induction by confirmation, can be criticised. Firstly, I will illustrate how the logical asymmetry, giving rise to falsificationism, is to be restricted to a certain domain of assertations and, secondly, that falsificationism, in my opinion, does not resolve the fundamental problem of induction, by limiting itself to what one may call 'disconfirmatory induction'.

Criticism of Falsificationism Restricting its Domain of Applicability

One objection to a falsificationism as a *general* methodology is that the logical asymmetry of falsification and verification is not valid for all kinds of logical propositions:

Firstly, existential statements, like 'genetic engineering has produced a red raven', can even never be falsified and only be verified (Quine, 1974; cf. Popper 1934, 40). Moreover, sentences like 'at least *n* objects of *p* are *q*' can only be verified and not falsified.

Secondly, Quine (1974) conceded that the asymmetry is valid for single universal statements with a single quantification, but he objected that the asymmetry does not equally hold for statements that are more complex. Universal statements that are quantified in multiple ways cannot be falsified in an unambiguous way. Moreover, scientific theories have normally a complex structure and they are quantified in multiple ways. For these logical reasons, Quine (1974) concluded that Popper's negative methodology cannot be a general methodology of science, but only a methodology for single universal statements (with one quantification). This is connected to the next point.

Thirdly, according to the Quine-Duhem problem of falsification (and confirmation), an empirical hypothesis (H_T) is never tested alone; it is always tested together with some auxiliary hypothesis (H_A) for instance concerning the process of measurement etc. Hence, an observed falsification ($\neg O$) needs not to be attributed to falsity of that theory – it may equally be attributed to falsity of only the auxiliary hypothesis: $H_T \wedge H_A \rightarrow O \wedge \neg O \Rightarrow \neg H_T \vee H_A$ (cf., e.g., Chalmers, 2001, 74 f.).

Fourthly, the asymmetry is not valid for probabilistic laws, for which we would predict only a certain amount of correct observations. Hence, falsificationism cannot be applied in a straightforward way to probabilistic hypotheses and it should be noted, that almost all psychological hypotheses belong to this class. Likewise, if one aims to test a deterministic functional law, one generally makes use of a region of tolerance, which renders a strict application of a falsificationist methodology impossible.[8]

Fifthly, Putnam (1974) and Lakatos (1974) have pointed out that in the history of science theories were almost never discarded because of a single recalcitrant falsification. Moreover, it would have always been disadvantageous to discard them on this basis since almost all valid theories would have been prematurely discarded (cf. also Chalmers, 2001, 76). Similarly, the Göttingen scientist and aphorist, Georg Lichtenberg (1742-1799) wittily pointed out much earlier: “One should not take note of contradictory experiences until there are enough of them to make constructing a new system worthwhile.”

For these reasons, falsificationism cannot count as a methodology for science with general applicability. But this does not entitle us to conclude that falsification is not the correct norm of hypothesis testing if we have to test a simple universal and deterministic hypothesis and if the auxiliary hypotheses can be neglected. Because we will be concerned with such hypothesis, I outline an even more fundamental criticism in order to discard falsificationism.

The Fundamental Problem of Falsification

In my opinion, falsificationism does not solve Hume’s problem of how to transfer past knowledge to the future, and hence falsificationism itself falls pray to this problem. First the transferability of falsifications from past to future will be discussed, before the focus is switched to Popper’s theory of corroboration.

[8] The theory of statistical tests by Sir Ronald A. Fisher (1890-1962) is sometimes interpreted as a falsificationist approach in the broad sense, since it is concerned with the *refutation* of a null hypothesis only. One may argue that this refined ‘falsificationism’ can obviously handle probabilistic hypotheses. However, alternatively one may argue that Fisher’s theory is fundamentally one-sided and the theory of Jerzy Neyman (1894-1981) and Egon Pearson (1895-1980) provides a preferable and a more balanced approach considering both kinds of possible decision errors (cf. Hager, 2004).

(a) Opposing Popper, one may apply Hume's sceptical argument not only to confirmations but also to falsifications of a universal hypothesis: a falsification implies nothing about whether the hypothesis will be false *in respect of future instances* or not. Hume himself (1739, Book I, Part III, Section 12, pp. 134 f.) discussed evidence which is contrary to a tested hypothesis. To use disconfirming and falsifying evidence for future predictions makes again the illegitimate assumption that "instances, of which we have no experience, must necessarily resemble those of which we have" (p. 135). Popper seems to ignore this point, but the problem of applying past knowledge to the future is equally valid for falsifications.

For example, let us assume that two million years ago a hypothetical scientist put forward the optimistic hypothesis that "all adult and living members of the genus *Homo* have a brain size of over 1000 cm^3". Our hypothetical scientist would have no problem to falsify this hypothesis, since it is generally accepted today that presumably every member of the genus *Homo* at that time would have falsified this claim. Today it is assumed that the *Homo habilis* had a brain size of about 650 cm^3 (varying between 500 and 800 cm^3). Nonetheless, the claim has become true today: *Homo sapiens sapiens* has a brain size of about 1350 cm^3 (varying between 1300 and 1500 cm^3). One may object that the use of a different hypothesis, a hypothesis without time restriction, would allow us to argue that logically the hypothesis has nonetheless been falsified and remains false. This is correct. However, would we have learned anything from a falsification used only in this way? According to this usage, falsifications would only restate a particular past (negative) observation, without learning *anything* for predicting or explaining the future. In this sense falsificationism does not solve Hume's problem of transferring past experiences to the future in a rational way.

(b) Falsificationism – like inductionism – obviously needs to provide some measure which of two or more competing theories (none of which have been falsified) should be preferred. If no such measure had been proposed, falsificationism would not be able to distinguish well-established theories (say the theory of quantum mechanics) from other theories that – ceteris paribus – have not been tested at all. In my view, two opposed positions on this matter can be found in Popper's writings, both of which lead to fundamental problems.

On the one hand, Popper (1934, 1972) advocated that one should prefer the theory that has been corroborated more frequently and that has a higher resulting 'verismilitude'. A theory is corroborated if it has been tested and has not been falsified in this test. Putnam (1974, 222-223) and Lakatos (1974, 256, 261) have argued that Popper's theory of corroboration is a pseudo-theory of confirmation. This theory still transfered degrees of corroboration from the past to the future. But since Popper claims to follow Hume in not allowing for such an inference his approach is inconsistent.

On the other hand, in other passages Popper (1972, 18-19; cf. 22) makes clear that different degrees of corroboration also do not provide a basis upon which to make predictions. The degree of corroboration would be a "report of past performance only", saying "nothing whatever about future performance, or about 'reliability' of a theory" (p. 18). But why then should we use Popper's degree of corroboration to assess competing theories? If the degree of corroboration says nothing about the probability of a theory being corroborated in future events, it would be absurd to use it as a criterion to distinguish between theories. Putnam (1974) criticised this in a clear way: if there "were no suggestion at all that a law which has withstood severe tests is likely to withstand further tests", no theory would be more confirmed than any other, and "science would be a wholly unimportant activity". A Popperian may perhaps object that past corroborations are still relevant to describe our past experience. Nevertheless, according to Popper's radical Humean account the description of the past is completely useless for most purposes, since it conveys no knowledge at all to access the future. With regard to the future there would be no reason at all to keep track of past corroborations and falsifications. Hence, the application of Hume's problem to Popper's position renders past knowledge – even knowledge about falsifications – completely useless, if it is to be applied to the future. We have again derived at what I want to call the 'fundamental problem of falsificationism' or the 'fundamental falsificationist problem of knowledge'. It follows from Popper's falsificationism that science and all empirical knowledge are completely useless for the prediction of any future event. To put it bluntly, falsificationism implies that it would make no difference if we forgot all our knowledge about the world and the sciences – our predictions and resulting actions would be no more irrational than they are now. But this consequence is no less absurd and paradoxical than Hume's original formulation of the problem of induction.

Hence, falsificationism provides no cure against Humean irrationalism, but it is rather to be seen as an expression of this irrationalist position. In conclusion, to allow for a rational account of knowledge acquisition a positive solution of Hume's problem of induction remains indispensable.

Broad's Contribution

Drawing on the early contribution of C. D. Broad (1887-1971) it will first be made clear that any inductive generalisation needs to have a *probabilistic* character, but that this is not sufficient to solve the problem of induction. Secondly, I will briefly discuss Broad's alternative resolution of the problem of induction, which is based on the concept of substance. Subsequently, I will expound my own knowledge-based proposal.

The Probabilistic Nature of Induction

Hume and Popper both emphasised that a universal hypothesis can never be completely verified. C. D. Broad as early as 1918/1920 pointed out that any positive account of induction needs to formulate less ambitious conclusions. A conclusion like 'all *S*'s are *P*' indeed could never be justified by cases of *S* being *P*. But, he continues, suppose "the conclusion becomes: It is highly probable according to the observed data that all *S*'s are *P*. There is then no illicit process. We argue from a *certain* proposition about some *S*'s to the *probability* of a proposition about all *S*'s. This is perfectly legitimate" (p. 3).

According to such a line of argument, it may appear legitimate to use probability theory or inductive logic as basis of a theory of induction (cf. e. g. Carnap, Reichenbach, or Hintikka). For instance, the sunrise problem, mentioned in the beginning of this Chapter, in the 18th century was treated by Pierre-Simon Laplace's probabilistic rule of succession. However, independent of the mathematical formulation of probabilistic theory of induction we face a more fundamental problem.

The Unresolved Problem of the Assumed Uniformity of Nature

Although an application of probabilistic reasoning indeed appears to be a necessary precondition to solve the problem of induction, this is clearly not sufficient to solve it. An application of the probability calculus to empiric matters itself assumes the uniformity of nature. This has been accurately noted by Broad: "the degree of belief which we actually attach to the conclusions of well-established inductions cannot be justified by any known principle of probability, unless some further premise about the physical world be assumed“ (1918, p. 1). This brings us back to Hume's problem: "there can be no demonstrative arguments to prove that those instances, of which we have had no experience, resemble those, of which we have had experiences. We can at least conceive a change in the course of nature […]" (Part III, Section VI, p. 89, cf. p. 91).

Please note, the uniformity assumption is not a feature unique to temporal inferences from the past to the future; it is in my view equally required in spatial and in conceptual respects. For example, we presume a special uniformity assumption if we reason from the colour of the (present) ravens in England to the (present) ravens in Scotland, and we are perhaps less convinced of this assumption if we reasoned to the ravens in Africa. Hence, there is not only an unjustified uniformity assumption with regard to temporal inferences from the past to the future, but also with regard to other conceptual aspects that can be used to differentiate between objects.

Discussion of Broad's Essentialist Solution of the Problem

Broad (1920) proposed overcoming the unresolved fundamental problem of induction by simply *assuming* that nature is uniform or constant, at least at some level of description. Broad argued that unchanging substances or essences need to be postulated in order to allow for rational induction and hence for empirical knowledge. One may accept the price of essentialism to escape Humean scepticism, but this proposal in my view is neither necessary nor sufficient to overcome Hume's problem.

Firstly, from an *a priori* postulation of the existence of substances one does not know *what* entities have to be seen as substances and when actually to assume uniformity. Hence, the required constancy assumptions, needed to justify induction, cannot be derived from the abstract postulation of substances or essences. An essentialist position is not sufficient to justify induction.

Secondly, Broad's essentialist proposal is, in my view, also not a necessary precondition of any rational concept of knowledge acquisition. My alternative proposal to the problem of induction, which will be laid out in the next section, does not rely on essentialism.

Synduction – Knowledge *and* Induction

Here a knowledge-based solution to the fundamental problem of induction will be proposed. In contrast to the claim of naïve empiricism that any use of knowledge in the objective and context free process of induction can only distort this process, I am going to argue that a knowledge-based understanding of induction provides a last resort against rendering induction irrational. Philosophically the advocated position would allow us to steer a middle course between radical empiricism (e.g., Locke) and radical rationalism (e.g. Leibniz).

Is There an Inductive Solution of the Problem of Induction?

Falsificationism and Broad's essentialist position cannot solve the fundamental problem of induction, because they do not provide reasons to justify any specific assumption about the uniformity of nature, which would be necessary for any inductive inference concerned with that particular case. The only way which, in my opinion, remains open, is to justify the preconditions of an inductive inference using knowledge that is itself based on induction. An inductive theory of induction is needed.

I will start with a simple example to illustrate that at least actual human inferences may often be understood to be knowledge-based. From the black colour of ravens in England, we probabilistically infer that the ravens in Scotland are probably black as well. However, since we generally know that the colours of animals differ in different climate regions, we would be more cautious about betting on whether all ravens in Africa are black as well (actually in Africa there are species of non-black ravens or crows: *Corvus ruficollis*, *Corvus albicollis,* and *Corvus albus*).

However, an inductive justification of induction is methodologically not unproblematic and also a tentative proposal of this possibility needs to be expounded in more detail.

From a Neo-Kantian perspective, the philosopher and psychologist Benno Erdmann (founder of the Department of Psychology in Bonn) advocated a similar position in his *Logik* (1892). An inductive justification of induction on some level abandons the distinction of the result of knowledge acquisition and its precondition. Erdmann noted that such a justification is circular (1892, § 91, pp. 550-568) and this leads us back to Hume's objection: "The same principle cannot be both the cause and effect of another" (Hume, 1739, Book I, Part III, Section 6, p. 90). However, Erdmann (1892, § 91, p. 566) rejects this objection and argued in a Neo-Kantian manner that the principle of induction is a necessary precondition for any knowledge acquisition and hence needs to be postulated *a priori*. He compared this to logic, which can only be justified based on logic, and needs to be postulated as an *a priori* truth necessary for rational reasoning.

In my view, any argument in favour of an inductive justification of induction needs to distinguish between two levels of argumentation. We will first discuss the justification of the general principle of induction and then, in a second step, a justification of specific inductions.

The General Principle of Induction and Its Modification

One may try to justify the *general principle of induction* in two ways, an empirical way and a mathematical (a priori) way. I think the *a priori* path is indeed more promising with respect to the *general* principle of induction, but this leads to a modified principle, which additionally requires a justification of specific inductions.

(a) An *empirical justification* of the *general* principle of induction is problematic. Chalmers (2001), for example, briefly criticised any empirical justification of the principle of induction. The basic principle may be formulated as the general claim that confirming evidence supports a hypothesis. Chalmers (2001, p. 43) argues that an inductive justification of the principle of induction is the inference from the observation “the principle of induction was successful in case 1” and the observation “the principle of induction was successful in case 2” to the conclusion “the principle of induction is always successful”.

In my view, this formulation is inappropriate, since it even renders the general principle of induction more implausible than it actually is. Firstly, following Broad’s (1920) probabilistic view, confirmatory observations can only induce an enhanced *probability* of the confirmed hypothesis. Secondly, we obviously know of instances where induction (and hence the principle of induction) failed. Hence, we have to add, for instance, the following premise to the above premises: “the principle of induction was unsuccessful in case 3”. Because of this, one can empirically at best infer the more moderate conclusion that “with a probability *x* the principle of induction has a certain probability *y* to be successful” (leading to a probability distribution of probabilities of successfulness).

Although such a revised argument would improve the empirical justification of induction presented by Chalmers, it is not only still prone to the criticism of circularity (cf. Chalmers, 2001), which will be tried to resolve subsequently, but also leads to a much too general conclusion. My main criticism would be that it is applicable in an undifferentiated way to any object. To apply induction with the same probability to every instance of induction is absurd. It becomes impossible to distinguish between the two probabilities either of reasoning from black ravens in England to those in Scotland or to those in Africa. Hence, an empirical justification of the *general* principle of induction leads to absurd results or needs to be formulated differently.

(b) Alternatively, I think one may justify a modified principle of induction on purely rational grounds (and hence perhaps also on empirical grounds). Erdmann (1892) argued in a Kantian way that the principle of induction is a *necessary* means of rationally acquiring empirical knowledge and hence a truth a priori. Although my argument is not opposed to Kantian credentials, it does not rely on them. The probability calculus is a mathematical truth, but it is applicable to the world only under the condition that its axioms are true. (This is, of course, equally true for formal logic.) The probabilistic inferences about the empirical world are based on the precondition that the axioms of the probability calculus are applicable. Although the Kolmogorov axioms of the probability calculus do not even explicitly mention this, they clearly presuppose the constancy of probabilities, and, applied to the empirical world they assume the constancy of nature. This, for instance, becomes apparent in the statistical law of large numbers, postulating that the average of a random sample from a population will converge at the average of that population. Of course, if the probabilities in the population change, the law of great numbers is not applicable (violating the assumption of a true random sample).

However, it is a general truth a priori that "if the constancy of probabilities can be assumed, one can rationally apply the principle of induction based on the probabilistic calculus". I think we should make use of this true principle to overcome the problem of induction, although this principle is not a *universal* principle, but a *conditional* principle (valid if its premises are given). I will call this the general 'principle of conditional induction'. It is plausible that this idea may also be extended to probabilistic knowledge about the validity of preconditions, implying a 'principle of probabilistic conditional induction': "if the constancy of probabilities can be assumed with a probability *x,* the principle of induction based on the probabilistic calculus is applicable with exactly this probability."

Firstly, these principles of conditional induction are still general principles – although based on conditions, they represent a general truth. Secondly, the principles of conditional induction are here formulated for the probability calculus, but it should be noted that analogous principles might be formulated for other axiomatic systems that make a constancy assumption as well. Thirdly, since rational induction can only be justified in terms that are conditional to certain preconditions, this account requires that these preconditions could be justified for specific cases of induction in a way that is not completely circular. Thus, we now turn to discuss the justification of single cases of induction.

Knowledge-Based Justification of Specific Inductions

I now argue that *single cases of induction* can be rationally based on knowledge (about its preconditions), which is itself based on induction, without rendering this process a vicious circle (*circulus vitiosus*). We may test the preconditions of a particular induction in a way that is not identical to the induction in question. Such a test would not beg the question, but would appear to allow for a knowledge-based concept of induction.

Let us look at a prototypical example, the prediction of throwing of a die several times. Assume that a die has been thrown 6000 times and the ‘6’ has been rolled in roughly 1000 cases (as supposed for a uniformly distributed die). It appears warranted to bet that if we roll the die another 6000 times that this will yield roughly the same portion of one sixth of the cases (assuming a region of acceptance between, say, 500 to 1500 cases). According to probability theory, this outcome is indeed highly probable – but this textbook example is only applicable to the empirical world if we assume that the probability of rolling a ‘6’ has remained constant in between the two games. This makes this inductive inference dependent on a variety of possible physical and even psychological assumptions, for instance whether another player may have manipulated the die. If another player were suspected to have loaded the die one would of course not bet that the previously calculated probability still holds. To test this we may actually role the die another 6000 times. Nonetheless, this does not resolve the problem of circularity, since the prediction would here be replaced by a direct evidence exactly of the predicted event. However, alternatively one may also test the loading of the die in different ways and then bet on the outcome of throwing it. For instance, one may check whether the die has a strictly cubical shape or whether the material used for the die is transformable or magnetic. The test of these properties may use all sorts of technical devices (e.g., X-rays) and again refers to further assumptions about the physics of throwing a dice and the physics of the technical devices. What is important here is that the preconditions for the inductive probabilistic inference can be measured in a way, which is not identical to the inferred prediction.

This knowledge-based approach is coherent with the intuition we have that inferring from black ravens in England that ravens in Scotland are also black has a higher probability than inferring that ravens in Africa are black too. Even without any knowledge about ravens in Scotland and in Africa, one may be able to make use of common implicit or explicit theories about the plumage and fur of animals in different climate regions. Of course, the probabilistic conclusions of an inductive inference can always turn out to be false, firstly, because the conclusion is by definition probabilistic and, secondly, the assumed constancy (and inconstancy) assumptions may have been false. In this sense all our empirical knowledge remains completely fallible. However, *given* any knowledge regarding constancy assumptions, it would be irrational not to apply them as thoroughly as possible to enable an induction that is as rational as possible. It will be shown that this is consistent with the general idea of Bayesian induction (cf. also the General Discussion of Part IV). From the advocated perspective the processs of induction itself is not static but improvable exactly because it is knowledge-based.

I have mentioned the use of additional theories external to the inductive inference in question. This has been treated in psychology (cf. Holland, Holyoak, Nisbett, & Thagard, 1986) under a set of different headings, for instant analogical reasoning (e.g., Holyoak & Thagard, 1995), causal learning (e.g., Waldmann, 1996) and category-based induction (e.g., Osherson, Smith, Wilkie, López & Shafir, 1990; cf. generally: Lien & Cheng, 2000; Waldmann & Hagmayer, in press). But here I cannot go into this related psychological literature (for a general overview of these fields, cf. e. g. Waldmann & Sydow, in press; on induction cf. Westermann & Gerjets, 1994; Lütkemeier, Westermann, Gerjets, 2003).

However, it should be noted that although, for instance, analogical inferences may often appear to be more problematic than empirical generalisations about throwing a die, the proposed account would treat both of them in principle under the same heading of knowledge-based inductions.

It has been shown in this section that the preconditions for induction may be justified in an inductive way without being fully circular. In the advocated view the same premises are not at the same time cause and effect of another (as Hume warned us); instead one induction relies on knowledge and, hence, on previous *other* inductions (confirmations *and* disconfirmations). In a larger context such an account of induction may indeed lead to circular dependencies, but as this circularity is always partial, I would rather call it systemic, reflective, or self-referential. In this sense, it has been made plausible that the circular aspect needs not to be a vicious circle, but may well be a self-corrective virtuosic circle (*circulus virtuosus*) which is increasingly based on empirical feedback. Further clarifications of this position are needed, but here it has been shown that a knowledge-based account of induction is a plausible candidate to overcome Hume's fundamental problem of induction.

Synduction and the Philosophy of the Knowledge-Based Account

The advocated knowledge-based resolution of the fundamental problem of induction is opposed to the radical irrationalism concerning confirmation (and as we have seen also concerning disconfirmation) found in Hume's writings. The advocated knowledge-based account of induction is neither a purely bottom-up account nor a top-down account, but it integrates both aspects. Hence, it appears false, to understand this resolution of the problem of induction as a mere rehabilitation of naïve empirism. This approach may indeed provide a solution for Hume's fundamental problem of induction and may provide a rational account of empirical knowledge acquisition. In this sense, it fosters a bottom-up approach; but knowledge, and hence top-down processes, likewise play an indispensable role. Hence, this clearly goes beyond naïve empiricism, advocating a tabula rasa concept of induction. The empirist account of induction has been characterised as a theory-independent and bottom-up process only going from the external to the internal (see pp. 22 f.). In contrast, the advocated knowledge-based account of knowledge acquisition, essentially adds context-dependent, internal-to-external and top-down aspects to this view (cf. the General Discussion of Part IV). In this sense knowledge is not only a result of learning but also a necessary and improvable means of learning. The advocated account necessarily involves data-driven *and* theory-driven aspects, bottom-up *and* top-down processes, internal-to-external *and* external-to-internal transfer of information and, in this sense, inductive *and* deductive components. Since empirical knowledge acquisition can neither be purely characterised as being inductive or deductive, I will call it a process of 'synduction', combining both aspects.

With this terminology at hand, the advocated approach can be characterised more concisely: Hume's problem of empirical knowledge acquisition indeed cannot be solved by induction alone, but only, perhaps, by an approach based on synduction (see particularly, pp. 52 f., 77 f.).

3.2 The Raven Paradox – a Specific Problem of Induction and its Bayesian Solution

The paradox of the ravens is a more specific problem of induction, showing that naïve inductivism becomes entangled in a paradox – even if the fundamental problem of induction is assumed to be solvable. Based on the asserted dichotomy of inductivism and falsificationism the raven paradox served as a justification for falsificationism. The discussion of this specific paradox of confirmation (and of the Goodman paradox, 1955/1963), largely replaced the discussion of Hume's fundamental problem of induction. There has been an extensive out-pouring of literature on the paradox of the ravens for over half of a century (e. g. Hosiasson-Lindenbaum, 1940; Hempel, 1945a, b, 1958; Watkins, 1957; Alexander, 1958; Mackie, 1963; Suppes, 1966; Hooker & Stove, 1967; Foster, 1994; Humberstone, 1994; Nickerson, 1996; Mahler, 1999; von Sydow, 2002, 2004b; Vranas, 2004).

The Bayesian resolution of the paradox is presented here because of four reasons: Firstly, if we want to discard falsificationism as the normative basis for hypothesis testing (and the WST), an alternative basis needs to be presented. Secondly, Bayesian WST models were partly antedated in the raven paradox debate. Thirdly, the advocated knowledge-based Bayesian solution of the paradox provides the background for the proposed knowledge-based Bayesian approach developed later for the WST. Finally, a WST experiment will be presented (see Chapter 5), which for the first time used ravens material, in order to connect the debate of the raven paradox and the debate on the WST more explicitly.

In the following sections, firstly, the raven paradox is introduced as a problem for naïve inductionism. Secondly, the paradox is presented as an argument for falsificationism, but it will be shown that this argument is inconclusive. Thirdly, the standard Bayesian solution of the paradox is outlined. Finally, it is argued that this solution is based on assumptions. As an alternative, a knowledge-based account is advocated also for this problem of induction, preparing a knowledge-based Bayesian account concerning the WST.

Hempel's Raven Paradox – a Specific Problem of Induction

The philosopher Carl Gustav Hempel (1905-1997) prominently formulated the paradox of the ravens in the context of his discussion of the confirmation of theories.[9] (The paradox also became called 'Hempel's paradox'.) The paradox starts with a few plausible assumptions, which seem necessary for any notion of empirical confirmation, and derives from them highly implausible consequences.

Any simple account of confirmation seems to make the following plausible assumptions (Panel 1).

Panel 1: Plausible Assumptions of a Theory of Confirmation

(a) *Assumption of confirmation*:

The hypothesis "All ravens are black" is confirmed by positive instances of black ravens. More generally, theories, laws or hypotheses of the form "All *p* are *q*" (formalised in predicate logic: $\forall x\ \ p(x) \rightarrow q(x)$) are confirmed by positive instances a of $p(a) \wedge q(a)$. This assumption has also been called "Nicod's criterion of confirmation" (cf. Nicod, 1924, 23).

(b) *Assumption of irrelevance*:

Black shoes or red herrings do not confirm (or disconfirm) "All ravens are black". More generally, evidence of $\neg p(d) \wedge \neg q(d)$ (or of $\neg p(c) \wedge\ q(c)$) does not confirm (or disconfirm) $\forall x\ \ p(x) \rightarrow q(x)$; instead these cases are irrelevant for the degree of confirmation of that hypothesis.

(c) *Assumption of equivalence*:

If evidence confirms (or disconfirms) one formulation of a hypothesis then it also confirms (or disconfirms) any logically equivalent formulation.

The three assumptions of confirmation all appear plausible if judged on their own. However, these assumptions are inconsistent:

Based on the assumption of confirmation the hypothesis "All non-black things are non-ravens" is confirmed by the case of non-black non-ravens, $\neg p(d) \wedge \neg q(d)$. The hypothesis "All non-black things are non-ravens", $\forall x\ \neg p(x) \rightarrow \neg q(x)$, is the logically equivalent contrapositive to the hypothesis "All ravens are black", $\forall x\ p(x) \rightarrow q(x)$.

[9] Actually Hosiasson-Lindenbaum formulated and discussed the paradox in a paper as early as 1940, attributing it to Hempel (pp. 136-140). Also Hempel (1945) referred to discussions with her (cf. also Hempel, 1937, 222).

Based on the assumption of equivalence, non-black non-ravens hence also support the original hypothesis “All ravens are black”. Therefore, any evidence of a non-black non-raven, for instance a red herring or a white swan, does confirm the hypothesis that all ravens are black. But this contradicts the assumption of irrelevance. One may confirm the hypothesis that ravens are black in one’s armchair only by observing the whiteness of the wallpaper, $\neg p(d) \wedge \neg q(d)$. The innocent and apparently indispensable assumptions of confirmation and equivalence lead to the bizarre conclusion, which stands in contradiction to the also plausible assumption of irrelevance. At least one of the three plausible assumptions must be mistaken and has to be dismissed.

This led to a controversial discussion. Different positions abandoned different assumptions. Hempel (1945) himself gave up the assumption of irrelevance. The falsificationist Watkins (1957, cf. 1958) instead completely dismissed the assumption of confirmation. Quine (1969) did not want to give up the irrelevance assumption as well, but he modified the confirmation assumption in the context of introducing his theory of natural kinds. The standard Bayesian resolution of the paradox, as we shall see, in a way retains all three propositions, but at the same time, it modifies all of them. The main theories of the matter, Hempel’s logical confirmation theory, the falsificationist abandoning of confirmation, and the Bayesian approach will now be outlined.[10]

Hempel’s Logical Confirmation Theory and its Problems

In order to resolve the outlined inconsistencies of a theory of confirmation and to retain the assumptions of confirmation and equivalence Hempel (1945a, 18f.; 1945b, 108-110) gave up the assumption of irrelevance. Hempel advocated that all three cases, a black raven ($p(a) \wedge q(a)$), a black non-raven ($\neg p(c) \wedge q(c)$) and a non-black non-raven ($\neg p(d) \wedge \neg q(d)$) have to be interpreted as being confirmative (Table 4).

Table 4
Confirmatory (+) or Disconfirmatory (-) Cases for the Hypothesis ‘All Ravens are Black’ According to Hempel (1945)

	Black (*q*)	Non-Black (*non-q*)
Raven (*p*)	+ cell a ($p \wedge q$)	- cell b ($p \wedge \neg q$)
Non-Raven (*non-p*)	+ cell c ($\neg p \wedge q$)	+ cell d ($\neg p \wedge \neg q$)

[10] Other resolutions of the paradox which are neither Hempelian nor Bayesian nor falsificationist, like those of Quine (1969), von Wright (1966), Forster (1994) or Mahler (1999), cannot be discussed here.

Here all true cases of a truth table of a hypothesis interpreted as an implication $p \rightarrow q$ are understood as confirmatory cases (cf. Table 1a). Although Hempel (1945, 1958) has not generalised this to other logical connectors, I would like to call this approach (which might equally be applied to other logical hypotheses) the 'logical confirmation theory.'[11]

However, this interesting logical confirmation theory also has its problems. It is questionable whether it really contributes to resolving the paradoxical character of the concept of confirmation.

Firstly, in my opinion a complete dismissal of the assumption of irrelevance remains counterintuitive and paradoxical. Hempel himself argued that the "paradoxical situation is not objectively founded; it is a psychological illusion." This illusion is based on knowledge we have in advance. In my view, his explanation of this point remains very general and is not very illuminating. With respect to the ravens problem he does not provide a particular explanation what knowledge may have led to the paradoxical intuition.[12] For this reason, Alexander (1958) criticised Hempel saying that it "is hardly enough just to accept the fact that the paradoxes disappear if we limit our evidence; we also want an account which will explain why the paradoxes arise in the first place."

Secondly, logical confirmation theory – without further refinement – entails another paradox, which I call the '*paradox of confirming contradicting hypotheses*'. For instance, evidence of a black non-raven ($\neg p(c) \land q(c)$) equally confirms the proposition "all ravens are black" ($\forall x\, p(x) \rightarrow q(x)$) and the proposition "all ravens are white" or even "all ravens are non-black" ($\forall x\, p(x) \rightarrow \neg q(x)$). But these hypotheses are commonly understood to be contradictory and even contrary to the hypothesis that all ravens are black. Hence, logical confirmation theory allows for the problematic result that the same evidence may confirm two claims which are commonly understood to be completely opposed to each other.

[11] Alternatively, this theory has also been called 'instantiation theory of confirmation' (Watkins, 1957, 117).

[12] Only in a footnote, Hempel (1945a, 21) refers to the approach of Hosiasson-Lindenbaum (1940), who provides an axiomatic mathematical system, which, in its basic features, seems actually to be largely equivalent with the more influential later Bayesian approach on this matter. But Hempel's account is not to be mixed up with the standard Bayesian approach. Firstly, Hempel (1958, 344) describes his notion of confirmation as "a non-quantitative relation between evidence sentence and an hypothesis". He does not discuss the relation of this theory to that of Hosiasson-Lindenbaum (1940). Secondly, in Hempel's (1945, 1955) account also black non-ravens (cases of cell c; cf. Table 4) confirm the conditional. We will see that this is at odds with the standard Bayesian resolution of the paradox (cf. p. 52 f.).

Falsificationism – a Cure for the Paradox of the Ravens?

The falsificationist position on the paradox of the ravens. In the wake of Popper's falsificationist approach (1934/2002, 1972) it has been argued that the raven paradox provides another argument for this negative methodology (e.g., Watkins 1957, 1958; cf. pp. 26 f.). Such proposals have presumably contributed to the popularity of falsificationism in the psychology of reasoning (cf. Humberstone, 1994). Moreover, the falsificationist solution to the paradox would be consistent with a more general falsificationist metaphysics, claiming that any kind of knowledge acquisition, for instance in biology, psychology or economy, in principle – and not only empirically – consists of blind trials (or mutations) and selective error elimination (Popper, 1972; Campbell, 1974; Dawkins, 1982; Dennett, 1995; for a critique cf. von Sydow, 2001).

The paradox again has provided falsificationism with another reason to abandon the assumption of confirmation. It is argued that the outlined paradox arising from any naïve empiricist position can be resolved, if the assumption of confirmation is discarded. In this view, the observation of a black raven ($p(a) \wedge q(a)$) does not confirm the hypothesis that all ravens are black; strictly speaking, black ravens themselves are completely irrelevant for the truth or falsity of this hypothesis. Since a universal hypothesis can never be verified and only be falsified, only falsifying observations of non-black ravens ($p(b) \wedge \neg q(b)$) are regarded as informative (cell b cases, cf. Table 4). Without the assumption of confirmation no paradox would arise in the first place.

Problems of a falsificationist resolution of the raven paradox. In the context of Hume's problem of induction, I have already discussed what I called the fundamental problem of falsificationism (cf. pp. 27 f.). I have distinguished an interpretation of falsificationism that comes close to a crypto theory of confirmation (Putnam, 1974) and a radical interpretation that renders knowledge in general – even knowledge about past falsifications – completely irrelevant for predicting any future event, which cannot count as a solution of Hume's paradox. Here it should be shown that both possible interpretations of falsificationism are highly problematic for the raven paradox as well.

(a) Let us first assume that the radical aspects of Popper's writings are decisive: This would suggest that his theory of corroboration is without any relevance for differentiating between different theories when we make predictions. As we have seen, in some parts of his work Popper emphasised that corroboration of a theory says „nothing whatever about future performance, or about the 'reliability' of a theory" (1972, 18).

This 'pure falsificationism' indeed resolves the inconsistency of the three assumptions of induction. Firstly, the confirmation assumption is clearly abandoned. Secondly, the equivalence assumption is valid for falsifying evidence: the logically equivalent propositions "all ravens are black" ($\forall x\ p(x) \to q(x)$) and "all non-black things are non-ravens" ($\forall x\ \neg q(x) \to \neg p(x)$) are both (only) falsified by the conjunction of the confirmed antecedent and the negated consequent, which is in both cases a non-black raven. Thirdly, the irrelevance assumption is retained.

Nonetheless, as I pointed out earlier, strict falsificationism in my view leads to absurd consequences (cf. pp. 27 f.). It follows from radical falsificationism that it does not matter whether one has collected many observations of ravens being black, or whether one has collected no such observation at all. Similarly, we would have no more reason to base our behaviour and our predictions on the assumption that the sun will rise tomorrow than as if this had never been observed before.

(b) Alternatively, if one stresses the moderate aspects in Popper's work, i.e. his theory of corroboration and verisimilitude, falsificationism itself, in my view, cannot escape from the paradox of the ravens.

Popper's theory of corroboration claims that we should prefer a theory to another if – ceteris paribus – this theory has been corroborated (tested without being falsified) more frequently.

But if this is accepted, each finding of any logically confirmative case (sensu Hempel) has to be seen as being corroborative. The finding of a black raven ($p(a) \wedge q(a)$) is a corroboration, since there might have been a falsifying white raven ($p(b) \wedge q(b)$) instead. Likewise, the finding of a black shoe ($\neg p(c) \wedge q(c)$) in the closet may well be corroborative, since there might also have been a white raven instead. Finally, the finding of a white swan ($\neg p(d) \wedge \neg q(d)$), say in an avarium, corroborates the hypothesis, since there might have been white ravens instead (Alexander, 1959, 229). As the paradox of the ravens has demonstrated, naïve inductionism cannot delineate which confirmative observations are relevant, just as falsificationism cannot delineate which observations are relevant for corroborating a hypothesis. The observation of white wallpaper, while sitting in an armchair, would corroborate that all ravens are black. Hence, with respect to the paradox of the ravens, the notion of corroboration leads to the same problems as the notion of confirmation or induction. This critique is consistent with Putnam's (1974) claim that falsificationism is a theory of confirmation in disguise.

In conclusion, falsificationism itself both in its strong and in its weak interpretation always leads to paradoxical consequences and thus falsificationism cannot count as a resolution of the paradox of the ravens.

A Bayesian Resolution of the Raven Paradox

The Bayesian standard resolution of the raven paradox (Alexander, 1958; Mackie, 1963; Suppes, 1966; Howson & Urbach, 1993; Nickerson, 1996) aims at maintaining all discussed basic assumptions of induction, the assumption of confirmation, the assumption of irrelevance, and the assumption of equivalence (cf. Panel 1), but modifies all of them a little bit. In particular, as we shall see, some concept of confirmation is retained, by conceding that seemingly irrelevant non-black non-ravens are confirmative – but only to a very minute degree.

Such a solution becomes possible, because the Bayesian account, in contrast to falsificationist accounts, *additionally* takes knowledge about frequencies into account. Later we will be concerned with knowledge about further constraints.

How can the raven paradox be resolved by introducing knowledge about probabilities?

(a) It is plausible to assume that we often (roughly) know the probabilities for the entities we talk about. In our case, we will have an estimate for the probability of an entity to be a raven, $P(p)$: If the universe of discourse is large, for instance considering all animals, $P(p)$ can be assumed to be very low. If the universe of discourse is smaller, for instance if we only consider birds, $P(p)$ would be a bit higher, but normally it can still be assumed that $P(q) << .5$. The Bayesian account of the raven paradox normally makes the following frequency assumptions:

$$P(raven) >> P(non\text{-}raven);\ P(black) > P(non\text{-}black) \quad (1)$$

$$P(non\text{-}raven) = 1 - P(raven);\ P(non\text{-}black) = 1 - P(non\text{-}black) \text{ [by definition]} \quad (2)$$

(b) Alexander (1958) and Mackie (1963) have shown that in a simple world, with knowledge only about the mentioned probabilities, the paradox of the ravens can be resolved, if we assume that people test the two following competing hypotheses against each other, the hypothesis that the claim is true or hypothesis of dependence, H_D (later *model* of dependence M_D), and the hypothesis that the claim is false or hypothesis of independence ($H_{I,}$ later M_I). This results in two probability matrices, one for H_D and one for H_I (Table 5).[13]

Table 5

Dependence and Independence Model of the Different (Conjunctive) Observations $O_{i,}$ First Proposed by Alexander (1958) to Resolve the Raven Paradox

	O_a: Raven ∧ Black	O_b: Raven ∧ Non-Black	O_c: Non-Raven ∧ Black	O_d: Non-Raven ∧ Non-Black
H_D	r	0	$s - r$	$1 - s$
H_I	$r \times s$	$r \times (1 - s)$	$(1 - r) \times s$	$(1 - r) \times (1 - s)$

Note: r := p(raven); s := p(black)

(c) Whereas the true/false description of two-valued logic would imply differences only for the observation of case *b*, here all four kinds of observations, *a, b, c,* and *d*, have different parameters given that either the dependence or the independence model is valid. Hence, even without further Bayesian calculations it becomes apparent that all of these observations may become informative. Conditional on the additional background knowledge, *K,* the probabilities for all observations obviously differ dependent on whether the hypothesis *H* is true (together with K resulting in H_D) or false (this latter case here corresponds to H_I).

$$\forall(i)\ P(O_i \mid H \wedge K) \neq P(B_i \mid K) \text{ for } r \neq 0 \wedge s \neq 0 \text{ (cf. Table 5)} \qquad (3)$$

It can even be generally stated whether a particular observation provides confirmatory or disconfirmatory evidence, based on the following criteria of confirmation. An observation O_i is generally

- *confirmatory* for a hypothesis iff $P(O_i \mid H \wedge K) > P(O_i \mid K)$,
- *disconfirmatory* for a hypothesis iff $P(O_i \mid H \wedge K) < P(O_i \mid K)$,
- *neutral* for a hypothesis iff $P(O_i \mid H \wedge K) = P(O_i \mid K)$.

[13] We will later also discuss alternative models of dependence and independence. Please note, that the model discussed here makes the assumption that the alternative Model H_A is the independence model ($H_A = H_I$) and that both marginal probabilities are fixed. The role of these preconditions has in my view not sufficiently been stressed by Alexander (1958), Mackie (1963) and Nickerson (1996).
My knowledge-based modification of the model of Oaksford & Chater, 1994, in the WST discussion, corresponds to this model in the ravens literature.

If we apply these criteria to the above model (Table 5) it is apparent that the observation of a white raven (O_b) is generally disconfirmatory (if $r \neq 0 \wedge s \neq 0$). It also is apparent that a black raven (O_a) remains generally confirmative for the claim that all ravens are black:

$$p(O_a \mid H \wedge K) > p(O_a \mid K) \text{ because } r > r \times s \text{ for } r \neq 0 \wedge s \neq 0 \qquad (4)$$

Hence, despite the quantification the assumption of confirmation (cf. Panel 1) still appears to be valid.

Likewise, the observation of white swans (O_d) is always confirmative (without, at this point, specifying the degree of confirmation):

$$P(O_d \mid H \wedge K) > P(O_d \mid K) \text{ because } 1 - s > (1 - r) \times (1 - s) \text{ for } r \neq 0 \wedge s \neq 0 \qquad (5)$$

However, finally, it needs to be noted that the observation of black shoes (O_c) is generally disconfirmatory in this standard Bayesian resolution of the raven paradox:

$$P(O_c \mid H \wedge K) < P(O_c \mid K) \text{ because } s - r < \mathrm{s} - r \times s \text{ for } r \neq 0 \wedge s \neq 0 \qquad (6)$$

This differs from Hempel's logical approach to confirmation and this may be counterintuitive. I will come back to this point later (pp. 52 f.). However, please note that this result rules out the outlined *paradox of confirming contradicting evidence*, which results from Hempel's position (cf. p. 42). In this model it is excluded that an observation which supports "all ravens are black" ($\forall x\ p(x) \rightarrow q(x)$) can at the same time also support "all ravens are non-black" ($\forall x\, p(x) \rightarrow \neg q(x)$). In this model, confirming evidence for one hypothesis always disconfirms the contrary hypothesis.

(d) After having shown whether observations are confirmatory or disconfirmatory it needs to be shown that the confirmation of 'all ravens are black' by non-black non-ravens is minute, if compared to the confirmation by black ravens. As we have seen, together with the background knowledge K the truth of the hypothesis H results in the dependence model, H_D, and its falsity here results in the independence model H_I.

Firstly, an example for the resulting actual values in the two basic models, H_D and $H_{I,}$ is provided. It is a plausible assumption that speaking about ravens say in a context of birds in general, the universe of discourse is relatively large, so that a probability of an entity to be a raven is low (rarity assumption, see p. 69). Hence, we assume for *P*(*p*) that *P(raven)* = 0.01 and for *P*(*q*) that *P(black)* = 0.1. Due to the definition of the basic model (cf. Table 5) the resulting probabilities for the possible kinds of observations, O_i ,depend on H_D or H_I (see Table 6).

Table 6
Probabilities $P(O_i \mid H_D)$ for an Example with P(p) = 0.01, P(q) = 0.10.

$P(O_i \mid H_j)$	O_a: Raven ∧ Black	O_b: Raven ∧ Non-Black	O_c: Non-Raven ∧ Black	O_d: Non-Raven ∧ Non-Black
H_D	0.010	0	0.090	0.900
H_I	0.001	0.009	0.099	0.891

Now further modelling steps are necessary to determine whether a black raven provides a larger information gain than a non-black non-raven. Although not all authors have completed this analysis one has to make use of the *Bayes-theorem*.

Alexander (1958, 232) and Mackie (1963) argued for a larger information gain of black ravens without the formal use of the Bayes theorem. Whereas Alexander was only concerned with the observations of conjunctions, like that of a black raven (*R* ∧ *B*), Mackie was also concerned with the case in which a raven is given and we additionally observe that it is black (*S* | *R*). To distinguish the conditional observations *S* | *R* and *R* | *S* is an important first step on the way to solve also the WST problem. Suppes (1966) did not use the above basic model (cf. Table 5), but he made formal use of the Bayes theorem. Although Nickerson (1996) also used a different basic model, he uses all steps of modelling provided here. His model is inspired by Oaksford and Chater (1994), who first provided a full Bayesian model of the WST (see Chapter 4). Von Sydow (2002) has tested the basic model of Alexander (1958) in the context of the WST debate, but constrained this model only to situations where its preconditions hold (see Chapter 5).

From the known *prior* probabilities, $P(O_i \mid H_D)$, we aim to derive the *posterior* probabilites, $P(H_D \mid O_i)$, the conditional probabilities for the hypothesis given a particular observation via Bayes theorem (the calculation is, of course, analogous for H_I):

$$P(H_D \mid O_i) = \frac{P(O_i \mid H_D)P(H_D)}{P(O_i \mid H_D)P(H_D) + P(O_i \mid H_I)p(H_I)} \tag{7}$$

To keep it simple we assume that we had no prior knowledge about whether the hypothesis that 'all ravens are black' is true or false: $P(H_D) = P(H_I) = 0.5$.

From these assumptions and the results of Table 6, we can derive the posterior probabilities using equation 7 (see Table 7).

Table 7

Posterior Probabilities $P(H_D \mid O_i)$ for $P(p) = .01$, $P(q) = 0.10$.

$P(H_j \mid O_i)$	O_a: Raven ∧ Black	O_b: Raven ∧ Non-Black	O_c: Non-Raven ∧ Black	O_d: Non-Raven ∧ Non-Black
H_D	0.91	0.00	0.48	0.51
H_I	0.09	1.00	0.52	0.49

It now seems to be apparent that the falsifying observation, O_b, and the confirming observation, O_a, differentiate much more clearly between the dependence and independence model than the confirming observations O_c or O_d.

(e) However, we still need a formal measure for the information gain of a particular observation. Here it is sufficient to use the measure proposed in the context of the raven's paradox literature (Nickerson, 1996). $P\ (H_D)_{\text{gain}} = P(H_D \mid O_i)$ - $P(H_D)$ and analogously $P(H_I)_{\text{gain}} = P(H_I \mid O_i)$ - $P(H_I)$. Later also more refined measures will be introduced (see Section 4.2). For our example, we can now calculate the different degrees of confirmation for the different observations O_i (Table 8).

Table 8

Degree of Confirmation (Positive Values) or Disconfirmation (Negative Values) for $P(p) = .01$, $P(q) = 0.10$

Observations	$p(H_D)_{\text{gain}}$	$p(H_I)_{\text{gain}}$
O_a: Raven ∧ Black	+0.41	-0.41
O_b: Raven ∧ Non-Black	-0.50	+0.50
O_c: Non-Raven ∧ Black	-0.02	+0.02
O_d: Non-Raven ∧ Non-Black	+0.01	-0.01

Firstly, Table 8 illustrates that for the two hypotheses each observation which confirms H_D disconfirms the alternative hypothesis H_I, and vice versa. This in my view resolves the paradox of confirming contrary hypotheses, which arises from Hempel's logical account of confirmation (cf. pp. 42, 47).

Secondly, and our main point here, Table 8 shows that a confirmation of a black raven ($p(a) \wedge q(a)$) or a falsificatory disconfirmation ($p(b) \wedge \neg q(b)$) has a larger impact than that of a non-black non-raven ($\neg p(d) \wedge \neg q(d)$), or than the disconfirmation of a black non-raven ($\neg p(c) \wedge q(c)$). This at least approximates the assumption of irrelevance.

One may still object that it remains counterintuitive that a white shoe should confirm "all ravens are black" in any degree! But the use of examples like white shoes implies that our universe of discourse includes almost any possible entity, resulting in a very low probability for an entity being a raven, e. g., $P(raven) = 10^{-12}$. Based on this quantitative assumption, which is plausible in the context of the ravens debate, we would even get the following degrees of confirmation (Table 9).

Table 9
Degree of Confirmation (Positive Values) or Disconfirmation (Negative Values) for $P(p) = 10^{-12}$, $P(q) = 0.10$

Observations	$p(H_D)_{\text{gain}}$	$p(H_I)_{\text{gain}}$
O_a: Raven ∧ Black	+0.41	-0.41
O_b: Raven ∧ Non-Black	-0.50	+0.50
O_c: Non-Raven ∧ Black	-0.00000002	+0.00000002
O_d: Non-Raven ∧ Non-Black	+0.0000000000003	-0.0000000000003

Table 9 clearly shows that the observations of non-black non-ravens provide only such a minute confirmation that this can be considered to be practically irrelevant for the actual testing of the ravens hypothesis.

(f) In conclusion, a Bayesian approach allows us to resolve the paradox of the ravens by retaining the assumption that black ravens confirm the hypothesis "all ravens are black" while at the same time also upholding the equivalence assumption and the irrelevance assumption. Although none of the assumptions of a naïve concept of induction (cf. Panel 1) has been completely abandoned, it should be noted that all assumptions have become modified:

The *assumption of confirmation* has become quantified. Not each confirmative instance provides a confirmation of the same amount. The enumerative understanding of confirmation is replaced by a quantified understanding of confirmation depending on additional previous knowledge, particularly about the probabilities of particular cases.

The *assumption of irrelevance* is still valid as long as we are concerned with the standard very low values $p(raven) = p(p)$ and high values of $p(non\text{-}raven) = p(non\text{-}p)$. Nonetheless, contrast to the original assumption, there are conditions under which *non-p* cases (non-black as well as black non-ravens) can become relevant for testing the truth or falsity of ‘all ravens are black’. This will be shown in an experiment with material using ravens in the related context of a WST (pp. 96 f.).

Table 10a, b
Example of the Degrees of Confirmation or Disconfirmation of Different Types of Observations O_i for (a) the Hypothesis ‘All ravens are Black’ and (b) the Logically Equivalent Hypothesis ‘All Non-black Things are Non-ravens’

(a) *p*(*black*) → *q*(*raven*) (‘0.01 → 0.10’)	
Gain of O_a (Raven ∧ Black): +0.41	Gain of O_b (Raven ∧ non-Black): -0.50
Gain of O_c (non-Raven ∧ Black): -0.02	Gain of O_d (non-Raven ∧ Black): +0.01

(b) *p*(*non-black*) → *q*(*non-raven*) ‘0.90 → 0.99’	
Gain of O_a (non-Raven ∧ non-Black): +0.01	Gain of O_b (Raven ∧ non-Black): -0.50
Gain of O_c (non-Raven ∧ Black): -0.02	Gain O_d (Raven ∧ Black): + 0.41

The *assumption of equivalence* has been modified as well. On the one hand, for instance the observation of a black raven, O_a, confirms the hypothesis “all ravens are black” as much as it would confirm the logically equivalent hypothesis “all non-black things are non-ravens”. Based on the probabilities *p(raven)* and *p*(*black*) of our original example, Table 10 provides the information gain values for *both* hypotheses. It becomes apparent that the equivalence condition is sustained in the respect of any observed object, O_i. On the other hand, the equivalence assumption is obviously violated with respect to the logical classes of observed objects. In testing the hypothesis ‘$\forall x$ *raven* (*x*) → *black* (*x*)’ the observed black raven is logically an instant for an observation O_a, $p \wedge q$, but when testing ‘$\forall x$ *non-black thing* (*x*) → *non-raven* (*x*)’ it is an instant for an observation $O_{d,}$ *non-p* ∧ *non-q* (cf. Table 10). Although a ‘black raven’ confirmation confirms both equivalent hypotheses to an equal degree (if both are formulated in regard of the same situation), observations with the same logical satus (‘black raven’ for ‘all ravens are black’ and ‘non-black non-ravens’ for ‘all non-black things are non-ravens’) have different degrees of confirmation. This is the case, because Bayesian models of the two hypotheses are based on different probabilities for the parameters. If the first model is based on the parameters *P(antecedent)* = *P*(*raven*) = 0.01 and *P(consequent)* = *P*(*black*) = 0.10, the model of the contrapositive would be based on the parameters *P(antecedent)* = *P*(*non-black*) = 0.90 and *P(consequent)* = *P*(*non-raven*) = 0.99. Therefore, the logically *non-p & non-q* case (O_d), for instance, refers to either a black raven or a non-black non-raven, and has a different impact on the two equivalent hypotheses (Table 10). Only in this sense are both hypotheses not equivalent.

The resulting Bayesian resolution of the raven paradox (cf. critically Hooker & Stove, 1967, Foster, 1994; Maher, 1999) provides a rational alterative to falsifycationism as well as to naïve inductionism. It has been shown before that the paradox of the ravens bewitches both approaches. By rehabilitating confirmation and at the same time providing an alternative to naïve inductionism, in my opinion the Bayesian approach steers a middle course between a bottom-up and a top-down approach. The Bayesian approach does not only rehabilitate confirmation (a bottom-up process) but also necessitates the use of previous knowledge about frequencies (top-down process). This knowledge goes beyond the mere use of logical classes. Based on the Bayesian resolution of the paradox, not all confirming (or disconfirming) evidence is treated equally; instead, different degrees of confirmation of disconfirmation are to be distinguished. Here the additional usage of knowledge does not immunise the tested hypothesis, but it allows us to test in a more efficient way.

Hence, even this standard Bayesian solution of the paradox of the ravens can be interpreted to be knowledge-based. This fits in with our earlier discussion of the fundamental problem of induction (cf. 3.1), where it was argued that the fundamental problem of induction – which is generally ignored by Bayesian accounts of the WST – can in principle only be resolved by a knowledge-based concept of induction. Since induction is traditionally seen as a purely bottom-up phenomena, I have introduced the concept of synduction, to stress that hypothesis testing necessarily combines bottom-up and top-down aspects. Although previous knowledge, of course, can also distort an uninterested and fair assessment of evidence, the use of knowledge is – applied in the correct way – also indispensable in resolving the paradox of the ravens.

The Forgotten Preconditions of the Bayesian Standard Resolution of the Paradox

The above Bayesian standard resolution of the paradox of the ravens depends on further tacit assumptions, again revealing the synductive nature of confirmation (cf. 3.1).

(a) In most cases the described Bayesian resolution of the raven paradox is falsely presented as a universal Bayesian account, independent of other knowledge about the situation (e. g. Alexander, 1958, cf. Table 5, p. 46). The preconditions of the outlined solution are often only mentioned briefly, not elaborating the consequences of the possibility of alternative preconditions (Mackie, 1963; Nickerson, 1996). However, the presuppositions normally made, $P(\text{raven} \mid H_D) = P(\text{raven} \mid H_I)$ and $P(\text{black} \mid H_D) = P(\text{black} \mid H_I)$, are only applicable to certain situations and they lead to quite different predictions (cf. v. Sydow, 2001; 2004c; Vranas, 2004). This will be discussed below. I will call the assumptions which equate parameters under the condition of the truth and the falsity of the hypothesis '*constancy assumptions'*. Additionally, we made '*independence assumptions'* to construct the alternative model: $P(\text{black} \mid H_I) = P(\text{black} \mid \text{raven} \wedge H_I) = p(\text{black} \mid \text{non-raven} \wedge H_I)$.

(b) Another problematic aspect of the Bayesian standard resolution is that the observation of black non-ravens (O_c: *non-p* $\wedge$ *q*) is generally disconfirmatory in this model (cf. p. 47). I will argue that the search for alternative models necessitates a knowledge-based approach.

The disconfirmatory status of the observation of black non-ravens appears problematic because this observation refers to a true case in the truth table (see Table 1, p. 5) of a conditional hypothesis (cf. similarly Hempel, 1945, 20; 1945b, 108, 110). Hence, Alexander (1958, p. 233) with his Bayesian credentials regarded this as the remaining puzzle of the Bayesian resolution of the paradox. He defended the disconfirmatory status of O_c this result by arguing that cases of O_c are anyway very unlikely to be considered, because of their low degree of disconfirmation (cf. Table 9). This is unconvincing, since there may be other parameter values where this is not the case. Consequently, this problem was even seen as a devastating point for a Bayesian resolution of the raven paradox and, hence, generally for Bayesian approaches (Hooker & Stove, 1967; Maher, 1999; cf. differently: v. Sydow 2001, 2004b, Vranas, 2004).

In order to resolve this problem one has tried to abandon the constancy assumptions which often have been made tacitly in the standard Bayesian resolution of the raven paradox: $P(p \mid H_D) = P(p \mid H_I)$, $P(q \mid H_D) = P(q \mid H_I)$. This has been discussed by Hooker and Stove (1967, pp. 307 f.) who aim to construct another *universal* Bayesian model for testing conditionals which is free from the problem that O_c is disconfirmatory.

To achieve this result, it appears necessary (cf. Hooker and Stove, 1967) to give up the assumption $P(\text{black} \mid H_D) = P(\text{black} \mid H_I)$ and to replace it, for instance, by $P(q \mid non\text{-}p \wedge H_D) = P(q \mid non\text{-}p \wedge H_I)$, that is by $P(\text{black} \mid \text{non-ravens} \wedge H_D) = P(\text{black} \mid \text{non-ravens} \wedge H_I)$. In such a model, all *non-p* cases (non-ravens) are irrelevant a priori and the *p* and *q* case (black ravens) remain confirmative. (In the context of the WST, such a model has independently been advocated by Oaksford and Chater, 1994; see Chapter 4.)

However, such a refined model has the consequence that any object characterised as black confirms 'all ravens are black'. Hooker and Stove (1967, 313) interpreted this consequence to be disastrous beyond repair.

Another problem of such a solution – if taken as a universal account – is that the assumption of equivalence is implicitly given up (v. Sydow, 2001, p. 71). As mentioned, the claim that all 'ravens are black' is confirmed or disconfirmed by the mere observation of black things and non-black things. In contrast this evidence is irrelevant for the logically equivalent contraposition 'all non-black things are non-ravens' – which should in this view be confirmed or disconfirmed by ravens and non-ravens.

The Knowledge-Based Bayesian Account and the Solutions of the Raven Paradox

Here a knowledge-based Bayesian account of the raven paradox is advocated. It should be argued here that only a knowledge-based account solves the remaining problems of the standard Bayesian account discussed in the last section. Moreover, this is in line with the knowledge-based resolution of Hume's fundamental problem of induction advocated further above (pp. 32 f.). The knowledge-based account goes beyond the outlined knowledge-based interpretation of the standard Bayesian account (interpreting the use of frequencies as the synductive use of prior knowledge, cf. pp. 45 f.) and advocates that there is not only *one* Bayesian structural model for testing the ravens hypothesis, but several. The knowledge-based account defends the standard Bayesian resolution of the paradox, but only does this conditional on its preconditions. Moreover, it will be argued that a second resolution (and perhaps more resolutions) of the paradox of the ravens can be derived, if we take a knowledge-based perspective.

(a) The objection against the standard Bayesian account that its assumptions are arbitrary, and that one may in principle also construct alternative models (Hooker & Stove, 1967), are coherent with a knowledge-based Bayesian account (v. Sydow, 2001, 2004c; Vranas, 2004). In my view, the search for one universal Bayesian model of the test of a conditional is misconceived; instead, conditionals may be tested in a variety of ways – based on a variety of models. To use the specific constancy assumptions which are valid in a given situation is not irrational at all and does not give up a concept of normativity (cf. General Discussion of Part IV). In principle, there are no mathematical constraints in assuming probability values for the expected observations O_i under the dependence model, H_D, and under some alternative model, H_A (cf. Hooker & Stove, 1967). No value is generally fixed, apart from $P(O_b \mid H_D) = 0$, which is given by the logical definition of a conditional hypothesis. Mathematically, only the consequences from such assumptions are rationally constrained.

Additionally, I want to argue that there are actually also different plausible situations, which correspond to real life contexts, which should elicit different structural constraints corresponding to different Bayesian models of testing conditionals. For instance, take the hypothesis "all women love a particular new fashion for wearing X". Here the parameter *P*(women) can normally be assumed to be constant and to be independent of the truth or falsity of the hypothesis. However, here the probability of *P*(wearing X) can plausibly be assumed to increase if the hypothesis is true and to decrease if it is false. Another model results, if, for instance, a manager of a department store knows in advance which dishware and which drinkware has been sold. Assume that he was interested in checking the additional hypothesis that "if a person bought dishware *p*, then that person also bought drinkware *q*". In this case, it would be rational to assume that $P(p)$ and $P(q)$ are constant, because he knew these aspects of the model in advance (cf. pp. 83 f., 98 f., Chapter 5 and 6). Hence, the existence of different Bayesian models of testing conditionals is not only theoretically but also psychologically plausible (cf. v. Sydow, 2004b).

(b) The second remaining problem of the standard Bayesian resolution of the raven paradox can in my view be resolved in a way which necessitates a knowledge-based approach.

We have discussed before that observations O_c are generally disconfirmatory in the standard model. This was seen as a problem because these cases are true instances of a logical implication. An alternative model was proposed which may overcome this problem, but this model violated the equivalence assumption (see pp. 53 f.). It has been outlined that a test of ‘all ravens are black’ and of the equivalent contrapositive “All non-black things are non-ravens” is in this model confirmed by different observations. In my opinion, this consequence and the violation of the equivalence assumption can only be ruled out, if the constancy assumptions are not taken as *general* premises for testing conditional hypotheses. Instead the preconditions need to be bound to particular situations or the particular knowledge about a situation. Instead of assuming the general constancy assumption $P(\mathit{consequent} \mid \mathit{negated\ antecedent} \wedge H_D) = P(\mathit{consequent} \mid \mathit{negated\ antecedent} \wedge H_I)$ one may alternatively assume that the constraint $P(\text{black} \mid \text{non-ravens} \wedge H_D) = P(\text{black} \mid \text{non-ravens} \wedge H_I)$ is a fixed matter of the situation. Applying this situational constraint to the equivalent contrapositive leads structurally to a different model based on $P(\mathit{non\text{-}antecedent} \mid \mathit{consequent} \wedge H_D) = P(\mathit{non\text{-}antecedent} \mid \mathit{consequent} \wedge H_I)$ with different parameterisations. However, this concept of a situational constraint renders the same instances in both equivalent models to be either confirmatory or disconfirmatory.[14]

This provides further reason to abandon the idea of universal constancy assumptions. Only in this case a model can be constructed where O_c is not always disconfirmatory, while retaining the equivalence assumption on the level of the situation.

Additionally, it should be noted that in principle even a situation can be constructed under which O_c becomes generally confirmative. Assume we have no (precise) prior knowledge about the probabilities of P(black) and P(raven) and the only alternative hypothesis to the tested conditional hypothesis is not an independence model, but a biconditional connective. In this case – but not in all other cases – O_c indeed becomes a confirmative case, as thought by Hempel (cf. pp. 41), by Hooker and Stove (1967) or Maher (1999, 61).

Hence, according to the advocated knowledge-based account, it depends on our previous knowledge and on the alternative hypothesis, whether O_c becomes confirmatory, disconfirmatory or irrelevant.

[14] Cf. Table 10a, b for a similar argument not in regard to the model structure but the model parameterisation.

(c) A knowledge-based approach implies an additional and perhaps even more plausible solution to the raven paradox.

Based on the outlined alternative model with the assumption $P(\text{black} \mid \text{non-ravens} \wedge H_D) = P(\text{black} \mid \text{non-ravens} \wedge H_I)$ any observation of non-ravens becomes irrelevant. In contrast to the standard solution, the irrelevance assumption remains valid in a strict sense. In the last section, it was shown that the equivalence assumption can be sustained if the constraint is not taken as a *general* constraint of testing conditional hypothesis but a constraint which is based on the given situation. Although non-ravens cases are generally irrelevant in this model other aspects of this model can still be modelled in a Bayesian way (cf. Section 3.2).

The traditional Bayesian resolution of the paradox of the ravens remains valid under the conditions it assumes. In the context of the WST debate, it will be investigated experimentally whether under such conditions non-ravens can indeed become subjectively relevant – and even more relevant than ravens.

Additionally, I have shown that within a knowledge-based Bayesian approach the raven paradox can also be resolved in a second way, which completely maintains the irrelevance assumption.

(d) All things considered, only a knowledge-based Bayesian approach can resolve the paradox of the ravens.

Firstly, the falsificationist but also the standard Bayesian approach of testing conditionals has to be replaced by a knowledge-based Bayesian approach. According to this view, the construction of a specific Bayesian model normatively relies on the knowledge about the situation. In the knowledge-based perspective Bayesian models are seen as domain-specific (or context-dependent), and nonetheless rational and normative models.

Secondly, the advocated knowledge-based Bayesian account implies at least a second resolution of the raven paradox. Crucial theoretical problems have been ruled out. Both resolutions of the paradox are consistent with the advocated approach.

Finally, the knowledge-based Bayesian account of knowledge acquisition has inductive and deductive components and can be described as a process of synduction. This is in line with the proposed general 'synductive' resolution of Hume's fundamental problem (see: pp. 32 f.). In this sense, also the treatment of the raven paradox philosophically steers a middle course between the Scylla of naïve empirism and the Charybdis of pure rationalism (cf. pp. 32 f.).

In this chapter, the two problems of induction, Hume's fundamental problem and Hempel's paradox of the ravens were discussed in detail. After having discarded falsificationism and having justified a knowledge-based Bayesian approach, we now come back to the WST debate.

4 Bayesian Hypothesis Testing in the WST – The Model of Oaksford and Chater (1994, 1998a)

In the WST debate (cf. Chapter 1.1) falsificationism has remained the most influential approach, even since the philosophical debate on the paradox of the ravens has come to consider a Bayesian resolution (cf. Chapter 3). Only in the last decade were probabilistic and Bayesian approaches applied to the WST debate (early proposals were, e.g., Kirby, 1994; Oaksford & Chater, 1994; Evans & Over, 1996).[15]

The optimal data selection model of Oaksford and Chater (1994, 1996, 1998a; Oaksford, Chater & Grainger, 1999; cf. Oaksford and Chater, 2003) represents the most refined Bayesian approach and has received most attention (e.g., Evans & Over, 1996; Laming, 1996; Klauer, 1999; Oberauer, Wilhelm, & Diaz, 1999; Osman & Laming, 2001; Yama, 2001; v. Sydow, 2002; Hattori, 2002; Oberauer, Weidenfeld & Hörnig, 2004). Hence, here concentrate on this approach.

Oaksford and Chater (1994, 1996, 1998a) turned against a domain-specific approach but also against falsificationism (cf. Oaksford & Chater, 1991; Garnham, 1993; Chater & Oaksford, 1993).

Positively, Oaksford and Chapter (1994, 1996, 1998a, 1998b, cf. 2001, 2003) proposed a general model of testing conditionals in a WST which is still concerned with *deterministic* if-*p*-then-always-*q* propositions. Although they also extended their theory to probabilistic conditionals (Oaksford & Chater, 1998b), their theory of deterministic conditionals is particularly interesting, since falsificationism has only made clear alternative predictions concerning this class of hypotheses. The Bayesian

[15] Kirby's (1994) probabilistic proposal for the WST was, in my view, still a falsificationist and not a truly Bayesian proposal (cf. Over & Evans 1994; Kirby 1994b). Likewise, the proposal of Klayman and Ha (1987), concerning the 2-4-6 task, may be said to represent a 'Bayesian falsificationist approach'.
In regard of the WST debate confer also: Evans & Over 1996a; Green & Over 1997; Green, Over & Pyne 1997; Over & Jessop 1998; Green & Over 1998; Klauer 1999 (see Section 4.3 and Chapter 5, 6).

approach claims to explain why WSTs in standard frequency conditions have led mainly to *p* and *q* selection patterns (first anomaly of the WST, see pp. 8 f.). But Oaksford and Chater (1994, 1998a, cf. 2003) have advocated a universal normative Bayesian approach for testing any conditional, and not a knowledge-based approach as I am going to advocate later (cf. Chapters 5 and 6).

Oaksford and Chater also proposed probabilistic models for syllogistic reasoning (Chater & Oaksford, 1999b) and conditional reasoning (Oaksford, Chater & Larkin, 2000), which will not be discussed here.

Chapter 4 is composed of three sections. In Section 4.1 the basic model of Oaksford and Chater (1994, 1998) will be introduced. This will also be of importance in the understanding of later modifications (Chapters 5, 6, and 7). In Section 4.2, the further model steps will be explained in detail. Here also the predictions of the model will be outlined. Finally, in Section 4.3 the crucial theoretical and empirical problems of the approach are discussed. Although some objections can be rejected, it will be shown that the account is essentially incomplete, which may, perhaps explain the problematic empirical situation of this universal Bayesian account.

4.1 The Basic Model

Oaksford and Chater's basic model (1994; 1998a) provides the basic computational level representation of a situation. It differs from a logicist-falsificationist view on testing conditionals in a WST in two respects: Firstly, the basic model of a conditional consists of two mathematical sub-models, representing either the truth or the falsity of the conditional hypothesis (dependence model and independence model). Secondly, based on additional knowledge about frequencies the models are quantified. Both models are characterised as a contingency table (Table 11).

The dependence model (M_D, cf. Table 11a), representing the states of the world in which the conditional is assumed to be true, is coherent with the logical definition of a deterministic understanding of an implication 'if *p* then *q*' (cf. Table 1, p. 5). Correspondingly, the probability $P(p \wedge non\text{-}q)$ is set to be zero (but cf. also Oaksford & Chater, 1998b).

Table 11

Basic Model of Oaksford and Chater (1994, 60; 1998a, 178) With a Dependence Model (M_D) and an Independence Model (M_I)

(a) M_D	q	*non-q*	marg.
p	p	0	p
non-p	$(1 - p)$ q'	$(1 - p)$ $(1 - q')$	$1 - p$
marg.	$q' + p - p q'$	$(1 - p)$ $(1 - q')$	1

(b) M_I	q	*non-q*	marg.
p	$p q'$	p $(1 - q')$	p
non-p	$(1 - p)$ q'	$(1 - p)$ $(1 - q')$	$1 - p$
marg.	q'	$1 - q'$	1

Note: In the cells of M_D and M_I the following notation is used:

$$p := P(p); q' := \frac{P(q) - P(p)P(M_D)}{1 - P(p)P(M_D)}$$

The alternative model is constructed as an independence model (M_I, cf. Table 11b), using a contingency table in which the occurrence of p and q is assumed to be statistically independent: $P(p \mid q) = P(p)$ and $P(q \mid p) = P(q)$. Thus, the joint probabilities in the cells of the contingency table of M_I are calculated by multiplying marginal probabilities.

In the dependence model the probability of the confirmative cell a, $P(p \wedge q) = P(C_a)$, is the base rate of the occurrence of the antecedent, $P(p)$, since the alternative b cell, $P(p \wedge \textit{non-q}) = P(C_b)$, was zero. This is based on the assumption that $P(p)$ is the same for both models M_D and M_I (first constancy assumption). Logicism has not quantified the probabilities of the logical cases and has regarded C_b to be equally 'true' whether the hypothesis is true or false. In contrast, C_a becomes informative for the Bayesian account because $P(C_a)$ differs in the two quantitative models M_D or M_I (cf. pp. 45 f.).

In constructing the dependence model Oaksford and Chater (1994; 1998a) additionally assumed that the *non-p* cases are not affected by truth or falsity of M_D. Hence, the two *non-p* cases have the same probabilities in the dependence model M_D as *well as* in the independence model M_I: $P(C_c \mid M_D) = P(C_c \mid M_I)$; $P(C_d \mid M_D) = P(C_d \mid M_I)$. Put differently, it is assumed that $P(q \mid \textit{non-p})$ is constant, independent of whether either M_D or M_I is true (second constancy assumption). Based on this general assumption, *non-p* cases – and the selection of a *non-p* card in an information selection task – are understood to be never informative for differentiating between the truth (M_D) or falsity (M_I) of the tested hypothesis (Oaksford & Chater 1994, 612).

The two postulated constancy assumptions, $P(p \mid M_D) = P(p \mid M_I)$ and $P((q \mid non\text{-}p) \mid M_D) = P((q \mid non\text{-}p) \mid M_I)$, are claimed to hold generally for the testing of all conditionals regardless of the specific context (Oaksford and Chater, 1994, 610; 1998b, 178-179). Oaksford and Chater justified the first assumption by arguing that otherwise, "observing *p* or *non-p* instances alone could provide evidence about whether the rule holds" (1994, 610). Indeed, it seems plausible that the number of ravens, $n(p)$, is normally independent from the truth or falsity of the hypothesis 'all ravens are black' (but cf. pp. 54 f.). The second constancy assumption, $P((q \mid non\text{-}p) \mid M_D) = P((q \mid non\text{-}p) \mid M_I)$, is based on the argument that the conditional makes "a claim about the consequent only if the antecedent occurs but makes no claim otherwise (Quine, 1959)" (Oaksford & Chater, 1994, 610).

We will come back to discuss these assumptions later on (Chapter 5, 6; cf. v. Sydow, 2002, 2004b); for the time being we will accept them as given.

A final assumption, which will not be focused here, is made to construct the basis model. Oaksford and Chater (1994, 612; cf. 1998a, 178-179) used a modified $P(q')$ value instead of $P(q)$:

$$P(q') := \frac{P(q) - P(p)P(M_D)}{1 - P(p)P(M_D)} \qquad (8)$$

In my opinion, the purpose of this modification was to rule out a problem resulting from the used constancy assumptions. Without this modification the resulting marginal probabilities $P(q)_{res}$ and $P(non\text{-}q)_{res}$ increase if the hypothesis is assumed to be true, $P(M_D) = 1$. Equation 8 allows that $P(q)_{res} = P(q)$ becomes valid for the dependence model if and only if $P(M_D) = 1$ and this becomes valid for the independence model if $P(M_D) = 0$. Hence, the modification allows us to assume the same probabilities $P(q)_{\mathrm{res}} = P(q)$ if the hypothesis is either definitely believed to be true or false. However, for all other values of $P(M_D)$ the $P(q)_{res}$ varies depending on the dependence or the independence model. Without this modification this would also be the case for $P(M_D) = 0$ or $P(M_D) = 1$. This modification also provides the basis for using $P(p)\ P(M_D) < P(q)$ as precondition for the testing of a conditional, instead of $P(p) < P(q)$.

Since we will not discuss this modification later on, I want to point out that this is not a necessary modification, even if one accepts the postulated constancy assumptions. Alternatively, one may skip Equation 8 and use $P(q)$ directly in the basic model (cf. Experiment 3). If this is done the resulting probability for q, $P(q)_{res}$, is indeed higher for the dependence model than for the original $P(q)$ also if $P(M_D) = 1$. But if subjects knew that the second constancy assumption, $P(q \mid non\text{-}p)$, is true, it would indeed by absurd to assume that $P(q)_{\mathrm{res}}$ should not increase if they additionally get to know that the conditional is true. The truth of the conditional here implies that there are additional *qs* which are 'caused' by *ps*.

Moreover, the resulting precondition, $P(p)\ P(M_D) < P(q)$, is not generally valid. For instance, the parameter values $P(p) = 0.6$, $P(q) = 0.4$, and $P(M_D) = 0.5$ fulfil the equation $P(p)\ P(M_D) < P(q)$. Nonetheless, it is evident that $P(p) < P(q)$ is violated and I do not see why it should not be rational to refuse any further test, because the frequencies alone – independent of one's belief in $P(M_D)$ – prove M_D to be false. Therefore, as a general assumption this modification is at least problematic.

Based on the outlined assumptions Oaksford and Chater (1994, 1998b) introduced the general basic model presented in Table 11. Although it is obvious that *q* cards in this probabilistic approach may become informative, since $P(C_a)$ differs in the dependence and independence model, further modelling steps are needed to provide a more precise expected information gain measure for selecting cards in the WST.

4.2 Further Steps in Modelling and Predictions

Further Steps in Modelling the WST

Subsequently the further modelling steps are outlined (Oaksford & Chater, 1994, 610, 1998a, 179-183; cf. v. Sydow, 2002), which are to be used also for the modified models (Chapter 5, 6 and 7).

Probabilities for Reversed Cards in the Dependence and Independence Model

The basic model described provides us with estimates of the probabilities of joint events C_i, conditional on the dependence model or the independence model. Additionally now the probability of getting a particular reverse side of a card needs to be determined, if we have selected a particular front side of a card, *p*, *non-p*, *q* or *non-q* (again conditional on the dependence or independence model). For instance, on the reverse side of a *p* card there can be either a *q* or a *non-q* side. The probability of finding a *q* side on a *p* card is given by the conditional probability $P(q \mid p)$, which can be derived by dividing the cases in which a conjunction of *p* and *q* is given ($P(C_a) = P(p \wedge q)$) by all cases in which *p* is given, $P(C_a) = P(p \wedge q)$ plus $P(C_b) = P(p \wedge non\text{-}q)$. Based on the basic model (Table 11) the following conditional probabilities are to be derived for the dependence and independence model:

$$P((q \mid p) \mid M_D) = \frac{P(C_a \mid M_D)}{P(C_a \mid M_D) + 0}; \quad P((q \mid p) \mid M_I) = \frac{P(C_a \mid M_I)}{P(C_a \mid M_I) + P(C_b \mid M_I)} \quad (9)$$

$$P((\neg q \mid p) \mid M_D) = \frac{0}{0 + P(C_a \mid M_D)}; \quad P((\neg q \mid p) \mid M_I) = \frac{P(C_b \mid M_I)}{P(C_a \mid M_I) + P(C_b \mid M_I)}$$

The conditional probabilities can equally be calculated for the other cards, *non-p*, *q*, and *non-q*. Interestingly, the resulting equations differ between the dependence and independence model not only for the falsificatory *p* and *non-q* cards but also for the *q* card ($P((p \mid q) \mid M_D) \neq P((p \mid q) \mid M_I)$ for $0 < P(q') < 1$ and $0 < \mathrm{P}(p)$). The general equations for the *q* card are:

$$P((p\,|\,q)\,|\,M_D)=\frac{P(C_a\,|\,M_D)}{P(C_a\,|\,M_D)+P(C_c\,|\,M_D)}; \qquad (10)$$

$$P((p\,|\,q)\,|\,M_I)=\frac{P(C_a\,|\,M_I)}{P(C_a\,|\,M_I)+P(C_c\,|\,M_I)};$$

$$P((\neg p\,|\,q)\,|\,M_D)=\frac{P(C_c\,|\,M_D)}{P(C_c\,|\,M_D)+P(C_a\,|\,M_D)};$$

$$P((\neg p\,|\,q)\,|\,M_I)=\frac{P(C_c\,|\,M_I)}{P(C_c\,|\,M_I)+P(C_a\,|\,M_I)}$$

To make the differences between M_D and M_I apparent, we need to insert the formulas from the basic model (Table 11):

$$P((p\,|\,q)\,|\,M_D)=\frac{P(p)}{P(p)+(1-P(p))P(q')}; \qquad (11)$$

$$P((p\,|\,q)\,|\,M_I)=\frac{P(p)P(q')}{P(p)P(q')+(1-P(p))P(q')};$$

$$P((\neg p\,|\,q)\,|\,M_D)=\frac{(1-P(p))P(q')}{P(p)+(1-P(p))P(q')};$$

$$P((\neg p\,|\,q)\,|\,M_I)=\frac{(1-P(p))P(q')}{P(p)P(q')+(1-P(p))P(q')}$$

This shows that $P(p\,|\,q)$ differs in the dependence and independence model. Moreover, $P(p\,|\,q)$ differs from $P(q\,|\,p)$ as there may be different probabilities for a raven being black and a black thing being a raven (in the context of the raven paradox debate cf. Mackie, 1963).

All in all, for each of the four cards and both possible outcomes there are equations for the dependence and independence model, resulting in sixteen equations for conditional probabilities.

Bayes' Theorem – Probability of a Model Conditional on the Data

Above the probabilities for possible sequences of data have been derived, conditional either on the dependence or independence model, $P(D_y\,|\,M_x)$. From the conditional probabilities and the prior probabilities $P(M_D)$ and $P(M_I)$ = 1 - $P(M_D)$ we can calculate the posterior probabilities, $P(D_y\,|\,M_x)$, by using Bayes' theorem :

$$P(M_D \mid D_y) = \frac{P(D_y \mid M_D)P(M_D)}{P(D_y \mid M_D)P(M_D) + P(D_y \mid M_I)P(M_I)}; \tag{12}$$

$$P(M_I \mid D_y) = \frac{P(D_Y \mid M_I)P(M_I)}{P(D_y \mid M_D)P(M_D) + P(D_Y \mid M_I)P(M_I)}$$

For a complete rational analysis, the two equations for the dependence and independence model (12) need to be computed for all eight possible sequences of possible data, D_y: $q \mid p$, $\neg q \mid p$, $q \mid \neg p$, $\neg q \mid \neg p$, $p \mid q$, $\neg p \mid q$, $p \mid \neg q$, and $\neg p \mid \neg q$. For the two falsifying cases, $\neg q \mid p$ and $p \mid \neg q$, the posterior probabilities are of course always $P(M_D \mid D_y) = 0$ and $P(M_I \mid D_y) = 1$. This reflects the (deterministic) definition of an implication, which is falsified by a single counterexample.[16] For the irrelevant sequences $q \mid \neg p$ and $\neg q \mid \neg p$ there are no differences between the dependence and the independence model, but all other data sequences, $q \mid p$, $p \mid q$, $\neg p \mid q$, and $\neg p \mid \neg q$, normally yield different results for the dependence and independence model.

Information Gain of a Sequence of Data (Shannon-Wiener Information)

It is assumed that a data sequence, D_i, which leads to the greatest reduction in uncertainty, provides the optimal selection. Shannon-Wiener information (Shannon & Weaver, 1949; Wiener, 1948) is a formalisation of this idea of uncertainty reduction.

Firstly, the initial uncertainty, depending on the prior degree of belief in the two models, $P(M_D)$ and $P(M_I)$, is quantified as entropy $E(M_x) = E(M_D) = E(M_I)$:

$$E(M_x) = P(M_D) \times \log_2 \frac{1}{P(M_D)} + P(M_I) \times \log_2 \frac{1}{P(M_I)} \tag{13}$$

Uncertainty or entropy, here of a hypothesis, is the sum, over the two assumed possible outcomes (here the truth of the hypothesis, M_D, or its falsity, M_I), of the product of the probability of these outcomes times the log of the inverse of these probabilities. If $P(M_D) = P(M_I) = .50$, then $E(M_x) = 1$. If $P(M_D) = .9$ and $P(M_I) = .1$ than uncertainty is reduced, $E(M_x) = .33$.

[16] Even if one would apply this formula successively, this falsification is not undermined by new applications of the Bayes theorem, since $P(M_D) = 0$ only allows for posterior probabilities $P(M_D \mid D_y) = 0$. Nonetheless, the model needs to be extended if applied to successive selection of cards like in the RAST (see Oaksford & Chater 1998b; cf. Klauer 1999).

Secondly, the posterior probabilities – $P(M_D | D_y)$ and $P(M_I | D_y)$ – which have been determined in the last section for all eight data sequences, D_y, can now be used to compute the resulting uncertainty for the model, given a sequence would have been observed:

$$E(M_x | D_y) = P(M_D | D_y) \times \log_2 \frac{1}{P(M_D | D_y)} + P(M_I | D_y) \times \log_2 \frac{1}{P(M_I | D_y)} \quad (14)$$

Thirdly, the difference between the initial uncertainty and the resulting uncertainty, after observing a sequence D_y, provides a measure for the information gain caused by the observation of that sequence:

$$IG(D_y) = E(M_I) - E(M_I | D_y) \quad (15)$$

Note: As an alternative measure of information gain, Kullback-Leibler distance has been proposed (Evans & Over, 1996a, 358; Laming, 1996, 366, 370-371, 373; Klauer, 1999; cf. Oaksford & Chater, 1996, Chater & Oaksford, 1999; Oaksford, Chater, & Grainger 1999; see p. 71 for details).

Expected Information Gain (EIG)

In the last section, we calculated information gain values for a given sequence of data, D_i, for instance, p | q. However, in the WST participants never actually come to turn over the selected cards. They never get direct evidence in the form of a selected sequence. If a card, for instance q, is selected, this hypothetically leads to *two* different outcomes p and *non-p*. Hence, the information gain values of both possible sequences, p | q and *non-p* | q, resulting from a card selection, need to be taken into account and need to be integrated.

Firstly, for each card selection we need to compute the probability that a sequence occurs, now simultaneously considering *both* possible models:

$$P(D_y) = P(M_D)P(D_y | M_D) + P(M_I)P(D | M_I) \quad (16)$$

Secondly, we can now calculate the expected resulting uncertainty values of a particular card, *EE*(card$_z$), by weighting the computed uncertainty values of the data sequences (Equation 14) using the probability of finding that sequence when selecting a card (Equation 16). (For the *p* card D_y stands for the sequence $q \mid p$; D'_y for the complementary sequence $\neg q \mid p$.) [17].

$$EE(M_x \mid card_z) = P(D_y)E(D_y) + P(D'_y)E(D'_y) \qquad (17)$$

Finally, the expected information gain values for each of the four cards in the WST, *EIG*(card), is calculated by subtracting[18] the expected uncertainty values, *EE*(M_x | *card*), from the initial uncertainty values, *E*(M_x) (Equation 13):

$$EIG(card_z) = E(M_x) - EE(M_x \mid card_z) \qquad (18)$$

Most of Oaksford and Chater's predictions (1994, Fig. 2, Fig. 3; 1998a, Fig. 10.2, Fig. 10.3) are based directly on these expected information gain values (see p. 67).

Error Value and Scaled Expected Information Gain (SEIG)

Oaksford and Chater (1994, 614) added two further aspects to their model when reanalysing previous results in WSTs.[19]

Firstly, a fixed error constant was added to the *EIG* value of each card to model that the subjects may choose a card by chance.

$$EIG'(card_z) = EIG(card_z) + 0.1 \qquad (19)$$

This was particularly done to account for *non-p* selections, which are generally deemed to be irrational (but cf. the next Chapter). Although this modification is plausible, it should not be interpreted as an aspect of the normative model itself.

Secondly, card selection is assumed to be competitive. Correspondingly, each resulting expected information gain value of a particular card is divided by the

[17] Oaksford and Chater (1998, 181) insert their Equation 9 instead Equation 7 into their Equation 10. However, in their example they used Equation 7 (p. 183). To me the latter procedure seem to be correct and only this is coherent with the original presentation of the theory (1994, Equation 6).

[18] The different formulations in Oaksford and Chater (1994, Equation 1, 2, 4, 5, 6) and later in Oaksford and Chater (1998, 179 f.) are equivalent (cf. Laming 1994, 370; Evans & Over 1996, 357, footnote 1).

[19] Furthermore, Oaksford and Chater (2003) – beside other modifications which will be discussed later – introduced a 'selection tendency function' (Hattori, 2002), which allows to achieve a better fit to the data. I will not adopt this modification, because it is only a descriptive means to improve the fit of the data by introducing another free parameter. I do not see a *normative* justification of this modification.

average expected information gain value of all cards. This value is referred to as *scaled expected information gain*, *SEIG*.

$$SEIG(card_z) = \frac{EIG'(card_z)}{(\sum_{i}^{4} EIG'(card_i))/4} \tag{20}$$

Predictions

Oaksford and Chater's model (1994, 1998a) for testing if-then hypotheses is determined by the three main parameters, $P(p)$, $P(q)$, and $P(M_D)$.[20] Oaksford and Chater have presented their predictions in a three-dimensional parameter space (Figure 3).

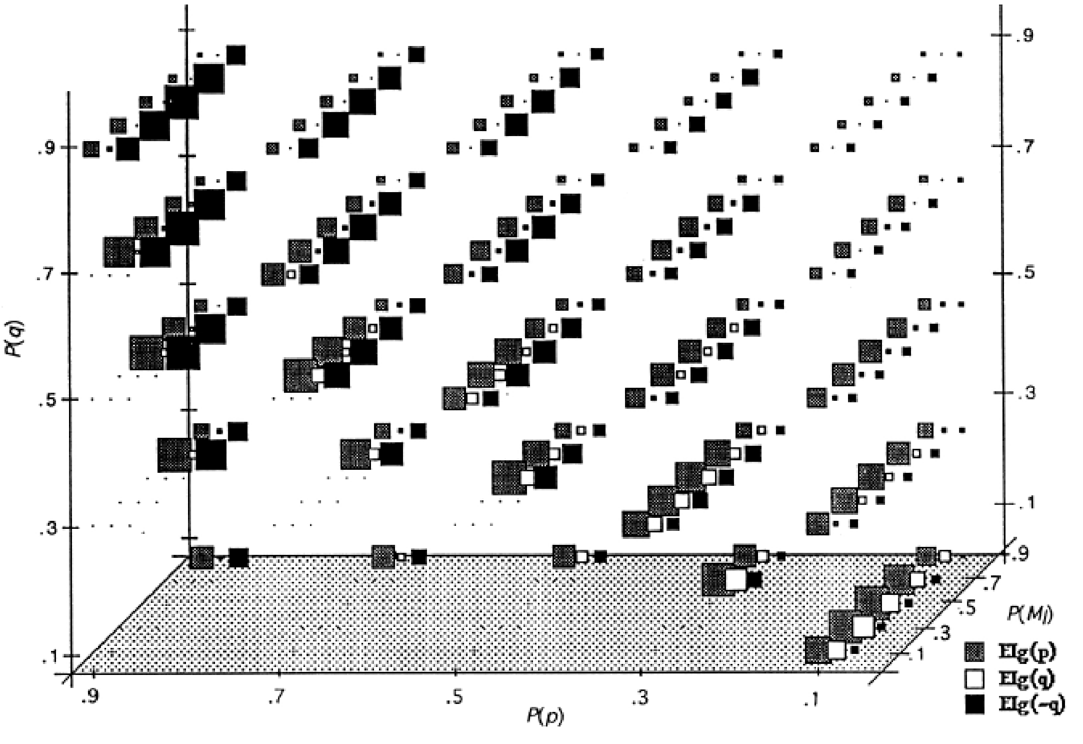


Figure 3. Model behaviour. At each co-ordinate ($P(p)$, $P(q)$, $P(M_I)$) the boxes represent the *EIG* value for the three cards, *p*, *q*, and *non-q*. Here the *EIG* values for the *non-p* card are always zero. The area of the boxes is proportional to the *EIG* values of the cards. Three dots indicate regions where the precondition of the model, $P(q) \geq P(q)P(M_D)$, is violated. (Taken from Oaksford & Chater, 1994, 611.)

[20] Additionally, the error constant introduc n the last section and variable for a probabilistic understanding of the rule (Oaksford & Chater, 1998b) can seen as a free parameter (Figure 3 represents the results without an error constant).

Oaksford and Chater's predictions can be summarised in the following way.

Firstly, the parameter $P(M_D)$ only rescales the *EIG* values of the cards, but does not change the relative order of their *EIG* values. Therefore, the relative informational value of each card is understood to be insensitive to the prior probabilities, $P(M_D)$, rendering this parameter largely irrelevant.[21]

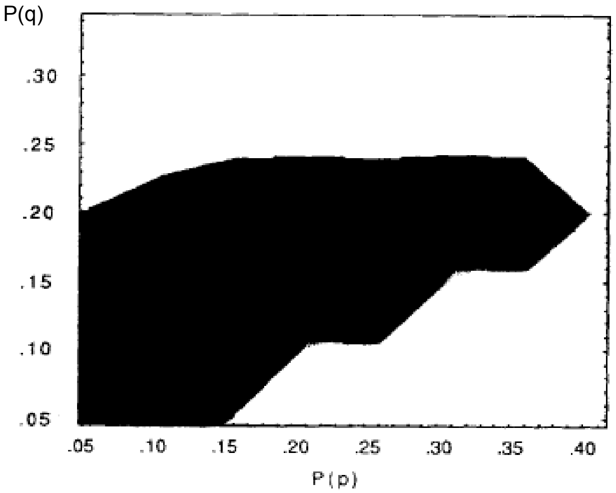


Figure 4. Plot of the parameters $P(p)$ against $P(q)$, with $P(M_D) = 0.5$, showing a region in black for which $EIG(q) > EIG(non\text{-}q)$. The white region below $P(q) = 0.20$ violates the precondition $P(q) \geq P(p)P(M_D)$. (Source: Oaksford & Chater 1994, 612; 1998a, 185.)

Secondly, the expected information gain of a *non-p* card, *EIG*(*non-p*), is always zero. This derives from the construction of the basic model and the fundamental constancy assumption that a conditional only makes a claim about situations in which its antecedent is true (but cf. Chapter 5).

Thirdly, the parameters $P(p)$ and $P(q)$ are postulated to have a strong impact on the EIG values of *q*- or *non-q* selections. For the region where $P(p)$ and $P(q)$ are small, it follows that $EIG(q) > EIG(non\text{-}q)$. Conversely, in the region where $P(p)$ or $P(q)$ is large, it follows that $EIG(q) < EIG(non\text{-}q)$. This becomes apparent in Figure 4 taken from Oaksford and Chater (1994).[22]

If we include the other cards into this relative preference prediction, the following order of *EIG* or *SEIG* values can be derived for the low probability region (cf. 994, 20):

$$SEIG(p) > SEIG(q) > SEIG(non\text{-}q) > SEIG(non\text{-}p) \qquad (21)$$

[21] This became a matter of controversy (Evans & Over 1996a; cf. Klauer 1999). However, particularly in this respect, Oaksford and Chater have successfully defended their model theoretically and empirically (Oaksford & Chater, 1996; Chater & Oaksford, 1999; Oaksford, et al. 1999; see pp. 71 f.).

[22] Oaksford and Chater (1994, 612) also made the prediction that the *non-q* card is informative "to the extent that $P(p)$ is large. It is independent of $P(q)$". This prediction refers to *Figure* 3, but seems to be inconsistent with *Figure* 4. This inconsistency can be resolved. In *Figure* 4 we are not concerned with absolute *EIG* values, but with the relative size of the *EIG* values of *q* and *non-q*. Because of the use of forced-choice instructions we will only consider these latter predictions (see p. 93). Similarly, Oaksford, Chater, and Grainger (1999, 200) explicitly expected „more *not-q* card selections when *P(q)* is high."

Conversely, if $P(p)$ and $P(q)$ are high, a different order results[23] (cf. Oaksford & Chater 1995a, 100-101; 1998b, 375):

$$SEIG(p) > SEIG(non\text{-}q) > SEIG(q) > SEIG(non\text{-}p) \tag{22}$$

Based on these predictions and some additional assumptions about the parameters, Oaksford and Chater (1994) explained previous results in the negation paradigm, in therapy experiments and particularly the empirically found predominance of *q* instead of *non-q* selections in testing conditionals in standard WSTs (first anomaly of the WST, see pp. 8 f.). Here, two additional assumptions should be mentioned.

Firstly, in the absence of information about the prior probabilities of the hypothesis they assumed that $P(M_D) = P(M_I) = 0.5$. From the viewpoint of Oaksford and Chater, this is not problematic (cf. Oaksford, et al. 1999).

More significantly, Oaksford and Chater (1994, 627; cf. Oaksford & Chater, 1998, 221) formulated a *rarity assumption*, stating that in normal tests of an if-*p*-then-*q* hypothesis, $P(p)$ and $P(q)$ are by default considered to be rare. This assumption differs from Keynes' (1921) principle of indifference, $P(p) = P(q) = 0.5$. Earlier the rarity assumption was advocated in the context of the 2-4-6 task (cf. Klayman & Ha, 1987; Crott, Giesel & Hoffmann, 1999). Moreover, the assumption is also a precondition to the standard resolution of the raven paradox (cf. 3.2). In testing ‚all ravens are black' we can normally reasonably assume that there are less ravens than non-ravens and less black things than non-black things (cf. Alexander 1958, 232, Mackie 1963, 266; Suppes 1966, 199). Actually, the rarity assumption has been directly tested by asking subjects whether they would formulate hypothesis in terms of rare or frequent events (McKenzie, Feerreira, Mikkelsen, McDermott, & Skrable, 2001, cf. McKenzie and Mikkelsen, 2000). The plausibility of the rarity assumption is vividly illustrated by the dictum of Bochenski: "The world is full of non-elephants." In the next section, direct empirical tests of this general approach will be discussed.

[23] This is a simplification: For very high $P(q)$ values the order of $SEIG(p)$ and $SEIG(non\text{-}q)$ is reverserved: $SEIG(non\text{-}q) > SEIG(p)$.

4.3 Theoretical and Empirical Problems of the Universal Basic Model

First the theoretical problems of the model of Oaksford and Chater (1994, 1998b, cf. 2003) will be discussed and it will be shown that while many objections have been ruled out the approach has not been freed of one fundamental problem, the universal formulation of its basic constancy assumptions. Secondly, the empirical situation will be reviewed. Although some frequency effects were found, the empirical evidence is rather negative for the precise predictions made by this model. It is claimed that additional explanatory factors seem to be needed to explain the variety of results found.

Theoretical Objections – With Particular Emphasis on Laming (1996)

The theoretical foundation of the normative and universal Bayesian approach of Oaksford and Chater (1994, cf. 1998a) has remained controversial. Many authors continued to pursue a logic-based mental model research programme or a domain-specific research programme (e.g., Fiddick, Cosmides & Tooby, 2000; Johnson-Laird & Byrne, 2002; Beller & Spada, 2003).

More importantly, Oaksford and Chater's approach also became the direct target of theoretical criticism. Some authors formulated the fundamental objection that the construction of the basic model is completely arbitrary (particularly Laming, 1996; cf. Evans & Over, 1996, v. Sydow, 2002), others criticised single aspects of the model or formulated modifications (e.g., Evans & Over, 1996; Green & Over, 1998; Klauer, 1999; cf. v. Sydow, 2002; Hattori, 2002). Oaksford und Chater have replied to many objections (Oaksford & Chater, 1996, 1998b; Oaksford, 1998; Chater & Oaksford, 1999a, 2003), but as it will be shown, they have not yet resolved the fundamental problem concerning the basic constancy assumptions of their model.

Before discussing this point I am going to outline three other disputes, firstly, one about relevance theory, secondly, one about the information gain measure that is used, and, thirdly, I will outline the dispute about the dependence and independence model, which prepares for the knowledge-based account. Finally, Laming's main additional critique, concerning the fundamental constancy assumptions in the basic model, will be discussed. In particularly the last point will be directly relevant for the further work.

A Dispute on Bayesian Hypothesis Testing and Relevance Theory

Sperber, Cara, and Girotto (1995) have objected that data provided in the context of their relevance theory (see pp. 17) is incoherent with Oaksford and Chater's approach (1994). Oaksford and Chater (1995a) rebutted that, conversely, one may reinterpret the data of Sperber et al. (1995) so that it favours their Bayesian approach, using Bayesian calculations as a more refined measure of relevance.

However, the empirical results when testing both theories against each other remained undecided (Almor & Sloman, 1996; cf. Hardman, 1998; Oaksford, Chater, & Grainger, 1999, Exp. 3 and 4). More recently Sperber and Girotto (2002, 2003) have provided evidence which is connected with what I will treat as focus effects in Part III, but without linking this to deontic logic (cf. Part III).

In my opinion, relevance theory may well uniquely explain some effects, but I also agree with Oaksford, Chater, Grainger and Larkin (1997, 455) that relevance theory cannot directly account for any effects of the manipulations of probabilities: Relevance theory "is framed in terms of cognitive effects and effort that do not appear to have any immediate application to [..] purely probabilistic manipulations". (Cf. pp. 177 f., 281 f.)

A Dispute on the Information Gain Measure

The Shannon-Wiener information gain measure, used by Oaksford and Chater (1994), has been criticised by several authors who instead favour the Kullback-Leibler distance (Evans & Over, 1996a, 358; Laming, 1996, 366, 370-371, 373; Klauer, 1999). Although Oaksford and Chater (1996, Chater & Oaksford, 1999a) have conceded problems in the Shannon-Wiener measure they have defended their formalisation as being generally adequate in the standard context of a disinterested inquiry. Oaksford, Chater & Grainger (1999) investigated this point empirically. Although their general results were rather disappointing for their model as a whole, the results clearly favoured their approach over the alternative information measure (Exp. 1 and 2). Hence, here I continue to use Shannon-Wiener information.

In any case, I try to use the rather unproblematic parameter value $P(H_D) = 0.5$. For this parameter value a theoretical problem of this measure can be ignored, which has not been removed by Oaksford and Chater. If the initial probability for the truth of the hypothesis is $P(H_D) = 0.25$ and the probability given the additional evidence is $P(H_D \mid evidence) = 0.75$ the Wiener-Shannon entropy (uncertainty) values of both probabilities are equal (cf. Equations 13, 14). Hence, the information gain of such evidence remains zero, which seems to be absurd (cf. Evans & Over, 1996; Oaksford & Chater, 1996). However, for the initial probability of the hypothesis $P(H_D)$ = 0.5, used here, this problem cannot occur.

Specific Modifications of the Basic Model

Some authors have supported a probabilistic approach but pointed out that the Oaksford and Chater's approach (1994) provides a too rigid and formalised account (Evans & Over, 1996; Green, Over & Pyne, 1997, 211; Green & Over, 1998, 191-192).[24] Although I agree with these authors in their criticism that the model is too rigid, I think that this technically refined approach is generally advantageous. It will be argued that only a context-independent application of this technical model is erroneous (cf. Chapter 5 and 6).

Additinally, objections that are more specific have elicited some valuable modifications of the model of Oaksford and Chater (1994), without leading to a modification of the fundamental constancy assumptions.

Firstly, it has been pointed out that conditionals do not always need to be understood to be deterministic (Green & Over, 1997; Over & Jessop, 1998). Although it should be noted that the logical-falsificationist norm of hypothesis testing is only applicable for non-exhaustive tests of *deterministic* hypothesis (normally implications), one may also think of a probabilistic conditional hypothesis with $P((p \wedge non\text{-}q) \mid M_D) > 0$. These proposals caused Oaksford and Chater to introduce a refined general model with an additional exception parameter (1998b).

Secondly, it has been objected that the alternative model does not always need to be a strict independence model (cf. Laming 1996, 369-370; Green & Over, 1997; Evans & Over, 1996; Green, Over & Pyne 1997, 214). This possibility has been theoretically conceded in a modification of the model by Oaksford and Chater (1998b; Green & Over, 1998; Oaksford, Chater, & Grainger, 1999, 221, 223; cf. Oaksford & Chater, 2003; Hattori, 2002, 1246-1248).

Although the predictions of the two modifications neither have been elaborated in detail theoretically, nor have been tested empirically, only being used in a post hoc way to explain some unexpected results, the modifications can be understood as steps towards a more flexible model. Nonetheless, Oaksford and Chater (1998a, 1998b, cf.

[24] However, please note that more recently Over & Evans, 2003, Evans, Handley & Over, 200, and Over, 2004, themselves have proposed a universal account of the meaning of a conditional.

even 2003) have not accounted for the more basic critique that the assumed fundamental constancy assumptions for the model are arbitrary (Laming 1996, cf. Green & Over 1997).

Laming's Fundamental Objections against the Basic Model

Laming (1996, 369) reasoned that the fundamental constancy assumptions between the cell probabilities of the two models of the basic model (cf. Table 11, p. 60) are arbitrary. Oaksford and Chater assumed that *P*(*p*) and *P*(*q* | *non-p*) are constant across the dependence and independence models. Laming objected that these assumptions could not be justified (cf. similarly, Oberauer, Wilhelm, and Diaz, 1999, 141). The "particular alignment of parameters between the two models is dictated by the need to make *K* [that is *non-p*] uninformative". If the empirical rank order of the card selection "had been other than it is, it would simply have dictated different parameter values and assumptions" (Laming, 1996, 371).

Laming (1996, 369) contended that one may have equally constructed a basic model using different constancy assumptions, leading to predictions which would be completely incoherent with previous data on the WST. The model in Table 12, like the original model, assumes a strict dependence model, the validity of the implication ($p \rightarrow q$ with $P(p \wedge non\text{-}q) = 0$), and a strict independence model. Additionally, the alternative construction of a basic model is based on other assumptions, $P(q \mid M_D) = P(q \mid M_I)$ and $P((p \mid q) \mid M_D) = P((p \mid q) \mid M_I)$. From this model, it follows that *q* cards should never become informative, which is completely opposed to the empirically found predominance of *p* and *q* selections in standard descriptive WSTs.

Table 12
Basic Model of Laming (1996) With an Unintuitive Dependence Model (M_D)

(a) M_D	*q*	*non-q*	
P	*p q*	0	*p q*
non-p	(1 - *p*) *q*	(1 - *q*)	*q-pq-* q^2+pq^2
	q	(1 - *q*)	1

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

Note: In the cells of M_D and M_I the following notation is used: $p := P(p); q := P(q)$; $p := \mathrm{P}(p), q := \mathrm{P}(q)$.

Laming (1996) concluded that the specific construction of Oaksford and Chater's Bayesian model cannot be warranted. The setting of parameters is interpreted as post hoc data fitting, designed to preclude the prediction of infrequent *non-p* selections and to accommodate for found *q* selections. Hence, Laming rejected Oaksford and Chater's approach and instead postulated a return to the falsificationist norm of correct *p* and *non-q* answers (cf. Gebauer & Laming 1997, Osman & Laming 2000).

This fundamental objection is not ruled out by introducing an exception parameter or by allowing for a different independence model (Oaksford and Chater, 1996, 1998b, cf. 2003). Whereas Oaksford and Chater (1996, particularly, p. 386) have defended their general constancy assumptions (cf. Oaksford and Chater, 2003), I would abandon the concept of a universal model of a conditional. However, the Bayesian model approach in my view needs not to be given up completely. Instead the used constancy assumptions have to be based on knowledge about the constraints given in a specific situation (see Chapter 5, 6, cf. v. Sydow, 2002).

Laming did not discuss the arguments of Oaksford and Chater (1994, 1996, 1998) in favour of the constancy assumption $P((q \mid non\text{-}p) \mid M_D) = P((q \mid non\text{-}p) \mid M_I)$. I first give a brief account of the reasons put forth by Oaksford and Chater, and, secondly, I lay out why Oaksford and Chater's arguments are inconclusive (v. Sydow, 2002).

Oaksford and Chater (1994, 610; 1998a, 240) with reference to Quine (1959) briefly reasoned that "a conditional rule makes a claim about the consequent only if the antecedent occurs but makes no claim otherwise". They continued that "conditional sentences do not assert a conditional but, rather, assert the consequent, *q*, conditional on the antecedent, *p*". In their reply to Laming (1996), in which they ruled out several other critical points, they in this regard only repeated their earlier argument: Psychologically the constancy assumptions reflect "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" (Oaksford & Chater, 1996, 386). In summary, the constancy assumptions of Oaksford and Chater (1994, 1998a) rest on the general assumption that the conditional 'if p then q' only makes a claim about *q* | *p*, and never about *non-p* cases. (Cf. recently Over, in press, and by Evans, Headley, and Over, 2003.)

However, the assumption that *non-p* cases are always psychologically irrelevant is neither empirically founded (cf., e.g., Kirby, 1994), nor normatively coherent (cf. v. Sydow, 2002). Firstly, there are context conditions under which the truth of a conditional can change the assumed probabilities of *non-p* cases as well (see Chapter 5, 6). Secondly, it is inconsistent with the spirit of the Bayesian approach to exclude *non-p* cases only because participants find them psychologically irrelevant, since Oaksford and Chater (1994) predicted also *non-q* selections although they were psychologically implausible. Finally, the postulated general model for testing conditionals violates the fundamental equivalence assumption that equivalent hypothesis are confirmed by the same instances (cf. Section 3.2; pp. 40 f.). It has been argued before that this problem can only be removed, if a knowledge-based approach is adopted (see pp. 53 f., 55 f.).

Empirical Problems – Many Experiments, Few Corroborations

Empirically, the formulation of the theory of Oaksford and Chater (1994, cf. 2003) has mainly been founded on meta-analyses (Oaksford & Chater, 1994, 1998a, 2003; Oaksford, 2002) and reinterpretations of WSTs, originally carried out by other authors (Oaksford & Chater, 1994, 1995, 1996, 1998a, 2003).

These reinterpretations are debatable and have remained controversial, since the experimenters themselves often interpreted their results differently (often even opposed to the Bayesian approach) and the assumed frequency manipulations had actually often been confounded with other factors (see, e.g., Sperber, Cara & Girotto, 1995, Oaksford and Chater, 1995a, but Oberauer, Wilhelm, & Diaz, 1999, 128; or Almor & Sloman, 1996, Oaksford & Chater, 1996, Almor & Sloman, 2000, 1061; or Feeney & Handley, 2000, Oaksford, 2002, Handley, Feeney & Harper 2002).

Likewise, also direct tests of frequency effects by other authors have remained inconclusive: Kirby, 1994a (cf. Oaksford & Chater, 1994; Over & Evans, 1994; Kirby, 1994b); Green, Over, and Pyne, 1997 (cf. Oaksford, 1998; Green & Over, 1998); Green and Over, 1998 (cf. Oberauer, Wilhelm & Diaz, 1999); and Yama (2001). Taken together these results do support the existence of some frequency effects, but these effects appear to be unstable and context dependent. None of these papers provided clear confirmation of Oaksford and Chater's (1994, 1998a) exact model.

Some evidence may be seen to be rather supportive for a revised model with fixed marginals, which will be treated in Chapter 5 (Oaksford & Chater, 2003; cf. v. Sydow, 2002, 2004). In that Chapter the previous results from other authors, which in part at least seem to provide support for this *revised* general Bayesian model, will be discussed in detail (Kirby, 1994; Greene, Over, & Pyne, 1997; Green & Over, 1997, 2000; Oaksford et al., 1999, Exp. 1; Oberauer et al., 1999, Exp. 1; Hattori, 2002; Oaksford & Wakefield, 2003; see pp. 83 f.). A closer examination of these experiments shows that all these studies remain problematic for a set of different reasons. The authors, for instance, used sequential designs that have different normative Bayesian solutions than the WST or, likewise, they made use of tasks for which it was totally reasonable to understand the conditional as a biconditional, which renders *non-p* answers rational even from a falsificationist viewpoint.

In any case, several papers even clearly disconfirmed the Bayesian account of Oaksford and Chater: Gebauer and Laming (1997), Moshman and Geil (1998), Oberauer, Wilhelm, and Diaz (1999), Stanovich and West (2000, cf. Oaksford & Sellen 2000), Osman and Laming (2001), Handley, Feeney and Harper (2002), and Oberauer, Weidenfeld and Hörnig (2004).

More important, even their own experimental tests (Oaksford et al., 1997, 1999, 2003) have failed to provide conclusive support for their model (Oakford and Chater, 1994, 1998a, or 2003). The results of Oaksford, Chater, Grainger und Larkin (1997) were achieved only by using a sequential RAST (reduced array selection tasks), not a WST; but this sequential task again requires a quite different Bayesian analysis so that it cannot count as direct evidence for the WST model postulated by Oaksford and Chater.[25] – In a second main experimental paper Oaksford, Chater and Grainger (1999) only clearly confirmed that the model parameter 'prior belief in the hypothesis', $P(M_D)$, had *no* influence on card selections. However, this result is only relevant for distinguishing between the information gain measures advocated either by Oaksford and Chater (1994) or by Evans and Over (1996; cf. Klauer, 1999; see pp. 71); in regard to the more fundamental question concerning the postulated frequency dependence of *q* and *non-q* selections the results were only mixed. Oaksford et al. (1999, 235) conceded that their "attempts to manipulate the probabilities in these experiments have only been partially successful. No single experiment in this sequence has produced all the effects predicted by probabilistic approaches." – Recently, Oaksford and Wakefield (2003) have reported one single experiment, which indeed seems to be fully in line with their refined model (cf. Chapter 5). However, it has been objected that again a sequential task was used which led to problematic confounds (Oberauer et al., 2004). Also in my view, this experiment is highly problematic and cannot count as a proper confirmation of that model (cf. pp. 89 f. for details).

This review of the empirical situation shows that the Bayesian model of Oaksford and Chater (1994, 1998a, cf. 2003) is highly debatable (for more details cf. pp. 83 f) and that support for it is mainly due to reinterpretations of former results – which are contentious themselves. On the one hand, there were many results, which seem to be incoherent with a Bayesian approach, while on the other some kind of frequency effects have been observed. There are frequency effects, but they are not as systematic and consistent as postulated by Oaksford and Chater (1994, 1998a, cf. 2003). Two conclusions can be derived from this review of the empirical situation. Firstly, more experiments are needed to investigate the context conditions under which one may

[25] In contrast, Oaksford et al. (1997) initially interpreted the results as directly supportive for their WST model. The results showed only minor discrepancies. However, a RAST is clearly a different task from a WST, because the same card may be selected more than once and subjects get feedback when cards are actully turned over. Oaksford and Chater (1998b, cf. Oaksford & Chater, 2003) conceded that the original model of the WST cannot directly be applied to this sequential sampling task (cf. also Klauer, 1999, Chater & Oaksford, 1999a). In my view the results of Oaksford et al. (1997) may well be due to the specific sequential sampling procedure and cannot count as a direct confirmation of their model of the WST.

perhaps obtain different results. Secondly, it appears reasonable to look for additional factors that may explain the found variety of results. One relevant potential confounding factor, in my view, may be the constancy assumptions participants actually make, based on their general knowledge or on features about the actual situation. This leads us to the knowledge-based account outlined and tested in Chapter 5.

5 Towards Knowledge-Based – but Normative – Bayesian Modelling

According to the knowledge-based Bayesian approach advocated here, the preconditions of the model by Oaksford and Chater should not be assumed general properties of testing a conditional, but properties that follow from a given situation or from previous knowledge activated in this situation. Philosophically, this knowledge-based approach is founded on the considerations in Chapter 3, where it was argued that only a knowledge-based Bayesian approach may solve Hume's and Hempel's problems of induction. More directly, this approach builds on the critique of the fundamental constancy assumptions, made by Laming in particular (1996; see pp. 73 f., cf. pp. 70 f.), and on the problematic empirical situation for any universal Bayesian proposal (see pp. 70 f.).

In this chapter, a Bayesian model with fixed marginal probabilities is advocated which additionally implies *p* versus *non-p* frequency effects (cf. v. Sydow, 2002; Oaksford & Chater, 2003). Advocating a knowledge-based account, the assumed constancy assumptions and other assumptions should be induced empirically using a *many cards selection task* (see pp. 83 f.). The aim is to obtain more positive evidence with this kind of task than has mostly been achieved with previous tasks, which have not intentionally induced the preconditions of this model (pp. 70 f., pp. 83 f.).

In Section 5.1 a model with fixed marginal probabilities is introduced and its predictions are presented. In Section 5.2 all previous results are discussed, showing that evidence up to now has rather been negative or inconclusive. In Section 5.3 the many cards selection task (MST) is described as a possibility to induce the model preconditions empirically. In Section 5.4 previous results of von Sydow (2002) are presented. In the Sections 5.5 and 5.6 two new experiments are reported, using MSTs. Experiment 1a and 1b uses ravens material for the first time, linking the debates about the paradox of the ravens and the WST debate. Experiment 2a and 2b uses letter-number material from the original WST by Wason (1966). The results are also discussed with respect to other theories of the WST.

5.1 Sydow (2002) Model – An Alternative Model with Fixed Marginal Frequencies

Von Sydow (2002) accepted Laming's critique (1996), but instead of giving up a Bayesian account altogether advocated to base Oaksford and Chater's Bayesian approach on knowledge about the constraints given in a situation.

Von Sydow (2002) first fully elaborated and tested the alternative Bayesian model with fixed marginal probabilities in the wake of Oaksford and Chater's refined quantitative modelling approach and achieved clearly positive results. This model of testing the conditional, $p \rightarrow q$, has fixed probabilities $P(p)$ and P(q) independent of the truth or falsity of the conditional. If a substance p is claimed to cause cancer q, one may know in advance the probability of cancer in a population, independent of the truth or the falsity of the conditional hypothesis. In this case, $P(cancer)$, as a model parameter, should not vary as a function of the truth or falsity of the hypothesis, but is given independently; $P(q)$ is fixed.

This model has actually long been known about from the philosophical literature on the raven paradox (cf. Chapter 3)[26]. Also in the WST debate, *non-p* cases were discussed earlier. This discussion was triggered by results of Kirby (1994a). Kirby, who still advocated a probabilistic-falsificationist approach, treated these *non-p* selections only as an indicator of a biconditional understanding of the conditional (Kirby, 1994b, cf. Green, Over and Pyne, 1997). Also other authors pointed out, that *non-p* cases might be informative for testing a conditional (Evans & Over, 1996, 360; Green, Over and Pyne, 1997, 219; Green & Over, 1998, 191, cf. particularly Green & Over, 1997, 466, 484, and more generally Over and Jessop, 1998).

Green and Over (2000) elaborated these rather informal suggestions more formally. But firstly, this was done in their own terms, not using the refined information gain formalism of Oaksford and Chater. Secondly, and most importantly, this account has been formulated and tested (cf. p. 83) in the context of *probabilistic* – not deterministic – causal conditionals and for probabilistic hypotheses even a falsificationist cannot normatively postulate *p* and *non-q* answers.[27] Thirdly, Green and Over have not formulated their model in terms of parameter-independent constancy assumptions. Fourthly, they have not empirically introduced the preconditions of such models and have not obtained results (apart from interesting post hoc explanations) which would clearly corroborate a particular a priori model (Section 5.2). Finally, some of the mentioned authors more recently seem to have come to advocate a rather

[26] Nickerson (1996) linked these debates and he was the first who at least implicitly transferred the dependence model with fixed marginals from the ravens debate to the WST debate. However, Nickerson only used example values corresponding to this model and not its general formulation. Moreover, he did not generally model the *EIG* values. Finally, he did not actually test his model empirically.

[27] Even without adopting Neyman-Pearson's test theory, also Fisher's 'falsficationist' test theory of course is concerned with testing *p* and *non-p* cases. This is essential to the general idea of an experimental condition and a control condition. Cf., e.g., Erdfelder & Bredenkamp, 1994; Hager, 2002.

universal account of understanding conditionals (cf. Evans, Handley & Over, 2003; Over & Evans, 2003; Over, 2004; Evans, Over, Handley 2005).

Von Sydow (2002) adopted the basic model for testing deterministic conditionals from the raven paradox debate and discussed its predictions in detail, building on Oaksford and Chater's (1994) refined Bayesian formalism and combined this with a knowledge approach (cf. Over and Jessop, 1998). At about the same time Hattori (2002, issue 4) also proposed a model with fixed marginals, likewise building on Oaksford and Chater's refined approach. Also Oaksford and Chater (2003) and Oaksford and Wakefield (2003) in a complete revision turned to this model for the WST.[28] However, these other proposals did not stress that the assumptions of the model for testing a conditional need to be bound to the constraints given in the situation. Both Hattori (2002) and Oaksford and Chater (2003) postulated a model with fixed marginals as a general model for testing any conditional and not as a particular knowledge-based model. Although this difference matters for the evaluation of experiments, this does not matter for the mathematical model itself.

In this section, I will describe the Bayesian model with fixed marginal probabilities, and then I will outline the resulting predictions.

The Sydow Model

Von Sydow (2002, cf. the other authors mentioned above) first elaborated and investigated a model of the WST in which the resulting marginal probabilities $P(p_{res})$ and $P(q_{res})$ are set to be constant in both sub-models. If the preconditions of this model are met, under certain conditions even a clear increase of *non-p* card selections can be predicted (see p. 81). The constancy assumption $P(p_{res} \mid M_D) = P(p_{res} \mid M_I)$ and $P(q_{res} \mid M_D) = P(q_{res} \mid M_I)$ differs from Oaksford and Chater's original assumptions (1994; 1998), which instead asserts that $P(p_{res})$ and $P(q \mid non\text{-}q)$ are identical in both submodels (cf. Table 11. p. 60).

Like in the original model of Oaksford and Chater (1994), also here some additional assumptions need to be made. Firstly, the alternative hypothesis, stating that the tested conditional is not true, is modelled by an independence model (M_I).

[28] Oaksford, Chater & Larkin (2000) introduced a seemingly similar model, but this was done for a different task (cf. v. Sydow, 2002). For the WST a revision has first been briefly mentioned – without any reasons and without own data – in an overview article (Oaksford & Chater 2001, p. 353), which cannot count as a full revision of their model. Oaksford and Chater (2002, issue 3) replied to Feeney and Handley (2000) and again used this refined model without much further discussion.

Secondly, the hypothesis that the conditional is true is understood deterministically here in the sense of an implication $P(p \wedge \mathit{non\text{-}q} \mid M_D) = 0$.[29]

Later, the aim will be to establish all these assumptions using corresponding instructions. Given these assumptions, the basic model in Table 13 can be derived.

Table 13
Basic Model of Sydow (2002, Oaksford & Chater, 2003) of the Conditional With Fixed Marginal Probabilities $P(p_{res})$ and $P(q_{res})$

(a) M_D	q	*non-q*	Marg.
p	p	0	p
non-p	$q - p$	$1 - q$	$1 - p$
Marg.	q	$(1 - q)$	1

(b) M_I	q	*non-q*	Marg.
p	$p\,q$	p $(1 - q)$	p
non-p	$(1 - p)$ q	$(1 - p)$ $(1 - q)$	$1 - p$
Marg.	q	$1 - q$	1

Note: In the cells of M_D and M_I the following notation is used: $p{:=}\,P(p)$, $q{:=}\,P(q)$

In a 2 × 2 matrix with fixed marginals the probabilities of the four cells only have one degree of freedom (1 *df* in each sub-model). Hence, with the additional constraint that $P(p \wedge \mathit{non\text{-}q} \mid M_D) = 0$, the above dependence model follows with necessity (see Table 16).

Table 14
Illustration of the Shifts within the Dependence Model Described in Terms of the Independence Model.

(a) M_D	Q	*non-q*	
p	$P(p \wedge q \mid M_I) +$ $P(p \wedge \mathit{non\text{-}q} \mid M_I)$	$P(p \wedge \mathit{non\text{-}q} \mid M_I)$ - $P(p \wedge \mathit{non\text{-}q} \mid M_I) = 0$	$P(p)$
non-p	$P(\mathit{non\text{-}p} \wedge q \mid M_I)$ - $P(p \wedge \mathit{non\text{-}q} \mid M_I)$	$P(\mathit{non\text{-}p} \wedge \mathit{non\text{-}q} \mid M_I) +$ $P(p \wedge \mathit{non\text{-}q} \mid M_I)$	$P(1-p)$
	$P(q)$	$P(1 - q)$	1

In the model with fixed marginals, in contrast to Oaksford and Chater's original model (1994, 1998a), the probabilities for *non-p & q* and for *non-p* ∧ *non-q* differ between the dependence and independence model (cf. Table 13 and Table 16).

[29] The advocated knowledge-based approach is in accordance with any knowledge-based variation also of these less fundamental aspects of the model. However, only for deterministic hypotheses there is a clear alternative norm how hypothesis should be tested.

The Predictions: *Non-P* Effects

In order to derive the prediction of this model with fixed marginal probabilities, further modelling steps were constructed analogously to Oaksford and Chater (1994, 1998) (see pp. 62 f; cf. pp. 71 f. for a critical discussion).

Figure 5 shows the resulting expected information gain values (*EIG*) for the Sydow model. The *scaled* expected information gain values (SEIG) would lead to the similar predictions and need not to be presented here. The predictions are given for the main model parameters *P*(*p*) and *P*(*q*).[30]

Figure 5a depicts the parameter values under which the model is applicable. *P*(*p*) ≤ *P*(*q*) needs to be assumed, since if there were more *p* cases than *q* cases there would necessarily be *p* cases without *q* on the reverse side, falsifying the implication '*p* → *q*'. Although the implication is formally not violated by *P*(*p*) = *P*(*q*), these cases are also excluded, because under this parameterisation the Bayesian model of a conditional becomes identical to the model of a biconditional.

In Figure 5b *P*(*q*) was varied, while *P*(*p*) = 0.1 was kept constant. In regard of the resulting *EIG*(*p*) versus *EIG*(*non-p*) information gain values, there is no change in the rank order between conditions; *EIG*(*p*) generally remains predominant. Nonetheless, there should be frequency effects for *EIG*(*q*) versus *EIG*(*non-q*) choices, since in the high probability conditions it becomes the case that *EIG*(*non-q*) > *EIG*(*q*). Additionally, *EIG*(*p*) at least approaches *EIG*(*non-q*) when *P*(*q*) is high. These latter effects concerning the two falsifying cards, *p* and *non-q*, might also be predicted on probabilistic falsificationist grounds (cf. Kirby, 1994a, 1994b), and should not be investigated here.

[30] The model parameter $P(H_D)$ should largely be irrelevant as in the model of Oaksford and Chater, 1994. For the present calculations and to model the experiments this parameter was set to 0.5. For more details see pp. 71 f.

Model Behaviour of the Sydow Model

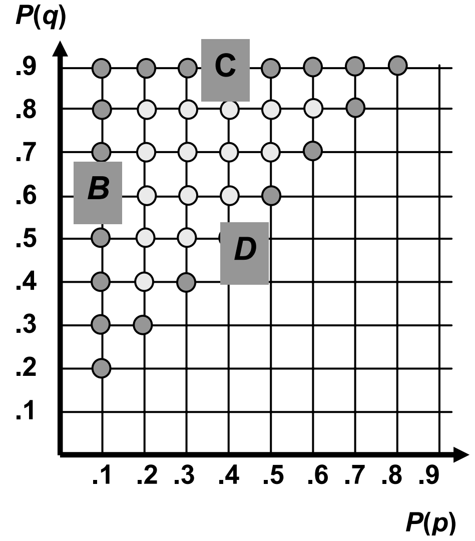


Figure 5a. At each co-ordinate of $P(p)$ and $P(q)$ a circle indicates where the precondition $P(q) > P(p) > 0$ is fulfilled. Predictions for the edges of the triangle are given in the figures 5b, c, d.

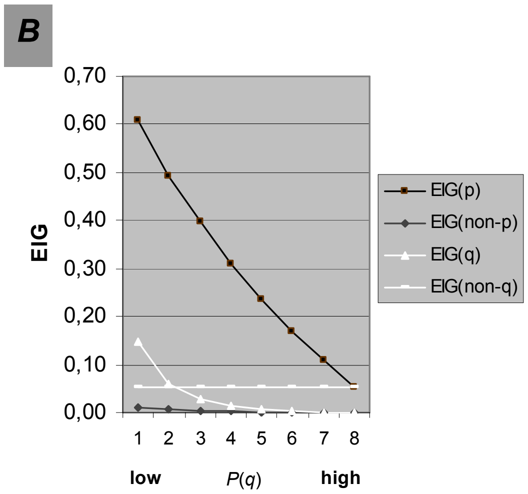


Figure 5b. *EIG* values of the four cards for the constant model parameter $P(p) = 0.1$ and a varying $P(q)$: '1' represents $P(q) = 0.1$; '2' represents $P(q) = 0.2$, etc.

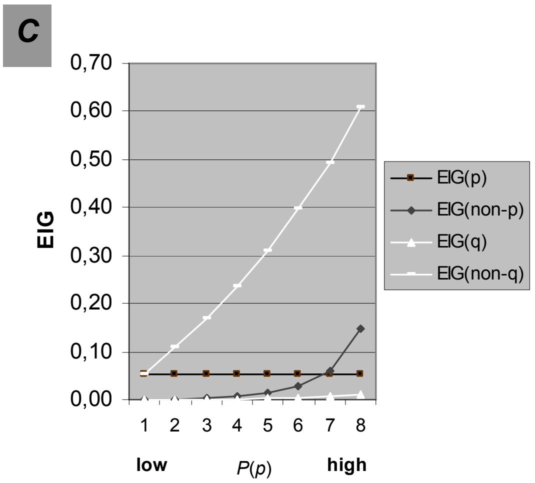


Figure 5c. *EIG* values of the four cards for the constant model parameter $P(q) = 0.9$ and a varying $P(p)$: '1' represents $P(p) = 0.1$; '2' represents $P(p) = 0.2$, etc.

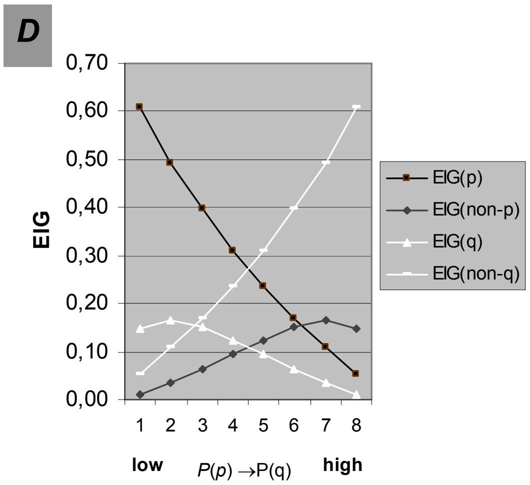


Figure 5d. *EIG* values for the four cards for the low and high model parameters $P(p)$ and $P(q)$: '1' represents $P(p) = 0.1$, $P(q) = 0.2$; '2' represents $P(p) = 0.2$, $P(q) = 0.3$, etc.

Figure 5c shows the *EIG* values if *P*(*p*) is varied, while P(*q*) = 0.9 was constantly kept. Although *EIG*(*non-q*) is always higher than *EIG*(*q*), in the high probability conditions, *EIG*(*non-p*) becomes more informative than *EIG*(*p*).

In a nutshell, the order of *EIG*(*p*) versus *EIG*(*non-p*) can be switched if *P*(*p*) is changed and the order of *EIG*(*q*) versus *EIG*(*non-q*) can be switched if *P*(*q*) is changed.

In Figure 5d the *P*(*p*) and *P*(*q*) parameters of '$P(p) \rightarrow P(q)$' are changed simultaneously ('0.1 → 0.2', '0.2 → 0.3', … , '0.8 → 0.9'). Here both the *EIG*(*p*) versus *EIG*(*non-p*) values and the *EIG*(*q*) versus *EIG*(*non-q*) switch from the low to high probability conditions. Hence, the main predictions of this model can be summarised in the following way:

- Low probability conditions: *EIG*(*p*) > *EIG*(*non-p*); *EIG*(q) > *EIG*(*non-q*)
- High probability conditions: *EIG*(*p*) < *EIG*(*non-p*); *EIG*(q) < *EIG*(*non-q*)

The main difference to the prediction of the original model of Oaksford and Chater (1994, 1998a) is the additional prediction of *non-p* effects if *P*(*p*) is high.

This may be seen to be a daring prediction, because it is incoherent with logical-falsificationist accounts, and since Laming (1996) has accused Oaksford and Chater (1994) of having chosen their original constancy assumptions exactly to fit the data of most previous experiments showing the irrelevance of *non-p*.

5.2 Former Empirical Results in Relation to the Model with Fixed Marginals

The empirical results for a Bayesian account of the WST have generally been reviewed before, showing a rather incoherent or problematic situation (see pp. 70 f.). Here we are concerned with the model with fixed marginals (v. Sydow, 2002; Oaksford & Chater, 2003), which makes the additional prediction of *p* versus *non-p* frequency effects and represents another deviation from falsificationist predictions. Here all main experiments, of which I am aware and in which *non-p* frequency effects were actually found, whether intentionally or unintentionally, will be discussed in detail. It will be shown that even for this subset with seemingly partially positive results, no experiment can count as clear evidence in favour of the model with fixed marginals. The results of von Sydow (2002), which were explicitly based on inducing fixed marginal probabilities *P*(*p*) and *P*(*q*) in the instruction, will be reported subsequently (see pp. 94 f.).

(a) Kirby (1994a) continued to advocate a 'probabilistic falsificationst' position and only expected frequency effects within the correct falsificatory *p* and *non-q* cards. In three experiments with descriptive abstract WSTs he varied *P*(*p*) without specifying *P*(*q*). Contrary to his predictions he not only found some frequency effects regarding *p* versus *non-q* (also these results only partly confirmed his predictions), but also other frequency effects as well. Actually significant *non-p* effects were found in two of the experiments. In retrospect, some results seem to correspond to patterns, which would be predicted by the Sydow model.

However, the results cannot be taken as clear positive evidence for that model. Firstly, none of the three experiments simultaneously supported *all* predictions of the model and some results clearly would have been incoherent with this model (e.g., significant *q* frequency effects in an unpredicted direction in Exp. 3). Correspondingly, the scenario describing a machine, which is printing cards, does not make clear whether fixed marginal probabilities should be assumed. Furthermore, because subjects were informed only about *P*(*p*) and not about *P*(*q*), and the card printing scenario suggests the assumption that $P(p \mid M_D) = P(q \mid M_D)$, subjects may well have adopted a biconditional interpretation of the hypothesis; and for a biconditional interpretation *non-p* selections would be rational even from a falsificationist viewpoint. Actually, Kirby himself (1994b, cf. Johnson-Laird & Byrne, 1991, 80; Over & Evans, 1994) proposed this explanation to account for the *non-p* selections in his experiments. Hence, the findings of Kirby provide no clear evidence with respect to the Bayesian Sydow model.

(b) Green, Over and Pyne (1997) investigated the testing of hypotheses like "if it is a work day, then the manager is in London". They also only varied *P(p)*, not *P*(*q*), using categories which differ in their probability (work days and weekend days). In Experiment 1, they found some probabilistic effects and indeed one significant *non-p* effect. However, for our purposes the experiment is problematic for two reasons. Firstly, neither the story nor the used probabilities excluded a biconditional understanding of the hypothesis. On the contrary, by explicitly asking the participants which combinations of cases may count as counterexamples, Green et al. was able to distinguish two groups of participants, one with a conditional and another with a biconditional understanding of the hypothesis. When Green et al. analysed the data separately for the conditional and the biconditional group, the *non-p* effect disappeared for the conditional group and only occurred in the group with a biconditional understanding. Hence, also a probabilistic falsificationist account (Kirby, 1994a, 1994b) may explain the data. Secondly, the direction of the found *non-p* effect in Experiment 1 was contrary to the predictions of the Sydow model and in Experiment 2 all probabilistic effects in the WST vanished. In my opinion, these results may be due to specific problems of the instructions (cf. Oaksford, 1998).

Hence, in any case this paper does not provide confirmation of the Sydow model or any other formalised Bayesian model of the WST.

(c) Green and Over (1997) also found *non-p* selections in two experiments testing the hypothesis "If a person has Zav's disease then they have a raised temperature." But also these experiments clearly cannot count as a confirmation of the Sydow model: Firstly, the selections in these tasks were not concerned with individual cards but with whole logical classes. However, to select, for instance, the classes of all *p* cards and of all *non-p* cards implies an exhaustive test of all individuals! Hence, even from a falsificationist viewpoint such a strategy would be rational and can be explained without a non-falsificationist Bayesian model. Secondly, although Green and Over's findings proved that participants also selected *non-p* cards, the results neither document frequency effects in contrast to other conditions nor a predominance of *non-p* selections. Hence, these experiments cannot be taken to provide a substantial confirmation of *p* versus *non-p* frequency effects.

(d) Green and Over (2000) indeed reported *non-p* frequency effects for a WST when participants had to test the rule "if you drink from the well then you will get cholera."

Nonetheless, also these findings cannot be interpreted to provide confirmation of the Sydow model.

Firstly, a biconditional interpretation is again plausible, since no precise probabilities for $P(p)$ and for $P(q)$ are provided and P(drink from the well) and P(get cholera) are modified by the use of identical words for both probabilities. This renders $P(p) = P(q)$ not only possible, but plausible. Moreover, no other cause of cholera is mentioned, which may have excluded this interpretation. Hence, like previously the full Bayesian account is not necessarily needed to explain the *non-p* selections found here (see, e.g., Kirby, 1994b, 249; Green, Over & Pyne, 1997; cf. Johnson-Laird & Byrne, 1991, 80).

Secondly, the causal rule, which was to be assessed, may plausibly be interpreted in a probabilistic way. But a probabilistic interpretation would anyway render the falsificationist normative solution inapplicable (see also p. 78, footnote 27).

Finally, the results were partially based on 'category' WSTs instead of 'individual' WSTs. In the 'category' conditions participants had to select whole classes of cards (like in Green & Over, 1997, above). In the category condition again, a *p* selection, testing all who drunk from the well, together with a *non-p*-selection, testing all those who didn't, provides the participants with complete knowledge about the diagnosis of all villagers (and all counterexamples of the conditional or biconditional). Hence, even from a falsificationist perspective the seemingly 'correct' *p* & *non-q* selection has no advantage over other selections, like *p* & *non-p*, in the category condition. Hence, found effects in this condition need not provide evidence for a Bayesian account. Additionally, Green and Over only tested the frequency effects for the *non-p* cards over all conditions (including category WSTs and individual WSTs), but a reanalysis of the data shows that the *non-p* effect was not significant in the standard individual selection conditions ($n = 78$, $\chi^2(1) = 1.01$).

Although these finding may be interesting in their own right, Green and Over (2000), hence, do not provide any support for *non-p* frequency effects in the sense of the Bayesian models investigated here.

(e) Oaksford, Chater and Grainger (1999) tested their original Bayesian model in four experiments but conceded that their "attempts to manipulate the probabilities in these experiments have only been partially successful. No single experiment in this sequence has produced all the effects predicted by probabilistic approaches" (p. 235). Nonetheless, although they had not predicted *non-p* effects by that time, Experiment 1 showed a strong increase of *non-p* selections in the high probability condition relative to the low probability condition, as it would have been expected on the basis of a model with fixed marginals (v. Sydow, 2002, Oaksford & Chater, 2003). But there are three objections to such an interpretation:

Firstly, this effect may be due to a biconditional understanding of the hypothesis. In this experiment, Oaksford et al. varied the cards and the formulation of the tested rules in order to introduce different frequency conditions. In the relevant high probability condition participants should test the hypothesis "if an item of furniture is heavy then it is big" by using the cards 'heavy', 'light', 'big', or 'small'. Here no information was given which indicated that $P(p) < P(q)$. Moreover, the empirical probability estimations of the participants themselves also indicate a biconditional understanding. But given a biconditional hypothesis, *non-p* selections are equally predictable from a falsificationist viewpoint and, hence, the results cannot be properly reinterpreted as being supportive for a model with fixed marginals.

Secondly, Oaksford and Chater advocated a universal account of testing conditionals and did not control for possible constancy assumptions in their experiments. Hence, from a knowledge-based perspective these experiments are difficult to interpret.

Thirdly, since Oaksford and Chater in their original and in their refined model (1998a, 2003) advocated a universal account of testing conditionals, they in principle fail to account for any differences between the experiments. In any case, no *non-p* frequency effect was found in the other three experiments.

In conclusion Oaksford, Chater and Grainger's results (1999) show some probability effects, but they neither clearly suggest the truth of their original universal model nor of their alternative universal model.

(f) McKenzie and Mickelsen (2000) have demonstrated that the informativity of the combination of *non-p* & *non-q* observations increases if these observations are rare (high probability conditions of *p* and *q*). This is coherent with the Sydow (2002) model, without formulating this model explicitly. Moreover, McKenzie and Mickelsen (2000) did not utilise a WST and worked with another task were no card needed to be selected (rather similar to a truth table task cf. Kao & Wassermann, 1993; Gebauer & Laming, 1997; White, 2000; Evans, Handley, Over, 2003; Barres & Johnson-Laird, 2003; see also p. 4).

Therefore, although McKenzie and Mickelsen's (2000) interesting results do indeed seem to be compatible with the Sydow model, their results do not provide any *direct* test of theories concerned with the WST.

(g) Oberauer, Wilhelm and Diaz (1999) directly tested Oaksford and Chater's account (1994) in three experiments and came to the conclusion that "optimal data selection does not explain the Wason selection task" (141). Although their comparatively complicated Experiments 2 and 3 appear to count against any Bayesian theory – the subjective probabilities of the cards did not affect card selections at all –, their Experiment 1, using a context story very similar to Kirby (1994a), led to *non-p* and *non-q* effects. Although this does not corroborate the original Oaksford and

Chater's model (1994), it may be reinterpreted as providing confirmation of the model with fixed marginal probabilities investigated here (v. Sydow, 2002; Oaksford & Chater, 2003). Oberauer et al. did not intend to check the assumed constancy conditions (the investigated model of Oaksford & Chater, 1994, even used different constancy assumptions). But from the knowledge-based perspective adopted here, it seems plausible that Oberauer et al. may involuntarily have induced these constraints in this scenario.

However, also this experiment cannot be interpreted as clear support for the model with fixed marginals either: Firstly, it is normatively not clear, which constrains should be assumed. Secondly, because they used probability manipulations for which $P(p \mid M_D) = P(q \mid M_D)$ can be asserted (in the relevant high probability condition), the found *non-p* effects may again be due to a biconditional interpretation only (cf. Kirby, 1994b).

(h) Hattori (2002), at about the same time as von Sydow (2002), proposed a model with fixed marginals and tested *non-p* effects. In one experiment, he also tested this model (a second experiment was concerned with different questions). Unlike von Sydow (2002), Hattori did not advocate that the fundamental constancy assumptions of the model with fixed marginals are knowledge-based. Indeed, in a knowledge-based perspective his metaanalysis, which did not distinguish between different constraints, would have been counterproductive.

In his experiment, Hattori was consistent in making no effort to establish any particular constancy assumption in the instruction: The probability information concerned a large pool of cards, from which a subset was taken, and from which in turn the four cards for the WST were taken. Hence, there was no precise knowledge about the card probabilities in the subset. More important, this scenario allowed also for the alternative constancy assumption $P(q_{res} \mid H_D) > P(q_{res} \mid H_I)$, instead of the assumed $P(q_{res} \mid H_D) = P(q_{res} \mid H_I)$. The scenario was concerned with a person who alleged an if-*p*-then-*q* rule about a hand of cards (referring to the Japanese card game Hanafuda). Here the truth of the claim may well imply an increased number of *q* cards. Corresponding to such a post hoc explanation the results would be consistent with a knowledge-based account, but not with Hattori's universal account, since globally no *non-p* effect was found. A *non-p* effect was only found for a subgroup. Additionally, Hattori himself discusses problems concerning a biconditional interpretation in this experiment. Hence, even the result for the subgroup may be problematic. In any case, Hattori's interesting results cannot count as providing clear confirmation for the model with fixed marginals.

(i) Oaksford and Wakefield (2003) report a single experiment in which they tested the refined general model with fixed marginals (Oaksford & Chater, 2003) and also predicting *non-p* effects. They used the same material as Oberauer et. al. (1999) in their third experiment, for which no probabilistic effect was found. Unlike Oberauer et al. (1999), they used a sequential sampling process. Based on this 'natural' sequential sampling process they provided data which seem to be fully consistent with the model with fixed marginals.

Although the results seem to be fully confirmatory I agree with Oberauer, Weidenfeld and Hörnig (2004, 522, 527) that the results cannot be interpreted as confirmation of the model with fixed marginals. A rational analysis of a sequential task is normatively not equivalent to a rational analysis of the WST. Oaksford and Wakefield (2003) themselves have conceded problems of the sequential procedure used in former RASTs (Oaksford et al., 1997; cf. Oaksford & Chater, 1998b). Although the task used by Oaksford and Wakefield ruled out some of these problems, they did not rule out all problems. The task remains a sequential task in the sense that subjects could select more than one card from each category. In the high probability condition with many *p*s, for instance with the sequence "*p*, *p*, *p*, ¬*p*, *p*, *p*, *p*, ¬*p*", a selection of, say, two *p* cards and one *non-p* card would count in favour of a higher portion of *non-p* cards than *p* cards. Hence, in the data of Oaksford et al. a 'dominance' of *non-p* card selections may still refer to a higher number of selected *p* cards than *non-p* cards. Therefore, even if subjects regarded it as more important to turn over one *p* card than one *non-p* card (if the dependent variable had been one single forced-choice selection), the used dependent variable may falsely indicate that *non-p* cards were preferred. Hence, the *non-p* effects under natural sampling conditions found by Oaksford and Wakefield (2003) may be an artefact of dividing the number of *p* selections in the high probability condition by a large number of opportunities in which a *p* card could have been selected.

Therefore, the experiment of Oaksford and Wakefield clearly cannot count as a proper confirmation of a model of the WST with fixed marginals (v. Sydow, 2002, Hattori, 2002, Oaksford & Chater, 2003).

(j) Recently, Oberauer, Weidenfeld and Hörnig (2004) reported two experiments, which removed the confounding of repeated selection options of the same card typs in the experiment of Oaksford and Wakefield (2003). Although they continued to use a sequential learning phase, as demanded by Oaksford and Wakefield (2003), they did not use the problematic sequential testing phase. In Experiment 1 their frequency manipulation had no effect at all. In Experiment 2 the instruction put more emphasis on the role of the probabilities learned. With these additional instructions *p* versus *non-p* frequency effects were actually found – but the difference went in the 'wrong' direction. Participants quite generally selected cards with a high and not a low probability. In my view, these disastrous results for the Sydow model may be due to the fact that in the learning phase selecting the card of the larger sample had been continuously reinforced upon subjects and they may have transferred this habit to the following WSTs, misinterpreting the additional instruction of this task.

However, although Experiment 2 of Oberauer et al. (2004) showed frequency effects, they only provided evidence against Oaksford and Wakefield's universal model and against the claim that natural sampling may provide the key to better results.

In summary, five main problems have been worked out, why even some studies, whose results may be reinterpreted as being supportive of the Sydow model, cannot count as proper tests of this Bayesian model with fixed marginals. Firstly, despite single positive results most papers remained inconsistent if considering all results together (Kirby, 1994a; Green, Over and Pyne, 1997; Oaksford, Chater and Grainger, 1999; Oberauer et al., 1999, 2004; cf. Chapter 4). Some significant *p* versus *non-p* effects even went in a direction, which is not predicted by this model (Kirby, 1994, Exp. 3; Green, Over and Pyne, 1997, Exp. 1; Oberauer et al., 2004, Exp. 2). Secondly, some experiments cannot count as testing Bayesian frequency effects (Green & Over, 1997, 2000), because the task instructions did not demand the selection of single cards, but of whole classes of cards. In such tasks, for instance, the combined selection of the class of *p* cases and the class of *non-p* cases exhaustively would test all cards. Here the otherwise surprising *non-p* selection becomes trivial; even from a falsificationist perspective any exhaustive test is perfectly rational. Thirdly, some experiments used *probabilistic* hypotheses (e.g., Green and Over, 2000). Although this is interesting in its own right, this cannot decide between a Bayesian and a falsificationist approach, since the falsificationist norm of hypothesis testing is only applicable to deterministic conditionals (implications). Fourthly, in many cases a biconditional understanding of the hypothesis 'if *p* then *q*' was induced by making $P(p) = P(q)$ plausible (Kirby, 1994a; Green, Over & Pyne, 1997; Oaksford, Chater & Grainger, 1999, Exp. 1; Oberauer et al., 1999, Exp. 1; Green & Over, 2000). But a biconditional interpretation of the conditional renders *non-p* selections rational even from a falsificationist perspective (Kirby, 1994b). Finally, the positive results of Oaksford and Wakefield (2003) may be due to a confounding factor, linked to sequential tests (cf. Oberauer et al. 2004).

Overall, none of the WSTs mentioned by Oaksford and Chater, 2003, provided clear support for the particular prediction of the Bayesian model with fixed marginals postulated by von Sydow (2002), Hattori (2002) and Oaksford and Chater (2003). We have seen that all seemingly positive results are vulnerable to criticism. Moreover, I also mentioned some of the directly negative results (see also pp. 83 f.). The problems to interpret the discussed experiments are all inherited to any metaanalysis that has used these data (Hattori, 2002; Oaksford & Chater, 2003). Hence, all studies in which *not-p* effects were claimed to be found remain inconclusive for many different reasons.

5.3 The Many Cards Selection Task (MST) – A Modified Experimental Paradigm

Persuing a knowledge-based approach, here a variant of the WST is introduced in order to induce the preconditions of the model with fixed marginals empirically and to avoid the problems of the tasks discussed in the last section (cf. von Sydow, 2002).

Seven Requirements for the MST

The variant of the WST was constructed to meet seven requirements:

Firstly, the task should provide salient knowledge about the probabilities in a frequency format (cf. Gigerenzer & Hoffrage, 1995). The probabilities should not be questionable as in the stack of cards selection tasks (SSTs) discussed above (Green, Over, Pyne, 1997; Oaksford, Chater, Grainger, 1999; Oberauer et al., 1999, Exp. 2 und 3; Hattori, 2002).

Secondly and importantly, the task should introduce the constancy assumptions of the model with fixed marginals, $P(p \mid H_D) = \mathrm{P}(p \mid H_I)$ and $P(q \mid H_D) = \mathrm{P}(q \mid H_I)$, in a simple and obvious way. In most SSTs this has not been given.

Thirdly, the task should not involve successive testing of cards from the same logical class. I have outlined before that sequential tasks (Oaksford et al. 1997, Oaksford & Wakefield, 2003) require a different normative analysis of the situation (cf. Oaksford & Chater, 1998b; Klauer, 1999; Oberauer et al. 2004; see p. 89).

Fourthly, subjects should select cards repesening individual instances, not whole categories (as investigated by Green, Over & Pyne, 1997; Green and Over, 1997).

Fifthly, it needs to be excluded that a biconditional interpretation is normatively correct. Here only frequencies which fulfil $P(p) < P(q)$ should be used, rendering a biconditional interpretation incorrect (cf. Kirby, 1994b).

Sixthly, a probabilistic interpretation of the tested hypothesis is to be excluded, since in this case the alternative falsificationist norm would anyway cease to be applicable. For this reason, the hypotheses will be formulated explicitly in a deterministic way (cf. p. 78, footnote 27).

Seventhly, the front sides and back sides of the cards need to be clearly distinguished and it should be obvious that *p* or *non-p* can only be on one side of the cards and *q* and *non-q* on the other (cf. Osman & Laming, 2001).

All points have been discussed before, particularly in the context of presenting the theoretical criticism of Laming (cf. pp. 70 f.) and when discussing former empirical results (pp. 83 f.). Interestingly, the fourth to the seventh requirement for a modified WST can also be required from a logical-falsificationist perspective (e.g., Osman and Laming, 2001).

Tasks with Many Visible Cards and Fixed Marginals

To achieve the above requirements von Sydow (2002) modified the WST and introduced a *many cards selection task* (MST).

In a MST – like in a WST – participants should normally test an if-*p*-then-*q* hypothesis (but cf. Experiments 4 to 7). The hypotheses are formulated deterministically as an 'if-*p*-the-*always*-*q*' sentence. We here use the standard letter-number example: "if there is an 'A' on one side of the card then there is *always* a '2' on the other side of a card".

In the MST, more than four cards are visible at the time of selection, although – in contrast to sequential tasks – maximally one card of each ofur logical classes (*p*, *q*, *non-p*, *non-q*) can be selected. The portion of cards in the different logical classes is used to manipulate the probabilities of the classes. There need to be enough cards to represent *P*(*p*) < *P*(*q*) and to secure that no exhaustive test of any class of cards is possible (by selecting only one card). The former requirement is needed to rule out a biconditional interpretation, the latter is needed to rule out an exhaustive test strategy.

Assume there were ten cards. It is pointed out that there is a letter on the one side of each card and a number on the other side. The cards are displayed twice, one time with the front sides (with *A* or *K*, corresponding to *p* and *non-p*) facing upwards and a second time with the back sides (with 2 and 7, corresponding to *q* and *non-q*) facing upwards (Figure 6a, b).

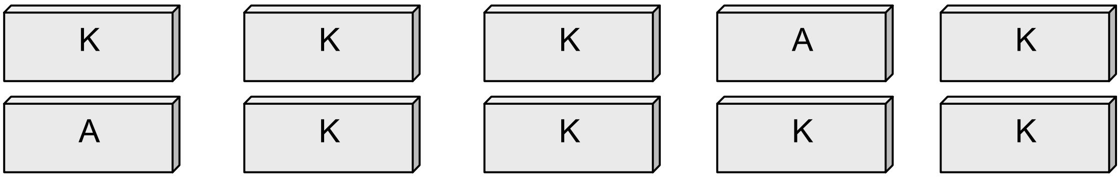


Figure 6a. First display of the front sides of the 10 cards showing *p* and *non-p* cases

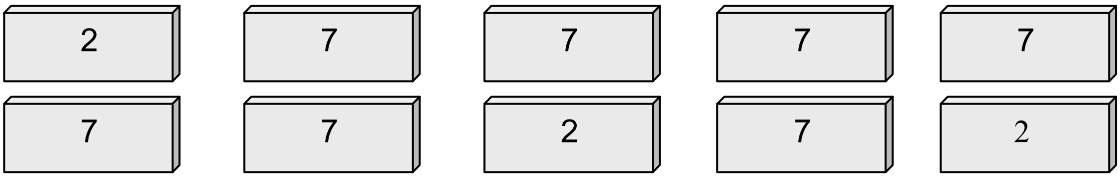


Figure 6b. Second display of the back sides of the cards showing *q* and *non-p* cases

It is made clear that the two displays refer to the same cards. It is stressed that the cards are mixed in between the displays and that the order of each card in the two displays is independent so that true mapping of front to back side is still completely unknown.

However, the display of both card sides should allow the participants to determine *P*(*p*), *P*(*non-p*), *P*(*q*), and *P*(*non-q*) exactly and it should allow us to vary these values experimentally. Using a paper and pencil task, the two card displays should normally remain visible during selection.

In this task and corresponding to the Sydow model, *P*(*p*) and *P*(*q*) can be assumed to be fixed, independent of whether the hypothesis is thought to be true or not – here *P*(*q*) should not increase if the hypothesis '$p \rightarrow q$' is assumed to be true.

Finally, the participants have to select the card(s) they would like to turn over. This is normally done by ticking cards already displayed. There are three reasons for this: Firstly, this prevents any confusion about the number of cards displayed. Secondly, the introduced frequency of cards remains salient during the process of selection. Thirdly, a particular card cannot be misinterpreted to represent a whole logical category of cards (the selection of the 'K' card does not mean a selection of all consonants).

Forced Choice Selection Tasks

In the MSTs used here participants will have to choose one of two cards within a larger display of, say, ten cards. They will either be asked whether they would select either *p* or *non-p*, or alternatively whether they would select either a *q* or *non-q* card. This is a forced-choice task.[31]

A first reason for this forced choice instruction is that one may interpret normal WST as being a sequential task. But in this case Oaksford and Chater's model (1994) may not be strictly valid (cf. Klauer, 1999). Although it may be controversial whether this is problematic for a WST, a forced-choice task is in any case not successive.

Moreover, a forced-choice test provides a more severe test, because any increase of, for instance, *non-p* selections does not only show that they become subjectively somehow informative (as in a WST), but that the subjectively expected information gain of *non-p* is higher, or at least equally high, than that of *p*.

Finally, a probabilistically extended falsificationism may explain frequency effects between *p* and *non-q,* since both still refer to a potentially falsifying card. The two kinds of forced-choice questions used here, which urge a decision between a *p* versus a *non-p* card*,* or between a *q* and a *non-q* card*,* exclude any kind of frequency effects which would be explicable by a refined falsificationism (Kirby, 1994a, b; Humberstone, 1994) as well.

5.4 The Results of von Sydow (2002)

Von Sydow (2002) was the first to fully elaborate the model with fixed marginals and at the same time provided confirmative results for that model, by using MSTs. These results can count as a clear corroboration of this Bayesian model, which is not affected by the problems discussed earlier.

In two MSTs the participants (*N = 128*) were asked to test the truth or falsity of the abstract assertion „if ‚+' than always ‚•' ". This abstract hypothesis was used to exclude effects of previous knowledge. (Even the original letter-number material may elicit assumptions about the probability of vowels and consonants.)

[31] Strictly speaking, this is not the case, because participants may normally also opt to choose neither of the cards. But since no theory predicts such a selection and there were almost no such selections this can be ignored in our present context.

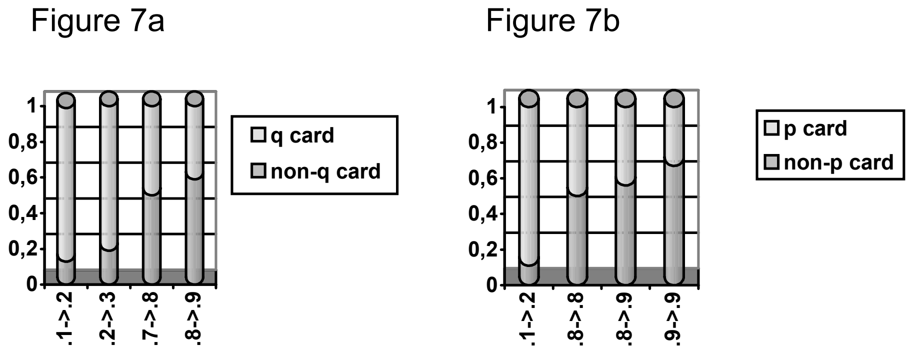


Figure 7a, b. Effect of card probabilities on card selections in von Sydow (2002). (a) Portion of *q* versus *non-q* selections. (b) Portion of *p* versus *non-p* selections.

The first MST was concerned with *q* versus *non-q* selections, varying the probabilities of the cards in four steps. In the first condition the probabilities $0.1 \rightarrow 0.2$ were used, brief for $P(p) = 0.1$ and $P(q) = 0.2$. Condition two used the probabilities $0.2 \rightarrow 0.3$, condition three $0.7 \rightarrow 0.8$, and condition four $0.8 \rightarrow 0.9$. From a falsificationist viewpoint all conditions should equally elicit *non-q* selections, whereas the advocated Bayesian model implies more *q* selections in the low probability conditions ($0.1 \rightarrow 0.2$; $0.2 \rightarrow 0.3$) and more *non-q* selection in the high probability conditions ($0.7 \rightarrow 0.8$; $0.8 \rightarrow 0.9$). The results of this experiment showed highly significant predicted differences between low and high probability conditions (see Figure 7a).

The second MST tested the prediction of an increased number of *non-p* (versus *p* selections) for the first time in a model with explicitly fixed marginals. Again, four conditions were used. Condition one: $0.1 \rightarrow 0.2$; condition two: $0.8 \rightarrow 0.8$; condition three: $0.8 \rightarrow 0.9$; and condition four: $0.9 \rightarrow 0.9$. Hence, we used a conditional ($P(p) < P(q)$) low and high probability condition and two biconditional high probability conditions ($P(p = P(q) \Rightarrow P(p \wedge \neg q) = P(\neg p \wedge q) = 0$). The results showed the predicted difference between the low probability conditional on the one hand and, on the other, the two biconditional conditions as well as the condition with a high probability conditional.

With these two studies, von Sydow (2002) provided evidence that strong *non-p* effects could be obtained if the marginals are fixed (even without sequential sampling). In the light of conflicting evidence in the literature (cf. pp. 70 f.), more support for this finding is needed.

Von Sydow (2002) reported two further experiments. The first experiment used the material of the raven paradox for the first time in a WST. But no *non-p* effects were achieved, since the marginal probabilites had not been fixed. In a fourth pilot study, the letter-number material was used from the original WST (Wason, 1966). Despite some positive effects some unpredicted effects became significant as well. Moreover, this pilot study was problematic to interpret due to possible transfer effects. Here these problems will be ruled out.

In the remainder of this chapter two new experiments will be reported, one with the original ravens material (but now with an MST with fixed marginal probabilities) and one with the original letter-number material (but without any possible transfer effects).

5.5 Experiment 1a, b – The Raven's Paradox and the WST

In Experiment 1 the Bayesian WST debate (Chapter 4 and 5) will be connected to the Raven's paradox debate (Chapter 3), by using raven material. In regard of both debates, I have argued that the preconditions for the Bayesian models need to be established by knowledge or a given context. In Experiment 1 a MST is used to establish the preconditions of the model (Section 5.3). It would be a novel finding to support the Bayesian resolution of the raven paradox by using ravens material in a WST and by showing an increase of non-raven selections in the predicted condition.

The Raven Paradox and the WST

The intimate connection between the raven paradox debate and the WST debate has been realised by only a few authors:

(a) On the theoretical level, Humberstone (1994) elaborated this analogy from a falsificationist perspective and Nickerson (1996) from a Bayesian perspective,[32] but the two debates continued in a rather disconnected way.

The raven paradox is concerned with the question of whether single conjunctive observations, $O_a = p \wedge q$, $O_b = p \wedge \neg q$, $O_c = \neg p \wedge q$, or $O_d = \neg p \wedge \neg q$, are confirmatory, disconfirmatory or irrelevant. In contrast, the WST debate is concerned

[32] Although Oaksford and Chater (1994) mentioned the ravens paradox, they did not point out that their original Bayesian approach was incoherent with the standard Bayesian resolution of the paradox (see pp. 53 f.).

with the turning over of cards, for instance a *q* card, and the resulting hypothetical observations, here either $p \mid q$ or $\neg p \mid q$. In the WST, as in the raven paradox, observations can be confirmatory, disconfirmatory or irrelevant. Moreover, mathematically the evaluation of these hypothetical conditional observations in the WST rests on the evaluation of the conjunctive observations relevant for the raven debate.

(b) Empirically, to my knowledge there has only been one earlier experiment by other authors that was directly concerned with the Bayesian resolution of the raven paradox (McKenzie & Mikkelsen, 2000). But this study neither used a WST nor ravens material.

Only one experiment conducted by von Sydow (2002) has directly investigated WSTs using ravens material. But in that experiment the marginal probabilities had not been clearly fixed and the results were not in accordance with the predictions. In four WSTs, the participants had to check the truth or falsity of the hypothesis ‘all ravens are black’. $P(p)$ and $P(q)$ was manipulated by changing the size of the universe of discourse. The size of the universe of discourse was manipulated subtly by using different labels for the cards. In the high probability (small universe of discourse) conditions, the cards were labelled ‘raven’ (*p*), ‘dove’ (*q*), ‘black bird’ (*non-p*) and ‘white bird’ (*non-q*). In the low probability (large universe of discourse conditions), they were labelled ‘raven’ (*p*), ‘table’ (*q*), ‘a black object’ (*non-p*) and ‘a white object’ (*non-q*). Additionally, the context story was varied describing the cards as referring either to ‘any object’ or to the ‘birds of a particular avarium’. The results showed no significant difference between the conditions. Only descriptively same small positive effects were observable.

Von Sydow (2002) argued that the deviations from the model with fixed marginals, might have been due to not explicitly fixing the marginal probabilities in the story. Additionally, a number of possible further problems were suggested as being potentially responsible for the negative results, most of them connected to the ravens material. For example, Humberstone (1994, 399) has suggested, that regarding ornithology in the field (as opposed to the second-hand ornithology of the birds-and-colours observation-recording cards), it is perhaps hard to take very seriously the distinction between observing a swan as white and observing a white bird as a swan. All you get to see is white swans.

(c) In the following Experiment 1 the aim was to remove all such possible misunderstandings. In contrast to the previous experiment with raven material (von Sydow, 2002) the following additional requirements (cf. p. 91) were met:

- The card-sides that report the species of the birds and the card-sides that report their colour were now presented as clearly distinct sides of observation-recording cards.
- The independence model was more clearly induced as most plausible alternative hypotheses to the independence model.
- In the used MST the probability of $P(p)$ and $P(q)$ are exactly known to the participants.
- The marginal probabilities, $P(p_{res})$ and $P(q_{res})$ were fixed by the instructions. In particular in this MST the constancy assumption $P(\text{black} \mid M_D) = P(\text{black} \mid M_I)$ is made plausible. The truth of the hypothesis that all ravens are black should not increase the assumed number of birds being black.

Method Experiment 1a, b

The goal of Experiment 1 was to obtain the predicted Bayesian frequency effects of the Sydow model for a WST with raven material.

Design and Participants

Experiment 1a was concerned with the dependent variable of *p* versus *non-p* selections, while Experiment 1b investigated the dependent variable of *q* versus *non-q* selections. Formally, two independent experiments (with different participants) were run to exclude any possible effects of sequential effects (cf. Klauer, 1999; cf. my requirement three, pp. 91 f, cf. 93 f.). For both dependent variable (Experiment 1a and Experiment 1b) a low ('0.10 → 0.20') versus a high probability condition ('0.80 → 0.85') was tested in a between-subjects design.

64 students of the University of Göttingen took part in the experiment (56 % female, 44 % male, mean age: 24 years). They were recruited on the campus and participated in the experiment voluntarily. They got a little present and were able to win a prize after finishing the task. The largest number of students studied law (36 %), humanities (11 %), and economics (9 %). None of the participants had prior knowledge of the selection task. 16 participants were randomly assigned to each of

the resulting four conditions of Experiment 1a and 1b. Five participants were excluded from further analysis since they did not follow the instructions. (Four participants selected a card, which they were not permitted to in their conditions. They selected *p* or *non-p* cards instead of *q* or *non-q* cards or vice versa.)

General predictions

Since the used instructions should induce the preconditions of the Sydow model (cf. 5.1) the advocated knowledge-based account predicts both frequency effects, *q* versus *non-q* effects (black versus white birds), and *p* versus *non-p* effects (ravens versus non-ravens). The predictions for the low ('0.10 → 0.20') versus high probability conditions ('0.80 → 0.85') can be derived from the model behaviour (see p. 81).

Table 15
Predicted Increased Portion of Selections Relative to the Corresponding Alternative Probability Condition of the Bayesian Model and the Predictions of Falsificationism and Naïve Inductionism

	Exp. 1a: raven (*p*) vs. non-raven (*non-p*)		Exp. 1b: black (*q*) vs. non-black (*non-q*)	
	Low probability condition	High probability condition	Low probability condition	High probability condition
Bayesian prediction	Raven (*p*)	Non-raven (*non-p*)	Black (*q*)	Non-black (*non-q*)
Falsificationist prediction	Raven (*p*)	Raven (*p*)	Non-black (*non-q*)	Non-black (*non-q*)
Naïve inductionism	Raven (*p*)	Raven (*p*)	Black (*q*)	Black (*q*)

In Table 15, the resulting general predictions are presented. In the high probability condition (relative to the low probability condition) increased numbers of *non-p* and *non-q* selections are expected. In contrast, falsificationism for both conditions equally demands *p* & *non-q* selections (cf. 1.2, 3.1). Likewise, naïve inductionism – without Bayesian extensions – would not predict any frequency effects (cf. Section 3.2).

Excursus: Procedure and Materials in All Experiments

In all experiments in this work, participants were given a little booklet. It always consisted of a front page, a general introduction, one or sometimes many task pages, and a final questionnaire concerning biographical data and comments. The pages of the booklets were very similar in all experiments, apart from the task page(s). Before describing the task pages of Experiment 1, a description of the other pages should be given once for all the experiments. Important differences from the general procedure will be reported for each experiment separately.

(a) On the *front page*, the title of the study and contact information was provided. The title of the experiments in Part II was 'Questionnaire – Styles of Human Hypothesis Testing' and additionally a name indicating the specific material used, for instance 'Raven Experiment'.

(b) The *general introduction page* informed the participants that the task is not an intelligence test, but that we are interested in their style of reasoning. This was done in order to obtain as 'natural' answers as possible. Participants were told to read the instruction carefully. They were asked to take as much time as they needed. They were allowed to use the margins for comments. Since some tasks were conducted in small groups, it was emphasised that the tasks have to be solved individually. This was enforced by the experimenter. The participants were encouraged to ask the experimenter, if they had any questions concerning understanding of the task. Finally, participants were informed that they could win prizes if they completely finished their task.

(c) In the *final questionnaire*, the participants were asked for biographical information about gender, age, and their field of study. It was inquired whether the task had been known to them. In the experiments of Part II, participants were asked whether they were versed in probability theory and formal logic and whether they used formal logic or probability theory, or their judgement/intuition to solve the task. They were always asked to comment on the task. Finally, they were asked whether they wanted information on the task and whether they wanted to take part in the lottery to win a prize.

Editorial note: When describing the material of the experiments, italics is used for any text passages which were somehow highlighted in the original instruction.

Procedure and Materials of Experiment 1

The task was placed in the setting of a zoo were the following deterministic rule was to be tested: "If a bird is a raven, then its feathering is always black". The participants should imagine visiting this zoo with two acquaintances, Carl and Gustav (named in honour of Carl Gustav Hempel, father of the raven paradox). These two acquaintances make two opposed claims, which should correspond exactly the two conflicting hypotheses to be tested in the Sydow model. The instruction read as follows (translation from German):

"With two acquaintances, Carl and Gustav, you are visiting an animal park. Carl describes the beautiful 'black ravens'. Gustav criticises that it is absurd to speak of 'black ravens', since all ravens are black anyway. Carl replied, in such an animal park there may well be ravens that are not black. Adding, it is not even clear, whether the majority of ravens in this park are black. Gustav insists that the ravens in this park are also all black. Finally, they decide to test this empirically.

In the keeper's cabin, there is a card file, in which all birds of the avarium are listed. In this file, 40 birds are listed on 40 cards. The cards are used on both sides. On *one* side of each card, the species is noted. Here it is only noted, whether the bird is a raven (🐦 = a silhouette of a raven) or not (~~R~~ = no raven). On the *other* side the feather colour of each bird is noted (● = black; ○ = symbol for a non-black, e.g., brown or white colour)."

You only aim to test, whether the following sentence is true or false (valid for the cards): *If a bird is a <u>raven</u>, then its feathering is always <u>black</u>.* [set in bold print]"

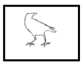
In order to provide information about *P(p)* and *P(q)* and make sure that this is taken as being independent of the truth or falsity of the hypothesis, 40 cards concerning birds were shown as in a standard MST (see p. 92). Firstly, an overview of the 40 *p* and *non-p* card sides was displayed, and then the reversed 40 *q* and *non-q* sides were shown. The number of shown card sides in the low and the high probability conditions is shown in Table 16. In the two displays the following symbols were used: Raven card, , non-raven card, ~~R~~, black card, ●, non-black card, ○.

Table 16
Number of the Displayed 40 Cards in the Low and High Probability Conditions of Experiment 1

	First display		Second display	
	Raven *(p)* cards	*Non raven* *(non-p)* cards	*Black* *(q)* cards	*Non black* *(non-q)* cards
Low probability conditions (‘0.10 → 0.20’)	4	36	8	32
High probability conditions (‘0.90 → 0.925’)	36	4	37	3

Note: For Experiment 1a and Experiment 1b identical displays were used.

The two displays of the 40 cards were each ordered in 4 rows × 10 columns. The two displays were said to be sorted independently of each other, so that the cards provide neither positive nor negative evidence for the hypothesis “raven → black”. The used card order is shown in Table 17. (No seldom card in the first display visually corresponded to another seldom card in the second display.)

Table 17
Position of the Seldom Cards in the Displays of the P / Non-P and of the Q and Non-Q Card Sides

		P and *non-p* display	*Q* and *non-q* display
Low probability conditions (‘0.10 → 0.20’)	Seldom cards	4 *p* cards	8 *q* cards
	Coordinates of these cards	1,7; 2,4; 3,9; 4,3	1,4; 1,8; 1,9; 2,1; 3,3; 3,10; 4,2; 4,5
High probability conditions (‘0.90 → 0.925’)	Seldom cards	4 *non-p* cards	3 *non-q* cards
	Coordinates of these cards	1,7; 2,4; 3,9; 4,3	1,8; 2,3; 4,5

Note: The positions of cards in the 4 rows × 10 columns matrix are shown. For instance, ‘1,7’ represents a position in the first row and in the seventh column of the matrix.

In Experiment 1a, which is concerned with *p* versus *non-p* selections, the first *p* and *non-p* display is the one, displayed later and from which one was allowed to make selection:

"The keeper lays out the 40 cards for you. On the 40 cards placed in front of you, only the side is visible which lists the species of the bird.

[Display of the *p* and *non-p* card sides.]

You also know something about the reversed sides of the cards, showing the colour of the feathering. Earlier on, the keeper displayed the same cards showing the reversed side of the cards, concerned with the colour of the birds.

[Display of the *q* and *non-q* card sides.]

Between the two displays of the cards, the keeper has collected up the cards. In this process the cards get shuffled completely. Hence, the order of the two displays need not correspond to each other. Unfortunately, the keeper is now not prepared to turn over many more cards. He only allows you to turn over *one* single card separately. – Please encircle the card that you would turn over in order to test (with only this sample) the truth or falsity of the controversial sentence. You are only allowed to turn over *one of the cards which are currently displayed* (above: '🐦' or '~~R~~')!"

The test of *q* versus *non-p* selections (Experiment 1b) necessitated differences in the context story, since the cards, from which one was allowed to make selections, needed to be in the temporarily final display. At the same time, the graphical presentation order was kept constant.

"The keeper lays out the same 40 cards twice for you. In the first display of the cards he puts the card with the name of the species facing upwards.

[Display of the *p* and *non-p* card sides.]

You also aim to see the backside of these cards but the keeper – before you could stop him – grabs all 40 cards, and lays them out anew, now with the sides with the feather colour facing upwards. The cards get totally mixed. In this second display, the card sides can be distinguished by the colour of the feathering of each bird. Because the cards have been mixed, their order no longer corresponds to the first display.

[Display of the *q* and *non-q* card sides.]

Unfortunately, the keeper is now not prepared to turn over many more cards. He only allows you to turn over *one* single card separately. – Please encircle the card that you would turn over in order to test (with this sample) the truth or falsity of the controversial sentence. You are only allowed to turn over *one of the cards which are currently displayed* (above: '●' or '○')!"

Results of Experiment 1

Main results. The number and percentage of found card selections in Experiment 1a and 1b are reported in Table 18. The results are presented both for the antecedent choice conditions (Exp. 1a: *p* versus *non-p*) and the consequent choice conditions (Exp. 1b: *q* versus *non-q*), contrasting the selections in the low and high probability conditions.

Table 18

Percentage and Number of Card Selections in the Low and High Probability Conditions of the P versus Non-P, and the Q versus Non-Q Forced Choice Conditions

Exp. 1a :	*P* versus *non-p*	
	Low probability	High probability
P	100 % 15	43 % 6
Non-p	0 % 0	57 % 8
n	15	14

Exp. 1b:	*Q* versus *non-q*	
	Low probability	High probability
Q	60 % 9	13 % 2
Non-q	40 % 6	87 % 13
n	15	15

Note. Selections which are predicted to increase are darkened.

Descriptively the results support the predicted increase of *non-p* and *non-q* selections in the high probability conditions. The changed proportions of card selections are also illustrated by bar graphs in Figure 8.

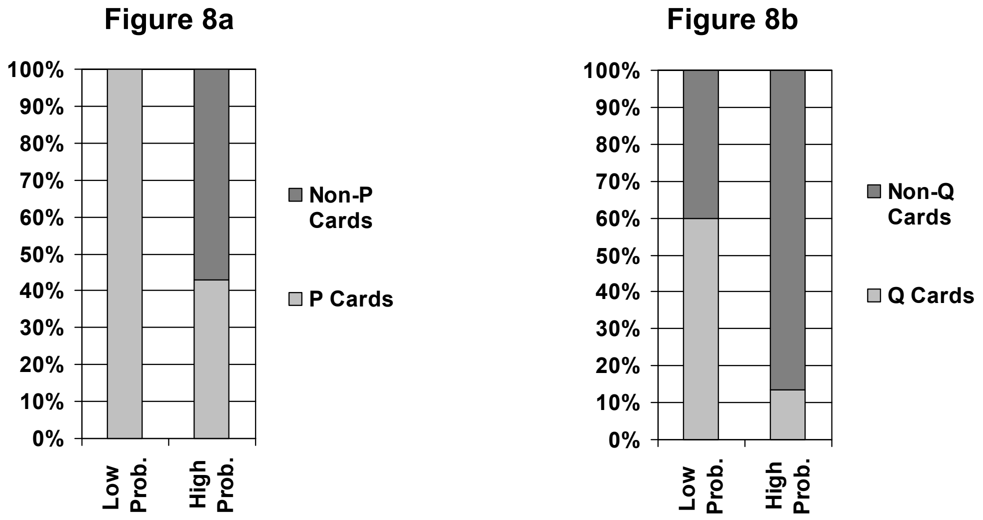


Figure 8a, b. Bar graphs of the percentage of (a) *p* or *non-p* selections and (b) of *q* or *non-q* selections in each corresponding low and high probability conditions.

The predicted changes in the proportions of card selections are statistically highly significant. The increase of *non-p* card selections relative to the *p* card selections in the high probability condition was reliable (exact Fisher test: df = 1, n = 29, p < 0.01; r_{φ} =.64). Also the increase of *non-q* versus *q* selections was highly significant (Pearson test: n = 30, $\chi^2_{(1)}$ = 7.03, p < 0.01; r_{φ} = .48).

Additional questions. In the questionnaire, the majority of participants stated they had no knowledge of probability theory (73 %) and no knowledge of formal logic (82 %). Over all conditions the assumed knowledge of probability theory and actual answers which confirmed to the Bayesian model were not positively correlated (r_{φ} = -.14). However, there was also no correlation between the knowledge of formal logic and answers which accorded to the logical-falsificationist norm (r_{φ} = -.03). In any case, when asked for their own strategy, only few participants mentioned logic (5 %) or probability theory (6 %) and most participants answered that they proceeded according to their own reasoning or intuition (75 %).

Discussion Experiment 1 –
Support for a Bayesian Solution of the Raven Paradox

The results of Experiment 1 are discussed here mainly with regard to the dispute concerning the raven paradox and only briefly with regard to the WST debate.

The WST debate will be considered in detail in Experiment 2 and alternative theories of the WST will be examined at the end of this chapter in the general discussion (Section 5.7).

Experiments 1a and 1b support the standard Bayesian resolution of the raven paradox (Section 3.2). In the WST context it has been shown here for the first time that for testing the hypothesis ‘if a bird is a raven then it is black’ changed probabilities, P(raven) and P(black), can increase the portion of non-ravens selections. Moreover, also the predicted *q* versus *non-q* effects were found.

It is consistent with the advocated knowledge-based account that these positive results were achieved in a context in which the marginal probabilities were fixed by the task instruction. In the only earlier WST also using ravens material (v. Sydow, 2002, Exp. 1), these preconditions – as has been outlined before – were not firmly established. The found contrast between the positive results obtained here on the one hand, and the negative result with the previous raven WST (Sydow, 2002; cf. pp. 94 f) or with the generally negative results for the model with fixed marginals (pp. 83 f) on the other hand, suggest that the improvement may be due to use of a task in which all preconditions for the model were actively induced. This at least indirectly supports the advocated knowledge-based approach.

In contrast, the results of Experiment 1 cannot be explained by a falsificationist psychological account of the raven paradox (Popper, 1934/2002, 1972, Watkins 1957, 1958). Strict falsificationism demands *p* and *non-q* selections, invariably under all conditions.

However, even a 'probabilistic falsificationism', which allows for frequency effects regarding the falsifying *p* versus *non-q* selections (cf. Humberstone, 1994; Kirby, 1994b), cannot explain the found *p* versus *non-p* and *q* versus *non-q* frequency effects.

Likewise, naïve tabula rasa empirism (cf. pp. 32 f.), without Bayesian modification, cannot account for the findings. An empiricism that denounces any use of prior knowledge and only advocates enumerative induction (cf. Nicod's criterion, cf. pp. 40 f.) would invariably predict *p* and *q* selections. Likewise, an account which advocates that negative instances are not natural kinds and, hence, assumes that negative instances are generally irrelevant for testing a positively formulated rule (see Quine, 1969; Oaksford and Chater, 1994, 1998b), cannot explain the increased number of *non-p* and *non-q* selections in the high probability condition.

With regard to the WST debate, the results provide us with one of the first clear corroborations of the Bayesian model with fixed marginals (von Sydow, 2002; cf. p. 78). Most previous results were negative or ambivalent (pp. 70 f., 83 f.). This contrast and also the found contrast to the older WST with raven material by v. Sydow (2002), where the marginal probabilites were not fixed, cannot be explained by Oaksford and Chater's universal approach (1994, 1998, 2003), but only by a knowledge-based account.

5.6 Experiment 2a, b – Letter-Number MST

The Letter-Number WST

The original WST of Wason (1966), with the hypothesis "if a card has an 'A' on one side, then it has a '2' on the other side", was the paradigmatic case of the research on the WST. The first anomaly for a falsificationist account (see pp. 8 f.) became apparent exactly for this kind of letter-number hypotheses. Even the first experimements of Wason and collegues showed a predominance of confirmative ‘A’ (*p*) and ‘2’ (*q*) selections (Johnson-Laird & Wason, 1970).

It would be particularly convincing for the alternative Bayesian approach to obtain *p* versus *non-p* and *q* versus *non-q* frequency effects in the Sydow model with this original letter-number material (see also the results of v. Sydow, 2002; pp. 94 f.).

But as the results with other material were rather negative (pp. 70 f.) or ambivalent (pp. 83 f.), the results specifically with letter-number material have not been much better (see particularly Oberauer, Wilhelm, and Diaz, 1999; cf. Feeney & Handley, 2000, Oaksford, 2002, Oaksford & Wakefield, 2003, Oberauer, Weidenfeld & Hörnig, 2004):

Oberauer, Wilhelm, and Diaz (1999) in two experiments used letter-number material when systematically testing frequency effects, but their results were clearly negative (Exp. 2, 3).

Oaksford and Wakefield (2003) suggest that a reason why Oberauer et al. (1999) did not find the predicted probability effects was that they did not use a natural sampling process, in which each data point is learned sequentially, one at a time (cf. Gigerenzer & Hoffrage, 1995). This might also explain the rather negative results of Oaksford, Chater, and Grainger (1999; see p. 86). However, Oaksford and Wakefield (2003) conceded that their earlier sequential test using a RAST (Oaksford, Chater, Grainger, & Larkin, 1997) could not count as a proper test of their theory of the WST, since cards were successively turned over (cf. Oaksford and Chater, 1998b; Klauer, 1999). In a new experiment, Oaksford and Wakefield used the same number-letter material as Oberauer et al. (1999) but employed a sequential natural sampling process without allowing participants to turn over any cards. However, participants could select not only one cards of a logical category (like in a WST) but many. Oaksford and Wakefield at least seemingly obtained *p* versus *non-p* and *q* versus *non-q* frequency effects. They concluded “that when natural probability manipulations are used people’s data selection behaviour is rational“ (p. 143).

However, Oberauer, Weidenfeld and Hörnig (2004, pp. 522, 527) objected, in my view correctly, that the manipulation used by Oaksford and Wakefield confounded the frequency manipulation with the number of opportunities to turn over cards from a particular logical category. Oaksford and Wakefield divided the number of selections by the numbers of opportunities to select a card. But based on such an analysis, *non-p* selections may appear to predominate the selections of the sequence "*p*, *p, p, non-p, p, p, p, non-p*" even if there were two *p* selections and only one *non-p* selection (cf. pp. 89 f.). Actually, if one reanalyses Oaksford and Wakefield's data without dividing the obtained selections by the selection opportunities, there is a decrease and not an increase in *non-p* or *non-q* selections in the high probability condition. Hence, I think one has to concede that their results do not provide corroboration of their Bayesian model of the WST.

Oberauer, Weidenfeld and Hörnig (2004) used a sequential learning phase, as required by Oaksford and Wakefield (2003), but tested frequency effects without confounding them with selection opportunities in the test phase. Their results were negative, even with significant results in the reversed direction. Hence, Oaksford and Wakefield (2003) seem to have been wrong; sequential sampling is clearly not sufficient to obtain positive Bayesian results.

The knowledge-based Bayesian account, advocated here, regards sequential sampling as neither necessary nor sufficient for obtaining positive results, although a clear frequency format may well be advantageous. Here it is argued that the preconditions of the Sydow model cannot simply be assumed as general features of testing conditional hypotheses, but that they need to be introduced actively by using for instance MSTs, also ruling out some other possible misunderstandings (cf. pp. 83 f.). Hence, from the knowledge-based perspective it is predicted that *p* versus *non-p* as well as *q* versus *non-q* frequency effects should be obtained if a task is used that actively introduces the preconditions of the tested model in a salient way, even without using sequential sampling.

Method Experiment 2a, b

Design and Participants

Experiments 2a and 2b were almost identical; they only differed in their dependent variable. Experiment 2a tested *p* versus *non-p* effects, Experiment 2b *q* versus *non-q* effects. Both Experiments were presented successively, varying the task order. Experiment 2 had three between-subject factors, one concerning the order of the experiments (Exp. 2a or Exp 2b first), one concerning the probabilities (low versus high probability condition) and one controlling for the tested rule ('vowel → even number' versus 'consonant → odd number' rule). The last factor served as a control to check for additional frequency effects caused by probability assumptions about atomic propositions mentioned in the rule (vowels and consonants might have different subjective probabilities regardless of the salient frequency manipulation).

Ninety-six students from the University of Göttingen participated in the experiment (72 % female, 28 % male; mean age 23 years). Most of the students studied psychology (53 %); the second largest group studied biology (10 %). The participants were randomly assigned to the resulting eight conditions.

In Experiment 2a seven participants and in Experiment 2b eight participants were excluded due to formal errors (they selected more cards or other cards than those they were formally allowed to select). Hence, 89 participants were analysed for Experiment 2a, and 88 participants for Experiment 2b.

Procedure and Materials

The order of Experiment 2a and 2b was varied as a separate factor but both tasks were each formulated almost identically independent of their serial order. If a task was presented in the first position, the instruction commenced "Below you see some cards". Tasks in the second position commenced "This is a new task, which is completely independent of the previous task. Below you see a *new* set of cards."

The tasks in Experiment 2a and 2b were formulated similarly. Both continued: "On one side of each card is a letter (consonant or vowel), on the other side is a number (even or odd). In this task you should check whether the following additional assertion is true or false". Then the rule was stated (the formulations of the rules were all set in bold print):

- In the *vowel condition*, the following rule was used: "If there is a vowel on the letter side of the card, then there is always an even number on the card side. *In brief: If vowel then even number*."
- In the *consonant condition*, the rule was formulated as follows: "If there is a consonant on the letter side of the card, then there is always an odd33 number on the card side. *In brief: If consonant then odd number*."

The instruction in *Experiment 2b* (*q* versus *non-q* selections) continued: "You should check, whether the cards correspond to this rule or not. [Break] The same 20 cards have been displayed twice; in the first display all letter sides were put upwards".

[33] Translation note: The German word for 'odd', 'ungerade', literary means 'uneven'.

Twenty cards were shown with letters facing upwards. “In the second display of the same cards now all number sides are shown.”

Twenty cards were shown with numbers facing upwards. In the two displays ‘A’ cards were used as vowels, ‘K’ cards as consonants, ‘2’ cards as even numbers, and, finally, ‘7’ cards as odd numbers. In the low and high probability condition two different card probabilities were used:

- In the low probability condition (‘.10 → .20’) there were 2 *p* card sides (10 %), 18 *non-p* card sides (90 %), 4 *q* card sides (20 %), and 16 *non-q* card sides (80 %).
- In the high probability condition (‘.80 → .90’) there were 16 *p* card sides (80 %), 4 *non-p* card sides (20 %), 18 *q* card sides (90 %), and 2 *non-q* card sides (10 %).

The order of the cards was analogous to Experiment 1 (see Table 16, p. 101). The instruction continued: “Between the two displays the cards were completely mixed; the card order does not correspond in the two displays. [Break] Please indicate by ticking a card, which single card you would turn over, to test the *truth or falsity* of the proposition in this sample. You are only allowed to turn over *one* currently displayed *gray* card (either a ‘2’ or a ‘7’).” In the described instruction of Experiment 2b the *q* and *non-q* card sides were gray, in order to make clear which cards can be chosen.

The instruction of *Experiment 2a* tasks was almost identical to the described Experiment 2b, only the first display with the letters facing upwards (*p* and *non-p* cards) was now described as representing the present display and the second display of the number sides (*q* and *non-q* cards) was described as being a previous display. Correspondingly, here the first display was gray. The final instruction again asked to select a presently displayed gray card, which here referred either to one of the ‘A’ or ‘K’ cards.

Results Experiment 2a and Experiment 2b

In all four conditions of each experiment the order of the task neither significantly influenced the *q* versus *non-q* selections (exact Fisher tests: $p = .67$, $p = 1.0$, $p = 1.0$, $p = 1.0$) nor the *p* versus *non-p* selections (exact Fisher tests: $p = .45$, $p = .66$, $p = .32$, $p = .66$). Hence, the results were collapsed across this factor.

Table 19 presents the *p* versus *non-p* selections found in Experiment 2a. Descriptively all differences went in the predicted direction. Table 21 presents the *q* versus *non-q* selections found in Experiment 2b.

Table 19
Experiment 2a: Percentage and Number of Card Selections Concerning P Versus Non-P Selections in the Low and High Probability Conditions

	Consonant → odd		Vowel → even		Overall	
	Low	High	Low	High	Low	High
P	95 %	48 %	78 %	67 %	87 %	57 %
	21	11	18	14	39	25
Non-P	5 %	52 %	22 %	33 %	13 %	43 %
	1	12	5	7	6	19
n	22	23	23	21	45	44

Note. Selections which are predicted to increase are darkened.

Table 20
Experiment 2b: Percentage and Number of Card Selections Concerning Q Versus Non-Q Selections in the Low and High Probability Conditions

	Consonant → odd		Vowel → even		Overall	
	Low	High	Low	High	Low	High
Q	64 %	43 %	76 %	41 %	70 %	42 %
	14	10	16	9	30	19
Non-Q	36 %	67 %	24 %	59 %	30 %	58 %
	8	13	5	13	13	26
n	22	23	21	22	43	45

Note. Selections which are predicted to increase are darkened.

There was no difference whether participants tested the 'consonant → odd' or the 'vowel → even' hypothesis, both in the low and the high probability conditions of Experiment 2a (Pearson $\chi^2_{(1)} = 1.58$, $p = .20$; exact Fisher test, $p = .19$). Likewise this had no effect in Experiment 2b (Pearson $\chi^2_{(1)} = .03$, $p = .86$; $\chi^2_{(1)} = .80$, $p = .37$). The explicit and salient frequency information in the MST seems to have expunged any remaining influence of previous knowledge about the frequency of vowels and consonants. Hence, the data can also be collapsed across this dimension.

The main test concerning the low versus the high frequency condition became significant for the predicted *p* versus *non-p* effect (Exp. 2a: Pearson $\chi^2_{(1)} = 9.81$, $p < 0.01$, $r_\varphi = .33$) as well as for the predicted *q* versus *non-q* effect (Exp. 2b: $\chi^2_{(1)} = 6.76$, $p < 0.01$, $r_\varphi = .28$).

According to the final questionnaire only 10 % of the participants claimed that they solved the task using formal logic, 71 % used their 'intuition or own reasoning'.

Discussion Experiment 2

The results of Experiment 2a and 2b both corroborated the predictions of the Bayesian model with fixed marginals (von Sydow, 2002; Oaksford & Chater, 2003). Although the effect sizes were lower than in Experiment 1 (see pp. 103 f.), when using 'realistic' ravens material, here too highly significant *p* versus *non-p* and *q* versus *non-q* effects were found.

For the first time, *p* versus *non-p* and *q* versus *non-q* frequency effects are were shown for a non-sequential selection task with the original letter-number material (Wason, 1966). We have seen before that Oaksford and Wakefield's results (2003) cannot count as proper confirmation of the postulated effects (cf. Oberauer, Weidenfeld and Hörnig, 2004, 522, 527). Oaksford and Wakefield's (2003) explanation, that earlier negative results (e.g., Oberauer, Wilhelm, Diaz, 1999; Oaksford, Chater, Grainger, 1999, Exp. 2-4, see p. 86) are presumably due to not using a natural sequential sampling process, appears to be incorrect. Firstly, Oberauer et al. (2004) have shown that supporting results of Oaksford and Wakefield (2003) may be due to confounding selection opportunities and selections, and if this confounding factor is removed, the results are also negative for a sequential task (Oberauer, Weidenfeld and Hörnig, 2004). Secondly, the current results show that the predicted results can be obtained if we use an MST in which all preconditions are clearly fulfilled. Reasons for the negative results in many previous tasks may have been that probabilities or constancy assumptions were not actively introduced, or that these preconditions were obscured by a complicated procedure. Of course, this needs not to be the only reason for the results. For instance, the interesting reversed effects of Oberauer, Weidenfeld and Hörnig (2004) may be due to a selection tendency learned in the first phase. In any case, here positive results have been achieved by using a non-sequential task in which the probabilities and constancy assumption of the model with fixed marginals were clearly given.

It will be shown in the General Discussion of Chapter 5 that other non-Bayesian accounts of the WST cannot explain the results.

5.7 General Discussion and Summary of Chapter 5

In this section, it is argued that the results of Chapter 5 favour the advocated Bayesian approach over all other theories of the Wason Selection Task. The implications for the debate on the raven paradox have ben considered in Section 5.5.

Support for the Bayesian Approach. In Chapter 5 it has been advocated that the Bayesian model with fixed marginals, which was first fully elaborated by von Sydow (2002, cf. Hattori, 2002, Oaksford and Chater, 2003), is only a valid model if the assumed constancy conditions are actually given in the situation. Since the model is not advocated universally here, but only in this knowledge-based sense, the normative and descriptive predictions of the model do not follow if its preconditions are not empirically and subjectively given. It has been shown that previous results were negative for a universal Bayesian account (cf. pp. 87 f.) and even all results which seem to show *p* versus *non-p* effects were shown to be incoherent or criticisable (pp. 89 f.). In order to exclude the problems of previous experiments and in order to induce a model with fixed marginals, a many card selection task (MST) has been proposed. If the knowledge-based account is correct, such a task should make it possible to obtain clear-cut frequency effects as predicted according to the Sydow model (cf. v. Sydow, 2002).

In this chapter, the results of two new experiments have been reported; both corroborate the predictions of the Sydow model (von Sydow, 2002; Oaksford & Chater, 2003). Experiment 1a and Experiment 1b for the first time showed the predicted *p* versus *non-p* and *q* versus *non-q* frequency effects using ravens material known from the raven paradox debate (see pp. 39 f.). This also corroborates the psychological adequacy of the standard Bayesian resolution of the paradox (pp. 104 f.). Moreover, Experiment 2a and 2b showed the predicted frequency effects for the first time for the original letter-number material of Wason (1966), without using the problematic sequential design used by Oaksford and Wakefield (2003; cf. p. 89; see also v. Sydow, 2002; cf. pp. 94 f.).

The found *non-p* effects are at odds with the older Bayesian model of Oaksford and Chater (1994, 1998b) and they seem to be problematic for the accounts which postulate that the truth or falsity of a conditional ‘if *p* then *q’* generally reflects only the likelihood of $P(q \mid p)$ (Evans, Handley & Over, 2003; Over & Evans, 2003; Over, 2004; Evans, Over, Handley 2005; but confer earlier, e.g., Evans & Over, 1996, 360;

Green, Over and Pyne, 1997, 219; Green and Over, 2000; see also Evans, 1972; cf. pp. 83 f.).

The results refute Oaksford and Wakefield's suggestion (2003) that the predominantly negative results in former selection tasks (e.g., Oberauer et al. 1999, Oaksford et al., 1999) are to be explained by not having used a 'natural' successive test. As argued before their own successive tasks (Oaksford et al. 1997, Oaksford & Wakefield, 2003) cannot count as a proper test of their theory (cf. the Discussion of Experiment 2). Moreover, Oberauer, Weidenfeld and Hörnig (2004, 522, 527) showed that a successive task is not sufficient to achieve the selection effects predicted by a model with fixed marginals. In contrast, the two non-successive selection tasks used and reported in this chapter clearly support the predictions of the Sydow model. Hence, in order to achieve positive results successive tests are neither sufficient, as shown by Oberauer et al. (2004), nor necessary, as shown here. In our experiments, the task used was constructed upon a knowledge-based account by inducing the assumed preconditions of the model with fixed marginals explicitly and by ruling out a number of problems within previous experiments.

Other Theories of the WST. The results of the two experiments reported in this chapter are inconsistent with mental logic theory of the WST and mental model theory. Although domain-specific theories have been mainly concerned with what I call 'prescriptive conditionals', the results are also inconsistent with some claims of these theories concerning descriptive conditionals. The General Discussion of Part II provides a more detailed discussion, also including relevance theory and matching bias heuristics.

Mental logic theory (ML theory, e. g., Braine, 1978; Rips, 1994; O'Brien, 1995; cf. pp. 10 f, 172 f.) has not predicted frequency effects and cannot explain the pattern found here.

The difference in difficulty to reason according to a modus ponens and a modus tollens can be explained using ML theory by postulating an incomplete set of rules not directly including the modus tollens. Also the difficulties achieving correct selections in standard WST (the first anomaly of the WST, cf. p. 8) and the difficulty achieving better results in 'therapy' experiments can be explained by postulating a complex reasoning process in order to derive the 'correct' falsificationist solution (without the direct use of a modus ponens). However, frequency effects should not affect the existence or non-existence of the postulated rules of a mental logic. Therefore, ML theory does not provide any positive explanation of the results of Experiment 1 and 2.

The *mental model theory of the WST* (MM theory, Johnson-Laird & Byrne, 1991, 2002) has, on the one hand, also generally maintained a falsificationist norm of testing conditionals and, on the other, has explained the deviations found in the WST by a set of incomplete representations (pp. 11 f. for details). "In short, the model theory predicts that people will select the card falsifying the consequent whenever the models are fleshed out with explicit representations of that card." (Johnson-Laird & Byrne, 1991, p. 80) MM theory has not advocated any frequency based account for the WST and hence all frequency effects obtained here seem to be problematic for MM theory.[34]

Nonetheless, one may try to defend MM theory, since the probability of finding a counterexample in a situation has been postulated to affect the fleshing out of mental models (cf. Green & Larking, 1995; Love & Kessler, 1995; Green 1998).

However, this integration of frequency effects into MM theory is problematic because there have been cases in which the obvious availability of counterexamples did not lead to 'correct' *p* and *non-q* selections. For example, "If I eat haddock then I drink gin" did not lead to facilitation effects (Manktelow & Evans, 1979), although counterexamples should be available. (In contrast, a Bayesian can account for the *p* and *q* selections by the plausible assumption that P(eat haddock) < 0.5 and P(drink gin) < 0.5.)

Moreover, even if the probability of thinking of a counterexample, P(counterexample), always led to 'a fleshing out' of the mental model, causing *p* and *non-q* selections, this would not imply any account of how conditions which vary exclusively in regard of $P(p)$ and $P(q)$ should be linked to P(counterexample) and the resulting fleshing out of a mental model. Hence, mental model theory cannot explain the results.

Additionally, even if we also tried to fill this gap and provided an account to link probabilities and the fleshing out of mental models, this, in my view, would not yield positive results for MM theory. It is not clear whether a high or a low frequency

[34] This is the case, although Johnson-Laird, Legrenzi, Girotto, Legrenzi & Caverni (1999) proposed a mental model theory of probabilities. I agree with Oaksford, Chater, & Larkin (2000, 898) that most aspects of such a theory do not intrinsically rely on the notion of a mental model. Moreover, Johnson-Laird et al.'s (1999) probabilistic extension of MM theory has not been applied to the WST. Furthermore, this theory, in my view, cannot account for *p* versus *non-p* effects anyway.

condition should enhance the probabilities of envisaging counterexamples. The most plausible and direct proposal would be to assume that the higher the number of visible and salient *non-p* and *non-q* cards in an MST, the higher should be the probability that the incomplete model (*p* & *q*) is fleshed out. But this straight proposal would imply that a low probability condition, with many *non-p* and *non-q* cards, should lead to more fleshed-out models than to a high probability condition, which is contrary to our results.

Furthermore, even if we assumed a more complex relationship between frequency information and the fleshing out of a mental model, this, in my view, would not explain the results. *P*(counterexample) may rationally be estimated using the independence model, M_I, also used in the Bayesian model (in the dependence model there are no counterexamples). Hence, *p* and *non-p* cases would have to be combined randomly with *q* and *non-q* cases. However, in this case with the parameters used in Experiment 1 and 2 an equal number of counterexamples, $P(p \;\&\; non\text{-}q \mid M_I)$, results in both the high and the low frequency conditions, and this would not allow for any frequency effects from the viewpoint of MM theory. If $P(p \;\&\; non\text{-}q \mid M_I)$ would alternatively be relativised by the number of those positive instances, which are represented in a mental model of a conditional, $P(p \;\&\; q)$, this would again yield predictions of an increase of *non-q* selections in the low frequency condition, which would be reversed to our findings (cf. also the General Discussion of Part II).

In any case, such complex calculation would be problematic from a MM viewpoint, since such calculations presupposes a much more complex representation to allow for the construction of a relatively simple model only made up of four instances. Any such account, if viable at all, would clearly go far beyond MM theory.

Finally, MM theory cannot explain the simultaneous increase of non-p selections which accures together with the increase of non-q selections (cf. General Discussion of Part II).

Therefore, current MM theory clearly cannot account for the results of Experiment 1 and Experiment 2.

The two main *domain-specific theories*, social contract theory and pragmatic reasoning schema theory, mainly made proposals concerned with social or deontic WSTs, and only took a minor interest in descriptive WSTs. Nonetheless, the results of this chapter also show that some of their claims and assumptions made about descriptive WSTs need to be given up.

As *Social contract theory* has been formulated, it has been argued that no thematic rule that was not a social contract had ever produced robust content effects (Cosmides, 1989, pp. 200; Cosmides & Tooby, 1992, p. 183; cf. later, Fiddick, Cosmides, & Tooby, 2000; Fiddick, 2004). In the current Experiment 1 and 2 clearly systematic ‘content effects’ have been achieved for rules which are not social contracts.

Moreover, Cosmides and colleagues have used descriptive WSTs as a yardstick against which to measure their findings with social contracts. They assumed and found a standard pattern of diffuse selections, generally predominated by *p* and *q* selections, in descriptive tasks. The findings of Experiment 1 and 2 show that this pattern is not an adequate general yardstick, since the selections in descriptive WSTs are shown to be dependent on *P*(*p*) and *P*(*q*).

Although *pragmatic reasoning schema theory* (Cheng & Holyoak, 1985; Holyoak & Cheng, 1995a, b, cf. pp. 13 f., 176 f., 270 f.) was almost completely concerned with the permission and the obligation schema, in principle it also allowed for schemata in the field of descriptive WSTs (see particularly Cheng & Holyoak, 1989, 306). However, Cheng and Holyoak have not developed a detailed positive theory on descriptive WSTs. In their writings on the WST they only briefly mentioned two schemas of causality and covariance which are claimed to lead generally to *p* and *q* selection patterns. In my view, this was nothing but a rediscription of the data which were at hand at that time. But this description is inconsistent with the frequency effects observed in Experiment 1 and 2.

Moreover, Cheng and Holyoak (1985, 396) argued that an "arbitrary rule, being unrelated to typical life experiences, will not reliably evoke any reasoning schema". Although the predicted frequency effects had a lower effect size for the abstract letter-number hypothesis than for the realistic ravens hypothesis, significant effects were obtained for both experiments (but see, pp. 176 f.).

Finally, also *other non-Bayesian mechanisms or approaches* like matching bias heuristics or relevance theory cannot explain the findings of Experiment 1 and 2 (cf. General Discussion of Part II).

6 Flexible Bayesian Models of the WST

Oaksford and Chater (1994, 1998a) formulated their original model as a universal Bayesian model for testing descriptive conditionals, not differentiating between different situations. Likewise, but in contrast to von Sydow (2002, 2004a), also Oaksford and Chater's revised model (2003) has been formulated as a universal Bayesian model.[35]

Hence, Laming's (1996) fundamental objection that no proper reasons have been provided to justify any general constancy assumption is not ruled out. Taking this

[35] Also the metaanalysis of Oaksford and Chater (2003) is based on this assumption, since otherwise one would have needed to differentiate between different models. Other problems of their metaanalysis and their reinterpretation of former experiments have been discussed previously, see pp. 83 f.

criticism seriously, von Sydow (2002) did not abandon the Bayesian approach but advocated a knowledge based account and for the first time investigated the model with fixed marginals while explicitly trying to introduce all preconditions of that model empirically (cf. also Chapter 5).

The advocated knowledge-based approach combines the refined modelling of Oaksford and Chater (1994, 1998) with objections that their research programme is "in danger of reducing the human agent to a technical cipher, the output of which is predicted only under a host of unchecked assumptions" (Green & Over, 1998, 192; cf. Evans & Over, 1996). Also Over and Jessop (1998) and Green and Over (2000), although only concerned with *probabilistic* conditionals, have theoretically pointed out the problem of different possible constancy assumptions (cf. pp. 70 f.). However, they have not formulated this in the elaborated modelling framework of Oaksford and Chater (1998a) and these interesting comments have not been explicitly pursued any further. Recent papers in the modelling tradition of Oaksford and Chater have ignored this problem (Oberauer et al. 1999; Hattori, 2002, despite pp. 126-128; Oaksford & Wakefield, 2003, Oberauer, 2004). Even Over and Evans (2003), Evans, Handley, and Over (2003), and Over (2004) have more recently appeared to come to a more universal position once again, claiming that the truth or falsity of a conditional is psychologically, generally only determined by the conditional probability $P(q \mid p)$ and not by the *non-p* cases (cf. Edgington, 2003; lately Evans, Over, Handley, 2005).[36]

Unlike any *general* Bayesian model of the WST, here a flexible modelling is advocated, while using the advanced modelling account of Oaksford and Chater. According to this knowledge-based account that model is normatively valid whose preconditions are fulfilled by the constraints given in a situation (cf. v. Sydow, 2002, 2004). Instead of asking whether the original model of Oaksford and Chater (1994) or the revised model with fixed marginals (von Sydow, 2002; Hattori, 2002, Oaksford & Chater, 2003) is the generally correct, it is predicted here that under different context conditions the one or the other model is the normatively correct and, corresponding to this normative claim, also the descriptively used.

[36] Paradoxically, this position rather refers to the original position of Oaksford and Chater (1994, 1998a), which they themselves have given up (Oaksford and Chater, 2003).

Table 21
Three Models Different Structural Models of a Deterministic If-then Hypothesis (see von Sydow, 2004a).

Table 21a
Simplified Model of Oaksford and Chater (1994) without q-modification (cf. p.60); P(p) and P(q | 1 - p) are set to be the same in both sub-models M_D and M_I.

M_D	*P*	*Non-q*	Marg.
P	p	0	p
Non-p	$(1 - p)$ q	$(1 - p)$ $(1 - q)$	$1 - p$
Marg.	$q + p$ $- p\,q$	$(1 - p)$ $(1 - q)$	1

M_I	*Q*	*Non-q*	Marg.
P	$p\,q$	p $(1 - q)$	p
Non-p	$(1 - p)$ q	$(1 - p)$ $(1 - q)$	$1 - p$
Marg.	q	$1 - q$	1

Table 21b
Model of von Sydow (2002; cf. p. 80), Oaksford and Wakefield (2003); P(p) and P(q) are set to be the same in both sub-models.

M_D	*Q*	*Non-q*	Marg.
P	p	0	p
Non-p	$q - p$	$1 - q$	$1 - p$
Marg.	q	$(1 - q)$	1

M_I	*Q*	*Non-q*	Marg.
P	$p\,q$	p $(1 - q)$	p
Non-p	$(1 - p)$ q	$(1 - p)$ $(1 - q)$	$1 - p$
Marg.	q	$1 - q$	1

Table 21c
Model of Laming (1996, cf. p. 73); P(q) and P(p | q) are set to be the same in both sub-models.

M_D	*Q*	*Non-q*	Marg.
P	$p\,q$	0	$p\,q$
Non-p	$(1 - p)\,q$	$1 - q$	$1 – p\,q$
Marg.	q	$1 - q$	1

M_I	*Q*	*Non-q*	Marg.
P	$p\,q$	$p\,(1 - q)$	p
Non-p	$(1 - p)\,q$	$(1 - p)$ $(1 - q)$	$1 - p$
Marg.	q	$1 - q$	1

In all sub-tables the following notation is used in the cells of M_D and M_I: $p := P(p)$, $q := P(q)$.

In contrast to *general* Bayesian approaches of testing conditionals (Oaksford & Chater, 1994, 1998a, 2003; cf. Over and Evans, 2003; Evans, Handley and Over, 2003, and Over, 2004) it is predicted here (cf. v. Sydow, 2004a, cf. theoretically von Sydow, 2002) that humans are not only sensitive to quantitative preconditions, but also to qualitative preconditions of different structural models.

Here the advocated knowledge-based Bayesian account of the WST is for the first time tested more directly, varying *different* structural preconditions in testing conditionals in a WST *within* an experiment (cf. v. Sydow, 2004). The original model of Oaksford and Chater (1994, 1998), the model of von Sydow (2002) and the model of Laming (1996) were modelled and tested along the same lines.

In Table 21, the three structural basic models are presented. Instead of the original model of Oaksford and Chater (1994, 1998; see Table 11, p. 60) here a slightly modified version is used, without their problematic modification of $P(q)$ (cf. p. 61). In this Oaksford model the constancy assumptions, $P(p \mid M_D) = P(p \mid M_I)$ and $P((q \mid 1 - p) \mid M_D) = P((q \mid 1 - p) \mid M_I))$, are assumed. In the Sydow model $P(p \mid M_D) = P(p \mid M_I)$ and $P(q \mid M_D) = P(q \mid M_I)$. In the Laming $P(q \mid M_D) = P(q \mid M_I)$ and $P((p \mid q) \mid M_D) = P((p \mid q) \mid M_I)$ are set to be constant. This last model was included although Laming (1996) only proposed this model in order to show that such a model would be absurd.

The further steps to model the WST (Bayes' Theorem, Wiener-Shannon-Information and the resulting *expected information gain* measure) were completely taken from Oaksford and Chater (1998a; see pp. 59 f., cf. pp. 71 f.).

I have not used the additional extensions, which were proposed by Hattori (2002) and Oaksford and Chater (2003) to achieve a better fit when modelling the data. Although these modifications may be psychologically plausible, they are not part of the normative model.

The modelling results are presented in Figure 9, showing the behaviour of the Sydow model (A), the Oaksford model without q modification (B), and the Laming model (C).

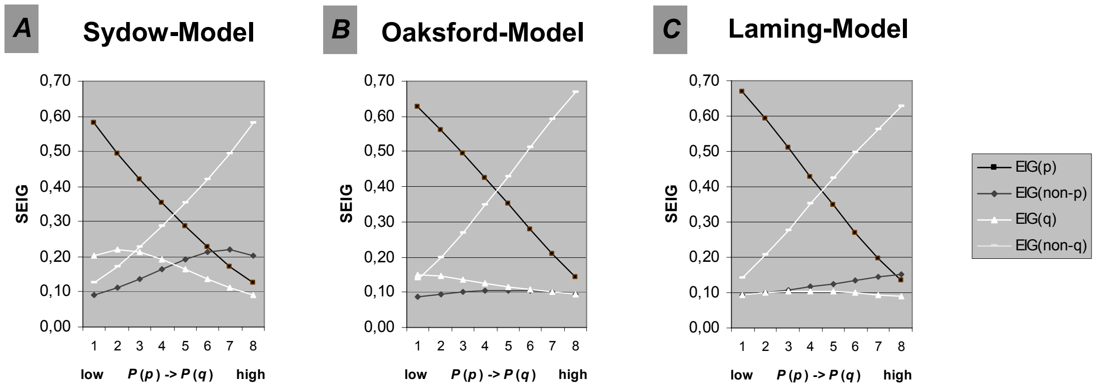


Figure 9. Model behaviour of (a) the Sydow model, (b) the Oaksford model (without q modification, cf. v. Sydow, 2004) and (c) the Laming model. The *SEIG* values of the four cards are plotted against the probabilities of '$P(p) \rightarrow P(q)$' ('1' stands for $P(p) = 0.1$, $P(q) = 0.2$; '2' for '$0.2 \rightarrow 0.3$' etc.).

The probabilities of the model parameters $P(p) \rightarrow P(q)$ ('.1 → .2', '.2 → .3', …, '.8 → .9') are plotted against the resulting scaled expected information gain values (SEIG) of the four cards.

6.1 Experiment 3a, b, c – Flexible Bayesian Models of Conditionals

In Experiment 3a, b, c the aim was to test the three structural models and to vary the probabilities of the parameters in the same experiment. The models are asserted to be relevant not only normatively, but also psychologically.

For example, if the conditional hypothesis "if virus then (always) symptom" is to be tested, it should make a difference whether we had independent knowledge about *P*(virus) and about *P*(symptom), or whether the virus is assumed to be new Asian birds virus. In the latter case, the truth of the hypothesis should lead to an increase of *P*(symptom), and to a decrease, if the hypothesis is false. The first case corresponds to the Sydow model, the second to the Oaksford model. After the experiment, the interpretations of the investigated models will be discussed in more detail.

In order to test the implications of the models more properly not only the card selections of the models will be accessed, but also the varying predictions for the resulting marginal probabilities.

Method

Design and Participants

The experiment had a 2 (low versus high probability condition) × 3 (structural models) between-subjects design. In all conditions the dependent variables were *p* versus *non-p* selections, *q* versus *non-q* selections and estimates for the resulting marginal probabilites given that the conditional is either true or false: $P(p_{res} \mid M_D)$, $P(q_{res} \mid M_D)$, $P(p_{res} \mid M_D)$, and $P(q_{res} \mid M_I)$.

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

Model Predictions for the Used Parameterisation

In Figure 9 (see above) the general model behaviour has been described. In Table 21 the predictions for the used parameter values are given (low probability condition:

$P(p) = .10$, $P(q) = .20$, $P(M_D) = .50$; and high probability condition: $P(p) = .80$, $P(q) = .90$, $P(M_D) = .50$).

Table 22

Expected Information Gain (EIG) and Scaled Expected Information Gain (SEIG) for the Cards and Resulting Marginal Probabilities $P(p_{res} \mid M_D)$, $P(q_{res} \mid M_D)$, $P(p_{res} \mid M_I)$, $P(q_{res} \mid M_I)$ for the Low Probability (.10 → .20) *and High Probability Conditions* (.80 → .90) *of the Different Structural Models*

	Low probability condition								High probability condition							
	EIG and SEIG				Probabilities				EIG and SEIG				Probabilities			
	for the cards				in M_D		in M_I		for the cards				in M_D		in M_I	
	p	$\neg p$	q	$\neg q$	p_{res}	q_{res}	p_{res}	q_{res}	p	$\neg p$	q	$\neg q$	p_{res}	q_{res}	p_{res}	q_{res}
Sydow	.61	.01	.15	.05	.10	.20	.10	.20	.05	.15	.01	.61	.80	.90	.80	.90
model	.58	.09	.20	.20					.12	.20	.09	.58				
Oaksford	.61	.00	.07	.05	.10	.28	.10	.20	.05	.00	.00	.61	.80	.98	.80	.90
model	.63	.09	.15	.13					.14	.09	.09	.67				
(changed)																
Oaksford	.67	.00	.10	.05	.10	.24	.10	.16	.09	.00	.00	.61	.80	.97	.80	.83
model	.63	.08	.16	.12					.17	.09	.09	.65				
(original)																
Laming	.61	.00	.00	.05	.02	.20	.10	.20	.05	.07	.00	.61	.72	90	.80	.90
model	.67	.09	.09	.16					.13	.15	.09	.63				

Table 21 shows the predictions for all three models, including two variants for the original and the modified Oaksford model (Oaksford and Chater 1994; see Table 11, p. 60). Beside the EIG and SEIG values for the card selections (cf. *Figure 9*), the resulting estimates for $P(p_{res} \mid M_D)$, $P(p_{res} \mid M_I)$, $P(q_{res} \mid M_D)$ and $P(q_{res} \mid M_I)$ are provided.

Materials and Procedure of Experiment 3a, b, c

Each participant was presented with a MST (cf. pp. 83 f.). In all conditions the same cover story was used. Participants were asked to suppose that they were physicians at a university hospital.

In all conditions 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 form of 100 patient cards. The front side each patient card provides information about tested viruses and, 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 did not have this virus (‘-’). Then she quickly took up the cards. Thereby the cards are completely mixed (bold print). Secondly she laid out the backsides of the cards, showing whether a patient has shown the specific symptom (‘●’) or not (‘○’).

Depending on the probability conditions, the cards that were shown were varied. The exact proportion of *p*- versus *non-p*-cards and *q*- versus *non-q*-cards resulted from how the parameters were set (low probability condition: $P(p) = 0.1$, $P(q) = 0.2$, high probability condition $P(p) = 0.8$, $P(q) = 0.9$).

In the structural Sydow condition with constant marginal probabilities $P(p \mid M_D) = P(p \mid M_I)$ and $P(q \mid M_D) = P(q \mid M_I)$ were induced by showing *all* fronts and backs of the cards (after mixing them in between). For the Oaksford-model, with $P(p \mid M_D) = P(p \mid M_I)$ and $P((q \mid non\text{-}p) \mid M_D) = P((q \mid non\text{-}p) \mid M_I)$, all cards were also first shown with the virus-side facing upwards ($P(p \mid M_x)$). But after mixing, the symptom-sides only those patient cards were shown, which had no virus ($P((q \mid non\text{-}p) \mid M_x)$). Thereby information is directly provided on $P(\neg p \wedge q)$ and $P(\neg p \wedge \neg q)$, which should remain constant in this model. In contrast, no direct information was provided for the *q* or *non-p* marginal probabilities, which are not constant in this model. Similarly, in the Laming condition, with $P(q \mid M_D) = P(q \mid M_I)$ and $P((p \mid q) \mid M_D) = P((p \mid q) \mid M_I)$, all cards were first shown with the symptom-sides facing upwards. After mixing, the virus-sides of the cards only of those patients were shown, who have had the specific symptom. Thereby information on $P(p \wedge q)$ and $P(\neg p \wedge q)$ was provided, which should be held constant in that model, but 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 separately. Participants were subsequently 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). 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 (○) (*q*-card, *non-q*-card). In both cases, they had to choose which card they would turn over.

Finally, four questions on a further page were used (in a frequency format), to survey the participant's estimation of the marginal probabilities resulting in each model: $P(p_{res}\,|\,M_D)$, $P(q_{res}\,|\,M_D)$, $P(p_{res}\,|\,M_I)$, $P(q_{res}\,|\,M_I)$. This was of especial interest to ascertain whether subjects had really understood the implications of the different models. It was mentioned that the previous instructions remained valid. Participants were first to suppose the hypothesis were true. 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*. After this were asked to assume that the hypothesis was false and were then asked the same questions. This additional dependent variable should allow to access whether the participants had really understood the implications of the models, which go beyond selecting particular cards.

Results and Discussion

Firstly, the card selections of the MSTs are reported and then the estimations of the marginal probabilities, which are of particular importance to access whether participants really grasped the implications of the three used models.

Card Selections

Table 23
Percentages (and Number) of Selections of the P or Non-P Cards and of Q or Non-Q Cards in the Low and High Probability Conditions of the Sydow Model, the Oaksford Model and the Laming Model (N = 72)

	(a) Sydow		(b) Oaksford		(c) 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)

Note. Predicted selections are darkened.

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 probability condition relative to the low probability condition. The descriptive results are shown in Table 23a, and they are visualized in Figure 10a.

Both differences were statistically significant, the *q* versus *non-q* effect (Pearson $\chi^2_{(1,\, n = 24)} = 6.0$, $p_{\text{one-tailed}} < .01$, $r_\varphi = .50$) as well as the *p* versus *non-p* effect (Pearson $\chi^2_{(1,\, n = 24)} = 10.9$, $p < .001$, $r_\varphi = .68$). The results of the *q* versus *non-q* selections were perfectly symmetrical. Also the *p* versus *non-p* effect had a high effect size, particularly for WSTs, and descriptively only shows a small *p* matching bias.[37]

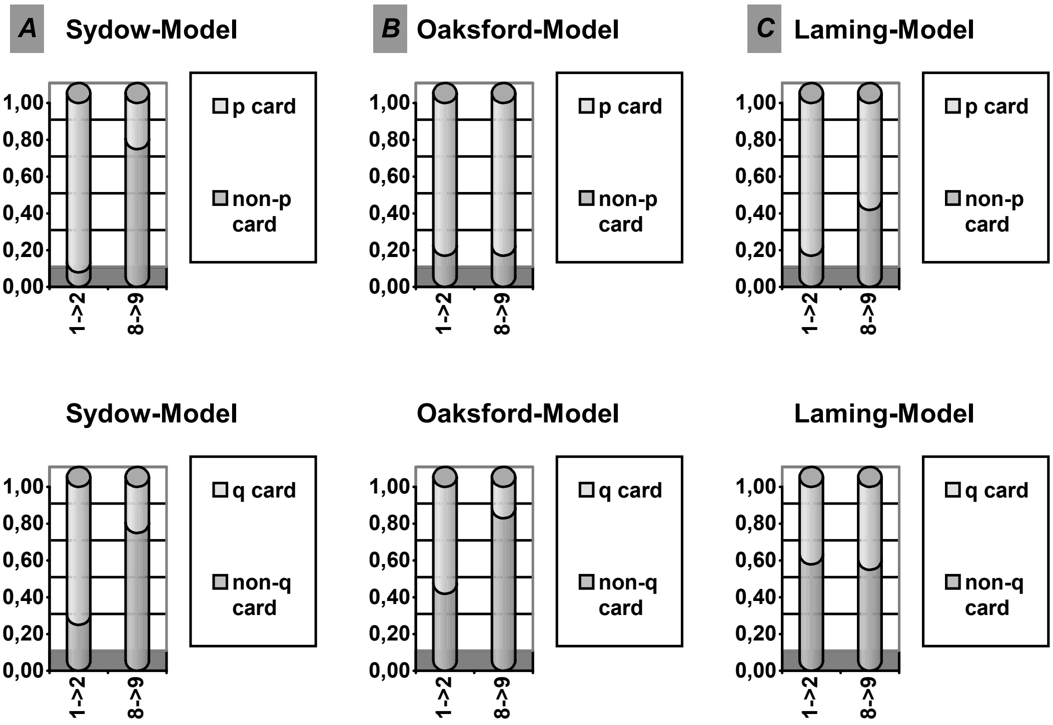


Figure 10. Bar graphs of the card *p* versus *non-p* and *q* versus *non-q* card selections in the low and high probability conditions of (a) the Sydow model, (b) the Oaksford model, and (c) the Laming model.

Oaksford and Chater (1994) model. According to this model, again an increase of *non-q card* selections is predicted in the high probability condition, but no increase of *non-p* selections is predicted. The results are shown in Table 23b and visualized in Figure 10b. For the *p* versus *non-p* cards there was indeed no difference between the low and high probability condition (exact Fisher test ($df = 1$, $n = 24$): $p = 1.00$, $r_\varphi = 0$). As hypothesized, the frequency of *non-q* card selections was significantly higher in the high probability condition than in the low probability condition (exact Fisher test (*1*, $n = 24$, one-tailed): $p < .05$, $r_\varphi = .43$). Even the high base rate of *non-q* card

[37] Remarkably, also w*ithin* conditions the differences were significant. For instance, in the high probability condition there were significantly more *non-p* than *p* and more *non-q* than *q*-selections (both: $\chi^2_{(1,\, n = 12)} = 3.0$, one-tailed, $p < .05$).

selections in the low probability condition, which may be surprising with respect to the standard selections in most descriptive WSTs, appears reasonable with regard to the *EIG* and *SEIG* values (cf. Figure 9*,* Table 22).

Laming (1996) model. Although Laming's model was originally only intended to provide an example of an absurd model, it was modelled and the prediction of a constantly high *non-q* card selection and a *p* versus *non*-p frequency effect was derived from it. As expected, no *q* versus *non-q* effect was found (exact Fisher test (1, $n = 23$): $p = 1.00$*,* $r_{\varphi} = .03$). But contrary to the predictions, the *p* versus *non-p* effect was not significant (exact Fisher test (1, $n = 24$, one-tailed): $p = .18$, $r_{\varphi} = .27$). However, even here the results pointed descriptively in the predicted direction: in the high probability condition over 40 % preferred a *non-p* card to a *p* card (Figure 10c).

In summary, the card selections clearly confirmed both the Sydow model and the Oaksford model, and in the Laming model they at least pointed in the predicted direction.

Estimates of Marginal Probabilities

The participants' estimates of the resulting marginal probabilities were used as another set of dependent variables to assess whether the participants really understood the implications of the induced models.

In Table 24 the means and modes of the subjective estimates of the marginal probabilities $P(p_{res})$ and $P(q_{res})$ are shown, conditional on the assumption either that the rule is 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 a few deviations can strongly influence the means, and since in some cases *two* types of answers were observed to dominate the selection patterns. Hence, Table 24 additionally shows the modes (two modes are shown when both had the same frequency or when their frequency differed only by one case). It was tested whether the number of cases represented by each mode (or by the two modes) is predominant relatively to all other cases not matching that mode (or these modes). This was tested for significance with one-dimensional χ^2 tests (*df=1, one-tailed,* $12 \geq n \geq 9$*).* The results of these tests are shown in Table 24.

In the von Sydow model there was only one mode for each case. All empirical modes matched the normative modes. 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 in all other estimations taken together (Table 24).

Table 24
Estimates of the Resulting Marginal Probabilities Given either the Truth (M_D) or Falsity (M_I) of the Hypothesis

$P(p_{res})$ $P(q_{res})$		Sydow			Oaksford			Laming		
		Normative	Mean	Mode	Normative	Mean	Mode	Normative	Mean	Mode
Low probability	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 probability	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[a] 90*	80 90	73 58	80* 90; 50	80 90	67 77	72 90*

Note. For each model the following values are shown: normative answers for $P(p_{res})$, then for $P(q_{res})$, means of these answers, modes of these answers. A mode (or two taken together) got an asterisk (*), if their predominance was also statistically significant ($p < .05$). (They also in each case united over 75% of the answers.)
[a] Also here 70 % (of 11 answers) matched the mode

In the Oaksford model and in the Laming model a relevant number of, but not all, estimations confirmed the predictions. But it will be shown that the deviations also showed an interesting consistence, at least for the Oaksford model.

In the Oaksford and the Laming model, the results significantly confirmed the predictions in regard to those marginal probabilities that had been held constant, this was $P(p_{res})$ in the Oaksford model, and $P(q_{res})$ in the Laming model. In both models there was only one mode for each of the four estimates. These modes outweighed all other estimates significantly. Also as predicted, a change of modes (between M_D and M_I) was found in the Oaksford model in regard to $P(q_{res})$, and conversely in the Laming model in regard to $P(p_{res})$. Hence, participants clearly understood which marginals were variable and which were not. But opposed to the predictions in both models two modes were found for the estimate of the variable marginal probability, given the hypothesis is assumed to be true. In the Oaksford model this is the case for the *q* estimates and in the Laming model this is the case for the *p* estimates. The two modes taken together in both models in each of the four estimates significantly outweighed all other estimations. In all these cases – independent of a high or a low probability condition – one of the two modes matched exactly or very closely the predicted mode. The other mode in all cases was consistent with an equivalence interpretation of the hypothesis. In the Oaksford model this second mode of $P(q_{res} \mid M_D)$ in the low and the high probability condition matched the predictions for $P(p_{res} \mid M_D)$. Conversely, in the Laming model the second mode of $P(p_{res} \mid M_D)$ exactly matched $P(q_{res} \mid M_D)$. This pattern explains the Oaksford model. In the Laming model there is an additional deviation found under the assumption that M_I is valid. Hence, at least in the Oaksford model one set of answers appears to show exactly the expected changes between M_D and M_I. Another set of answers seems to be consistent with an interpretation of the hypothesis as not representing an implication, but as a statement of equivalence. (Based on the low N no further analysis of this additional effect was possible.) The results for the Laming model also pointed to these two interpretations, but, as mentioned, showed additional problems.

In summary, also the results for the estimations show that participants distinguished the tested models. The results for the Sydow model were unambiguously positive. In the Laming model, participants understood which aspects of the model were flexible, but the estimates partly went in the wrong direction. For the Oaksford model, a substantial number of answers confirmed the predictions, although there was also a second group, whose estimates were 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.

6.2 General Discussion

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 confirmed the predicted differential effects at least of the Sydow and the Oaksford model. The descriptively positive results of the card selections for the Laming model did not reach significance. The estimates for the resulting marginal probabilities provided evidence that at least a substantial part of the participants in the Sydow and Oaksford model conditions additionally understood further aspects of these models.

Implications for Non-Bayesian Approaches

Approaches that are normatively based on standard formal logic and its falsificationist interpretation have clear normative predictions in all conditions of the experiment. In each and every case one equally ought to select *p* cards and *non-q* cards, since these are the only cards by which a (conclusive) falsification could be achieved.

The most influential psychological theory of the WST, the *mental model theory* (MM theory, Johnson-Laird & Byrne, 1991, 2002, see. pp. 11, 272) is normatively still tightly linked to the falsificationist research program. However, also the additional psychological assumptions of MM theory cannot explain the particular pattern of the frequency effects found.

Firstly, MM theory cannot explain the predicted and found *p* versus *non-p* effects in the Sydow model, since all postulated mental models of descriptive conditionals include *p* cases, which generally would lead to a selection of this case. A biconditional implication is implausible, since there were more *q* cases than *p* cases and subjects reported this difference in their frequency estimations even under the assumption that the hypothesis is true. (Even in case of a biconditional interpretation, it is not clear why these frequency effects should occur.)

Secondly, also *q* versus *non-q* frequency effects found in the Sydow and Oaksford model cannot be explained. It has been argued in detail in the General Discussion of Chapter 5 that MM theory provides predictions if one induces expectations about counterexamples, but that it does not make any clear prediction concerning the mere manipulation of $P(p)$ and $P(q)$. Additionally, it has been elaborated, that even if this gap were closed, MM theory would predict the reversed pattern of results (cf. Discussion of Chapter 5).

Thirdly, the simultaneous increase of *non-p* and of *non-q* in the Sydow model has never been predicted by mental model theory.

Fourthly, the contrast between the selections in the *different* models cannot presumably be explained by MM theory.

Last but not least, the additional probability estimations for $P(q_{res} \mid M_D)$ in the Sydow and Oaksford model cannot be explained by MM theory, since the representations of a conditional assumed by MM theory are much too basic to account for such more refined quantitative implications of different Bayesian models.

Also *mental logic theory* (see pp. 10 f.; Braine, 1978; Rips, 1994; O'Brien, 1995) cannot account for the found frequency effects, since a strictly deductive solution with a limited set of postulated mental inference rules does not account for any of the found frequency effects.

Likewise, the two domain-specific theories, *pragmatic reasoning schema theory* (see pp. 13 f.; Cheng & Holyoak, 1985, 1989; Holyoak & Cheng, 1995a, 1995b) and evolutionary *social contract theory* (see pp. 14 f.; Cosmides, 1989; Cosmides & Tooby, 1992, Gigerenzer & Hug, 1992), which both broke with any concept of normativity, cannot explain the fit of data to these Bayesian models. The empirical results of the experiment at least show the incompleteness of these theories. Moreover, although these approaches concentrated on the deontic domain, the systematic findings obtained here for descriptive conditionals are inconsistent with some of their claims that were made explicitly for the domain of testing descriptive conditionals as well. This has been discussed in detail in the General Discussion of Chapter 5 and will be discussed once more in the General Discussion of Part II (for other problematic aspects of these theories, cf. the General Discussion of Part III).

Additionally, other theories such as matching bias theory or relevance theory clearly cannot explain the found frequency effects (cf. General Discussion of Part II).

Hence, all non-Bayesian approaches are unable to explain the structural effects and frequency effects obtained here.

Implications for Bayesian Approaches – The Necessity of a Flexible Account

On the one hand, the results of Experiment 3 show that Oaksford and Chater's (1994, 1998a, 2003) discussed universal Bayesian approach to hypothesis testing needs to be extended by a structural component, which determines which aspect of the model is held constant independent of the truth or falsity of the conditional hypothesis. We were here concerned with the microstructure of a conditional; keeping the causal and logical macrostructure constant (see next section). On the other hand, although this necessitates an extension of this most refined Bayesian approach of the WST it (normatively as well as empirically) strengthens the Bayesian approach in general.

Earlier, in Chapter 5, it has been shown that positive results for the model with fixed marginals including *p* versus *non-p* effects can be obtained if the preconditions for this model have been introduced empirically using MSTs. In contrast to previous mostly negative (or ambivalent) results, this not only provided evidence in favour of a Bayesian approach but, indirectly, also in favour of the postulated knowledge-based approach.

Here this knowledge-based account was tested more directly, for the first time using explicitly varying different structural models in one experiment. The objection of Laming (1996), that the assumptions of the discussed basic models are licentious, is ruled out by introducing exactly the same preconditions of these models experimentally. By clearly fixing the preconditions of different models, support has been provided not only for the model von Sydow (2002; cf. Hattori, 2002; Oaksford & Chater, 2003) but also for the abandoned original model of Oaksford and Chater (1994, 1998, cf. similarly now Over & Evans, 2003; Evans, Handley & Over, 2003; and Over, 2004). (The problematic results for the Laming model are discussed in the next section.)

These confirmative results show the necessity of extending the universal Bayesian models. It follows that the concept of postulating only *one* universal Bayesian model for testing a conditional is false (e.g., Oaksford & Chater, 1994, 1998; also Oaksford & Wakefield, 2003; and even Over & Evans, 2003; Evans, et al. 2003; Over, 2004). Instead additional hidden preconditions need to be taken into account.

On the larger scale, the advocated synthesis of different models nonetheless sustains a concept of normativity; it even enables a consistent normative approach in the light of the criticism of Laming. Nevertheless, the synthesis allows for a plurality of preconditions, which seems to rather accord with a domain-specific account. The current results suggest a *domain-specific normative Bayesian approach* (cf. Chapter 0).

To evaluate this proposal a lot more research is needed and after an excursus on the causal interpretation of the models, a further model is proposed theoretically (cf. v. Sydow, 2006).

Excursus: Flexible Causal Models?

Although the Oaksford model, the Sydow model or the Laming model (Oaksford & Chater, 1994; von Sydow, 2002, Hattori, 2002; Oaksford & Chater, 2003; Laming, 1996) have all not been interpreted as causal models before, here a possible causal interpretation of these models should be presented (cf. von Sydow, 2004c). Such an interpretation may link two largely disconnected debates, the debate on the WST, and the debate on causal reasoning.

In regard of the current results, it is interesting to link both debates. From this two conclusions will be drawn. On the one hand, it shall be shown that such a link may provide a causal explanation of why the results for the Laming model were less adequate than those obtained for the Sydow or the Oaksford model. On the other hand, it shall be shown that the causal interpretation of these models also extends the standard visualisation of causal models in causal model theory (e.g., Waldmann, 1996; Waldmann & Hagmayer, 2001; cf. Cheng, 1997) or causal Bayes nets theories (e.g., Pearl, 1988, cf. 2000) (for an overview on theories of causal reasoning and causal learning see Hagmayer & Waldmann, in press, Meder, 2006).

(1.) First, we need to link causal graphs to our Bayesian models of the WST. In a nutshell, causal model theory and causal Bayes net theory both represent causal structures visually in the form of causal graphs, which guide learning and inferences (a detailed account of these theories is given by Meder, 2006). If this very basic idea, held by these theories, were applied to the three Bayesian models of information selection discussed in this chapter and instantiated in Experiment 3, which causal models would result?

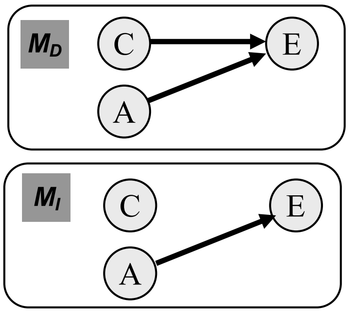


Figure 11. Standard causal graphs for the sub-models M_D and M_I for *all three* discussed models.

Actually, all three models would refer to the same pair of causal graphs (Figure 11). For all three Bayesian models investigated in Experiment 3 the potential cause (*c*) and the potential effect (*e*) was clearly ordered, the virus was a cause and the symptom was an effect. Moreover, in all three models it was questioned either whether a virus (*c*) caused a symptom (*e*), or whether, alternatively, *c* and *e* occur together only independently. Hence, all three models tested two competing hypotheses. There is the hypothesis of the deterministic truth of the conditional hypothesis (M_D, with $P((e \mid c) \mid M_D) = 1$). The alternative hypothesis has been an independence model (M_I), implying that the effects occur independently of the focused cause (*c*), and hence there needs to be an alternative cause (*a*) for these effects as well. Also in the dependence model an alternative cause (*a*) needs to be assumed since the relation between *c* and *e* is deterministic and it was obvious that $P(e) > P(c)$. This results in a common effect model. Of course, applying this to the information choice questions in the WST would in any case need further modelling steps, which are provided by the approach of Oaksford and Chater.

Since these considerations concerning the construction of the causal models, which correspond to M_D and M_I, are applicable to all three discussed basic models of the WST, the Sydow model, the Oaksford model and the Laming model, nothing seems to be gained by this causal representation. Moreover, the basic causal model representations, causal arrows, do not seem to account for the problematic findings of the Laming model. But even more problematic, these basic causal representations do not differentiate between the Sydow and the Oaksford model, and the corresponding different findings.

Indeed most earlier causal approaches that have been discussed in the field of logical (deterministic) reasoning (e.g., Cummins, 1996; cf. Krynski & Tenenbaum, 2004) have varied the actual causal structure (represented in a causal graph), for instance by adding alternative causes or disabling conditions (cf. Feeney, Handley, 2000). In contrast, here this basic causal structure was kept constant. In regard to previous variations of the causal structure in the WST context, Beller and Spada (1998, 2003) have convincingly argued that this can be reinterpreted as only adding logical premises. They concluded that these results may also be accounted for on a logical basis.

(Some experiments, that were discussed in the context of causal learning, e.g., Waldmann, 1996, really cannot be explained purely in logical terms, for instance, since they draw on the property of causal direction; cf., von Sydow, 2006.)

In any case, the additional qualitative and quantitative premises used here, cannot be understood as additional logical if-then statements or additional causal relations.

(2.) Hence, in order to account for the confirmed difference between the distinguished Bayesian models of information gain we need to add further structural symbols to the standard causal graphs (see Figure 12 and the following comments).

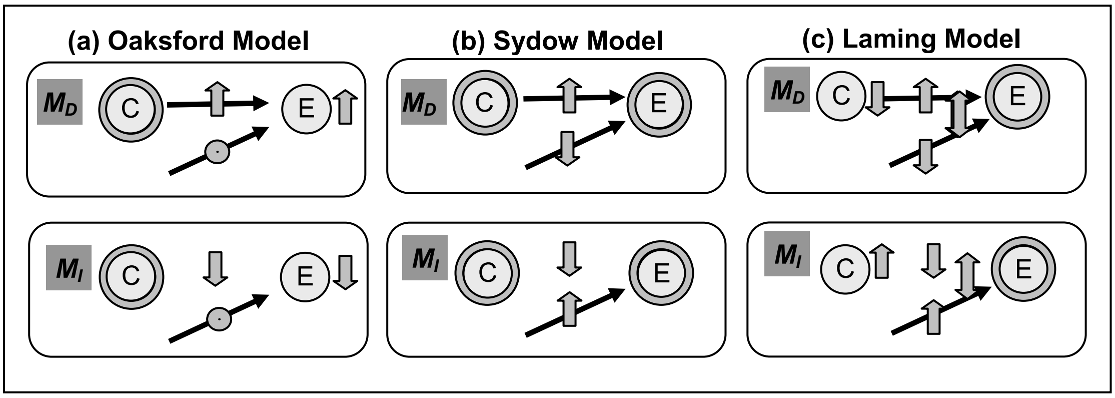


Figure 12. Causal graphs of the dependence submodels (M_D) and the independence submodels (M_I) of the Sydow model, Oaksford model, and Laming model. The upwards and downwards arrows show a relative increase or decrease in the probabilities of *c* or *e*, and they show an increase or decrease in the conditional probabilities of the causal arrows. The additional dark circles surrounding causes or effects, or the circles on the arrows show what is set to be constant independent of M_D or M_I.

In a causal perspective, all three models impose additional structural constraints going beyond the basic causal graphs (Figure 11). These constraints are structurally valid, regardless of the parameterisation of the model. They are valid in the low and high probability condition and they are given independently of whether the conditional is assumed to be true (M_D) or false (M_I).

A causal perspective, which would additionally consider these constraints, would interpret the models in the following way:

(a) The *Oaksford model* (Figure 12a) does not only describe that the truth or falsity of the conditional is in question (like in the other models as well), but additionally assumes that the impact of the alternative cause remains constant. In the Oaksford model, the probability $P(e \mid non\text{-}c)$ is fixed, independent of the truth or falsity of '$c \rightarrow e$'. The effects *e* in the *non-c* case have to be due to the alternative cause(s) *a*.[38] Hence, distinguishing the corresponding likelihood and the base rate of alternative cause *a,* the product $P(a) \times P(e \mid (a \wedge non\text{-}c))$ has to be constant (assuming the independence of cause *c* and cause *a*). Although also the probability of the cause is assumed to be constant, $P(c \mid M_D) = P(c \mid M_I)$, the probability of the effect (the number of symptoms of a disease) will increase in this model if the postulated deterministic causal relation is actually given, $P(e \mid M_D) > P(e \mid M_I)$ (if $P(c) > 0$).

The preconditions of this causal model should be fulfilled in real life situations, for instance, if a *new potential cause* occurs which can be assumed not to affect the probability of the other previous known alternative causes. For instance, a new virus may have occurred causing the symptom of sneezing. Here the probability of the symptom occurring without that virus will (presumably) not be affected by whether this new virus causes this symptom or not. But the number of those who get the symptom will depend on the truth or falsity of the causal claim. Such preconditions are given in presumably many situations.

Our results and the refined causal interpretation of the model (Oaksford & Chater, 1994, 1998a), which has been completely discarded by Oaksford and Chater (2003), shows that this model is not to be discarded.

Additionally, the model should now be derived from a causal perspective more formally, using standard causal modelling (Pearl, 1988, cf. 2000) and the additional

[38] Please note that our instruction did not distinguish between the base rate of the alternative cause $P(a)$ and the strength of the, perhaps probabilistic, causal arrow a ~> c, linked to the likelihood $P(e \mid (a \wedge non\text{-}c))$.

above premises. These premises include the concept of distinguishing two hypotheses (Oaksford and Chater, 1994), M_D and M_I in the process of testing a conditional and the adopted constraints.

Only the dependence sub-model (M_D) of the Oaksford model needs to be derived formally, since the independence model (M_I) seems to be rather trivial.

In the dependence model causal equations (facorisations corresponding to the modelled causal graph) for each of the four cells of the basic model need to be specified, in order to derive the same results for $P(c \wedge e)$, $P(c \wedge \neg e)$, $P(\neg c \wedge e)$, and $P(\neg c \wedge \neg e)$ as in the information gain model. Since effect e is assumed to be caused by cause c deterministically and also somehow by an independent hidden alternative cause a, both potential causes need to be considered.

Hence, in order to calculate the probability that cause c and effect e are present, $P(c \wedge e)$, the equations for a common effect graph would take into account all cases were c has been present and e occurred, regardless of whether the alternative cause is present or not:

$$P(c \wedge e) = P(c)P(a)P(e \mid (c \wedge a)) + P(c)P(\neg a)P(e \mid (c \wedge \neg a)) \qquad (23)$$

Since it is assumed here that c causes e deterministically, regardless of the alternative cause a, $P((e \mid c) \mid M_D) = 1$, the conjunctive likelihoods can be eliminated:

$$P(c \wedge e) = P(c)P(a) + P(c)P(\neg a) \qquad (24)$$

This is equivalent to $P(c)$, which corresponds to $P(p)$ in the Oaksford model (see Table 21a, p. 118).

For the second cell, the following equation has to be set up:

$$P(c \wedge \neg e) = P(c)P(a)P(\neg e \mid (c \wedge a)) + P(c)P(\neg a)P(\neg e \mid (c \wedge \neg a)) \qquad (25)$$

Since the causal relationship between c and e was postulated to be deterministic, it follows that $P(\neg e \mid (c \wedge M_D)) = 0$ and that $P(c \wedge \neg e) = 0$. For the third cell (and analogously for the fourth cell) the following equations are to be formulated:

$$P(\neg c \wedge e) = P(\neg c)P(a)\mathrm{P}(e \mid \neg c \wedge a) + P(\neg c)P(\neg a)\mathrm{P}(e \mid \neg c \wedge \neg a) \tag{26}$$

$$\Leftrightarrow P(\neg c \wedge e) = P(\neg c)[\, P(a)P(e \mid \neg c \wedge a) + P(\neg a)P(e \mid \neg c \wedge \neg a)] \tag{27}$$

The expression, $P(a)P(e \mid (\neg c \wedge a)) + P(\neg a)P(e \mid (\neg c \wedge \neg a))$, is equal to the originally given probability $P(e)$ or $P(q)$ in the independence model (opposed to $P(e_{res})$), since, according to the constancy assumptions of the causal Oaksford model (Figure 12a), the *non-c* cases in the dependence model (and the hidden causes) have exactly the same probability of eliciting the effects, as they have generally in the independence model. It follows that:

$$P(\neg c \wedge e) = P(\neg c)P(e) = (1 - P(c))P(e) \tag{28}$$

The argument to achieve $P(\neg c \wedge \neg e) = (1 - P(c))\,(1 - P(e))$ is analogous. In conclusion, all four expressions derived on the basis of causal models and our additional assumptions correspond to the Oaksford model (without *q*-modification, see Table 21a, p. 118). We now can proceed with discussing the constraints of the other models.

(b) The *Sydow model,* additionally to the standard causal graph of a common effect model (Figure 11), imposes the structural constraint that not only $P(c)$ but also $P(e)$ is fixed, independent of whether the potential cause elicits the effect or not (Figure 12b). Here I refrain from deriving the causal interpretation more formally.

The used constraints, that $P(c \mid M_D) = P(c \mid M_I)$ and $P(\mathrm{e} \mid M_D) = P(\mathrm{e} \mid M_I)$, are not only mathematically possible, but they also seem to be plausible for real life settings. For example, there may be two independent medical surveys, one concerning the prevalence of a particular virus *c*, and one concerning the prevalence of the symptom that someone is sneezing, symptom *e*. In this case, both $P(c)$ and $P(e)$ are known, independent of whether *c* really causes *e* or not.

Interestingly, it follows from the resulting Sydow model that the impact of the alternative cause *a,* $P(a) \times P(e \mid (a \wedge non\text{-}c)$, here is (epistemologically) *not* independent from the truth or falsity of the tested hypothesis '$c \rightarrow e$'. (One may perhaps say that the assumption of modularity, normally used for Bayes nets, is violated on an epistemological level.) If '$c \rightarrow e$' is true (M_D), fewer cases of *e* have to be caused by *a,* than if '$c \rightarrow e$' is false (M_I).

If we additionally assumed a fixed base rate for the alternative cause then the truth of *'c → e'* would imply a lowered causal strength of the alternative cause. This phenomenon on the type level of a causal relation resembles 'explaining away' effects, for long discussed on the token level in the causality debate. From a causal perspective, this lower impact of the alternative cause may either be attributed to a lower base rate of a, $P(a)$, or to a weaker causal strength of the probabilistic causal relation *a ~> e*, $P(e \mid (a \wedge \textit{non-c})$. If we alternatively assumed a deterministic causal relation between *a* and *e*, then the truth of the implication *'c → e'* would imply a reduced base rate of *a*. Hence, although the Sydow and the Oaksford model, in regard to the causal graphs, only tested the existence or non-existence of a causal arrow, we see that in the Sydow model – unlike the Oaksford model – the alternative cause (and hence also *non-c* cases) is in any case affected by the truth or falsity of '*c* → *e*'. This idea is reflected by the results obtained in Experiment 3.

(c) Likewise, the *Laming model* (Table 21c, p. 118) additionally to a standard common effect model assumes further constraints, which are normally not represented in causal graphs (Figure 12c*,* cf. Figure 11). In its causal interpretation, the Laming model assumes that $P(e)$ is fixed, and that the number of effect cases *e* is not affected by the truth or falsity of '*c* → *e*'. But unlike the Sydow model, here $P(c)$ is formally allowed to vary depending on the truth or falsity of '*c* → *e*'. Since, $P(c \mid e)$ is additionally assumed to be fixed (to be the same in M_D and M_I) $P(c)$ needs to be reduced if '*c* → *e*' is assumed to be true. This constraint is implausible in a causal domain.

However, let us assume a virus *c* is one of many potential causes of symptom *e*. Additionally, assume one has firm knowledge about the prevalence of a particular symptom *e*, $P(e)$, and about how many of the bearers of that symptom are infected by virus *c,* P(c | e) (without knowing whether *c* causes *e* or not). If one assumes that 'virus *c* → symptom *e*' is deterministically true, the overall number of persons carrying virus *c,* $P(c)$, is reduced, because the deterministic relation '*c* → *e*' denies the existence of any cases of '*c* ∧ *non-e*' and $P(c \wedge e)$ cannot have changed, since it was known anyway how many bearers of the symptom had the virus, $P(c \mid e)$.

Since the overall number of non-symptom bearers, $P(\textit{non-e})$, is also known in advance, a reduction in the number of non-symptom bearers on the side of virus *c* carriers additionally implies an increase in the number of non-symptom bearers on the side of persons who do not carry virus *c* (but an alternative virus *a*) and hence a lower causal impact of cause *a*.

Only if this long train of reasoning is understood, one can conclude that if '*c* → *e*' is true, there are less *c* cases and the impact of the alternative cause is reduced.

(3.) However, this is a difficult train of argument. Hence, from a causal perspective it is warranted to call the constraint $P(c \mid e)$ 'unnatural'. Correspondingly, the results of Experiment 3 show that the Laming model has only been partially been understood by the participants. These difficulties become understandable in a causal perspective, showing that the constraints in the Laming model lead to much more complicated causal model than the Oaksford or the Sydow model. In contrast, without such causal considerations the Oaksford and the Laming model are formally of equal complexity (see Table 21a, c, p. 118). Hence, the complexity of a (refined) causal model, and not of the logical models assumed by mental model theory, may provide an account to explain why the Laming model is more difficult to understand than the Oaksford or the Sydow model. In this sense the (refined) causal modelling in turn goes beyond the traditional information gain approach.

(4.) In summary, it has been proposed that the distinguished models of information choice, the Oaksford model, the Sydow model and the Laming model, may also be understood based on a causal model approach (e.g., Waldmann, 1996; Pearl, 1988, cf. 2000). However, all three investigated models lead to the same basic causal graphs. Only if the standard graphs are extended to allow for representation of the additional constraints, the differences between the models of information choice to test a conditional become visual in these graphs as well. It was derived formally that the original Oaksford model, discarded by Oaksford and Chater (2003), is equivalent to a plausible causal model. Moreover, the extended causal representations allow us to explain why participants had more problems with the Laming model than with the Oaksford model. It was suggested that a link between the knowledge-based understanding of the information gain approach of Oaksford and Chater (1994, 2003) and the causal model approach (e.g., Waldmann, 1996) may be mutually seminal.

On the one hand, the application of causal models may help to understand the differences in the difficulty between the postulated models of hypothesis testing. On the other hand, causal model theory is provided with an information gain account and is extended since the obtained results necessitate an extended representation of the causal graphs.

However, now another flexibilisation of the Bayesian test of the truth or falsity of a conditional will now be shown, which likewise cannot be directly derived from standard causal models, but which is naturally derivable from the more general matrix representation of the Bayesian models discussed before.

Further Flexibilisation – Ideas for Future Research

Not one single universal Bayesian model of testing a conditional in a WST can be derived from the advocated knowledge-based Bayesian approach to hypothesis testing, but several. These models should depend on the preconditions provided in the instruction (or on our knowledge about these preconditions).

Experiment 3 for the first time confirmed the different predictions for the Oaksford model (Oaksford and Chater, 1994, 1998) and the Sydow model (Sydow, 2002; Hattori, 2002; Oaksford & Chater, 2003) *within* the same experiment.

The idea to combine a Bayesian approach and different structural models may not only be relevant for the WST debate, but may also shed light on related research on the truth table tasks, or on *a*-, *b*-, *c*- or *d*-cell biases and different strategies in the evaluation of contingency tables (see e.g., Kao & Wassermann, 1993; White, 2000; McKenzie & Mikkelsen, 2000; Barres & Johnson-Laird, 2003; Evans, Handley & Over, 2003; cf. v. Sydow, 2006, pp. 16-18).

In regard to the WST, only one further model will be outlined here (for other proposals see v. Sydow, 2006). If the flexible or knowledge-based Bayesian account is right, I think, it is possible to construct a model of a conditional that should *structurally* lead to only logical-falsificationist *p* and *non-q* selections – also from a Bayesian perspective. Whereas logical *p* and *non-q* selections should be obtained in the Oaksford model in high probability conditions, here a model should be formulated for which these selections are *generally* correct, independent of the parameters $P(p)$, $P(q)$ and $P(H_D)$.

Table 25

Basic Model of the Test of a Deterministic Conditional in which all Cells but the B-Cell, $P(p \wedge non\text{-}q)$, are Held Constant Between M_D and M_I

(a) M_D	*Q*	*Non-q*	Marg.
P	$p\,q$	0	$p\,q$
Non-p	$(1 - p)$ q	$(1 - p)$ $(1 - q)$	$1 - p$
Marg.	q	$(1 - q)$ $(1 - q)$	$1 - p +$ $p\,q$

(b) M_I	*Q*	*Non-q*	Marg.
P	$p\,q$	p $(1 - q)$	p
Non-p	$(1 - p)$ q	$(1 - p)$ $(1 - q)$	$1 - p$
Marg.	q	$1 - q$	1

Note: In the cells of M_D and M_I the following notation is used: $p := P(p)$, $q := P(q)$

Table 25 exactly provides such a basic model of testing a conditional. By backward design, the dependence model M_D was constructed to differ from the independence model M_I, only in the falsificatory '$p\wedge\neg q$' case. The dependence model is simply constructed by removing all contradicting cases from the independence model (by setting $P(p \wedge \neg q) = 0$). In this model some marginal probabilities are, of course, constant ($P(non\text{-}p_{res}\,|\,\mathrm{M_D} = P(non\text{-}p_{res}\,|\,\mathrm{M_I})$, $P(non\text{-}q_{res}\,|\,\mathrm{M_D}) = P(non\text{-}q_{res}\,|\,\mathrm{M_I})$), whereas others vary ($P(p_{res}\,|\,\mathrm{M_D}) < P(p_{res}\,|\,\mathrm{M_I})$, $P(q_{res}\,|\,\mathrm{M_D}) < P(q_{res}\,|\,\mathrm{M_I})$). Moreover, it should be noted that the overall probability of all cases taken together also varies between the dependence and independence model.

There may well be real live situations in which such a model is applicable. Let us assume there is a flock of hundred sheep, which are either black (*p*) or white (*non-p*), and which have either pointed (*q*) or curved horns (*non-q*). Normally, these two physiological features can be assumed to be independent (M_I). Now someone alleges that a wolf has killed exactly those sheep, which are white and which (at the same time) have curved horns (M_D). In this situation the conditional, 'If a sheep in this population is white then it always has pointed horns', becomes true. If now someone asked you to test this conditional hypothesis, *p* and *non-q* cases would be the correct solution, not only from a falsificationist viewpoint, but also from a knowledge-based Bayesian viewpoint. For this model the advocated flexible Bayesian account structurally predicts the predominance of *p* and *non-q* patterns – independent of $P(p)$ and $P(q)$. Additionally subjects should understand that the truth of the hypothesis is linked to a reduction in the overall number of sheep in this flock.

This thought experiment makes apparent that there may even be Bayesian models, which would lead to predictions completely coherent with the falsificationist approach. In this perspective, the falsificationist prediction appears to be only a special case resulting from a more general, but knowledge-based, Bayesian account.

7 The Bayesian Logic of the WST

7.1 Theoretical Proposal of a Bayesian Logic of the WST

All previous Bayesian approaches to information selection in the WST have only been concerned with hypotheses in the form of an implication or a biconditional (e.g., Evans & Over, 1996; Evans, Over, & Handley, 2005; Green and Over, 1998; Hattori, 2002; Kirby, 1994; Klauer, 1999; Oaksford et al., 1994, 1997, 1998a, 1998b, 1999, 2001, 2003a, 2003b; Oberauer et al. 1999, 2004, Over et al., 1994, 1998, 2003; von Sydow, 2002, 2004a). The Bayesian account of information gain has neither theoretically nor empirically been extended to other logical connectors. This would provide another critical test for the Bayesian approach. We will introduce this approach in analogy to connectors of propositional (and predicate) logic.

The few investigations of other (deterministic) logical connectors in a WST context have yielded incoherent results. The alternative connectors were often not tested separately but in more complex expressions and together with negations (Wason & Johnson-Laird, 1969; van Duyne, 1974; cf. Evans, Legrenzi, Girotto, 1999; see also the THOG task on exclusive disjunctions, e.g., Wason, 1977, Griggs, Platt, Newstead, Jackson, 1998).

Here I will try for the first time to extend the Bayesian information gain approach for testing conditionals (Oaksford & Chater, 1994, 1998a, 2003; v. Sydow, 2002, 2004a) to other logical connectors and to test them without the use of negations. The different connectors are further structural models, differing from the conditional Sydow model in that they set different cells in the dependence model to zero. Such an extension makes the logical basis still inherent in the Bayesian approach apparent. On the other hand, the logical basis is combined with the additional use of prior knowledge about probabilities and structure. Although the advocated Bayesian approach goes beyond the traditional falsificationist logic of hypothesis testing (Popper, 1934/2002, 1972, 1974, 1996; cf. e. g. Wason, 1966; Johnson-Laird &

Byrne, 1991; Kirby, 1994), the logical basis of the resulting predictions allows us to call this approach a Bayesian logic of hypothesis testing.[39]

Table 26

The 16 Connectors of Propositional Logic and Their Quantitative Preconditions

Logical connector		Tautology / verum	Adjunction[a] (inclusive OR)	Replication (inv. impl.)	Implication, conditional	Exclusion (NAND)	Affirmation *p*	Affirmation *q*	Contravalence[a] (exclusive OR)	Equivalence, biconditional	Negation *q*	Negation *p*	Conjunction	*P* and *non-q*	*Non-p* and *q*	*Non-p* and *non-q* (NOR)	Contradiction / falsum
Number		C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9	C_{10}	C_{11}	C_{12}	C_{13}	C_{14}	C_{15}	C_{16}
Symbol		T	∨	←	→	↑	p	q	>-<	↔	¬q	¬p	∧			↓	⊥
p	*q*	0 F	1 false cell				2 false cells						3 false cells				4 F
W	w	w	w	w	w	f	w	w	f	w	f	f	w	f	f	f	f
W	f	w	w	w	f	w	w	f	w	f	w	f	f	w	f	f	f
F	w	w	w	f	w	w	f	w	w	f	f	w	f	f	w	f	f
F	f	w	f	w	w	w	f	f	f	w	w	w	f	f	f	w	f
Preconditions for the truth of the connectors		■	◥	◢	◤	◣	|	‾	\	/	_	|	▾	▴	▾	▴	
		always true	$p + q \geq 1$	$p \geq q$	$p \leq q$	$p + q \leq 1$	p=1, ¬p=0	q=1, ¬q=0	$p = \neg q, \neg p = q$	$p = q, \neg p = \neg q$	$q = 0, \neg q = 1$	$p = 0, \neg p = 1$	$p = q = 1, \neg p = \neg q = 0$	$p = \neg q = 1, \neg p = q = 0$	$\neg p = q = 1, p = \neg q = 0$	$\neg p = \neg q = 1, p = q = 0$	always false

Note: In the first row, the name of the connector is given; in the second row the connector, C_i, is numbered. In the third row a used symbol for the connector is depicted. The following five rows provide the truth table definition of the connectors (they are ordered by the number of forbidden cells, which also in our probabilistic models always take the value zero). In the final two rows, show the quantitative preconditions for the truth of the connectors. Firstly, they are illustrated graphically by possibility spaces (cf. Figure 13, p. 144) and, secondly, they are formulated mathematically (p: = $P(p)$ and q := $P(q)$).

[a] The inclusive OR and the exclusive OR are both also called disjunctions.

Here we will only be concerned with dyadic deterministic connectors. As before, the dependence hypothesis (M_D) will be tested against an independence model (M_I) – although other alternative models may again be possible. But there is only a small

[39] This Bayesian logic of hypothesis testing has to be distinguished from fuzzi logic, cf., e. g., Hajek, 2002.

subset of the 16 deterministic dyadic connectors that can reasonably be used for probabilistic hypothesis testing in a WST (cf. Table 26).

Firstly, most connectors have to be excluded, because their truth leaves no degree of freedom in regard of the frequency parameters $P(p)$ and $P(q)$ (Table 26). Their truth only allows for only one configuration of probabilities.

For example, let us consider the basic model for the logical conjunction (Table 26, C_{12}). The logical conjunction is only true if *p* AND *q* are true, and is false otherwise. In the following examples the word ‘and’ is interpreted in this way: “all physical entities always have the attributes extension and weight”; “When we went out, you always drunk wine and ate haddock”. [40]

In the first example only the observation of a physical entity without extension or without weight would falsify this conjunction, or, in the second example, a case in which you did not drink wine or in which you did not eat haddock. In the given universe of discourse the truth of the conjunction presupposes that $P(p) = P(q) = 1$ and that $P(non\text{-}p) = P(non\text{-}q) = 0$ (cf. Figure 13a, p. 144).

Table 27
Basic Model of the Conjunctive Connector (P AND Q)

(a) M_D	*q*	*non-q*	marg.
p	1	0	1
non-p	0	0	0
marg.	1	0	1

(b) M_I	*q*	*non-q*	marg.
p	$p\,q$	p $(1 - q)$	p
non-p	$(1 - p)$ q	$(1 - p)$ $(1 - q)$	$1 - p$
marg.	q	$1 - q$	1

Note: In the cells of M_D and M_I the following notation is used: $p{:=} P(p)$, $q{:=} P(q)$

If we construct a Bayesian model for the conjunction (see Table 27) analogous to previous models, the dependence model (M_D) has to be formulated completely independently of any frequencies (cf. the following Figure 13a). Moreover, we cannot construct a model with fixed marginals. All marginal probabilites in the dependence model differ from those in the independence model (if $P(p \mid M_I) <> 1$ and $P(q \mid M_D) <> 1$). Despite these problems, such a basic model may perhaps be relevant to some other task.

[40] For our present purpose the difference between propositional and predicate logic can be ignored (as it is often ignored in the psychology of reasoning).

Nevertheless, more seriously, this deterministic model cannot be applied to a selection task. The four visible cards sides in the WST show instances of the marginal cases *p*, *non-p*, *q* and *non-q*; hence, this procedure denies that *P*(*non-p*) and *P*(*non-q*) are asserted to be zero. To show a *non-p* card side or a *non-q* card side (with or without selecting them) falsifies the hypothesis 'always *p* AND *q*'. In this respect, the predictions made using a Bayesian approach for a corresponding WST would be largely identical to those of falsificationism – no cards need to be turned over since the hypothesis is obviously false anyway.[41]

Generally, all connectors C_i in which at least one of the marginals $P(p_{res})$, $P(non\text{-}p_{res})$, $P(q_{res})$, or $P(non\text{-}q_{res})$ is necessarily zero if the hypothesis is true (C_6, C_7, C_{10}, C_{11}, C_{12}, C_{14}, C_{15}, C_{16}), cannot reasonably be tested in a WST, since the cards shown in the task would prove the hypothesis false - before even starting any selection.

Of course, also the connector of a tautology, $C_{1,}$ that is any claim that is empirically empty and true merely because of its form, cannot be tested in a WST.

Hence, only the connectors C_2, C_3, C_4, C_5, C_8, and C_9 remain as candidates for possible WSTs (Table 26).

Of these six connectors, a further two may be excluded if our goal is to test the Bayesian predictions of connectors, since they do not allow for an independent variation of the probabilities *P*(*p*) and *P*(*q*). The equivalence and the contravalence (C_8 and C_9 in Table 26) both can be used in a WST and even in a WST with fixed marginals (a Sydow model), but both refer to two 'forbidden' cells. These two cases lead to quantitative constraints so that *P*(*p*) and *P*(*q*) cannot be varied independently from each other.

For example, the equivalence hypothesis '*p* if and only if *q*' ('*p* iff *q'*, '*p* is sufficient and necessary for *q'*, C_9) quantitatively presupposes *P*(*p*) = *P*(*q*) and *P*(*non-p*) = *P*(*non-q*) (see Figure 13b).

Likewise, the (possible) truth of the contravalence hypothesis 'either *p* or *q*' (C_8) presupposes *P*(*p*) = *P*(*non-q*) and *P*(*non-p*) = *P*(*q*).

Hence, these two connectors are also not suitable for varying the probabilities *P*(*p*) and *P*(*q*).

[41] However, I want to point out the possibility to extend the Bayesian model of deterministic connectors, like the conjunction, in a probabilistic way, analogously to the mentioned probabilistic extensions of the Bayesian model of the conditional, by adding an error term to each cell. In a different context, I have collected evidence in favour of such a proposal and I plan to elaborate on this possibility in my future research.

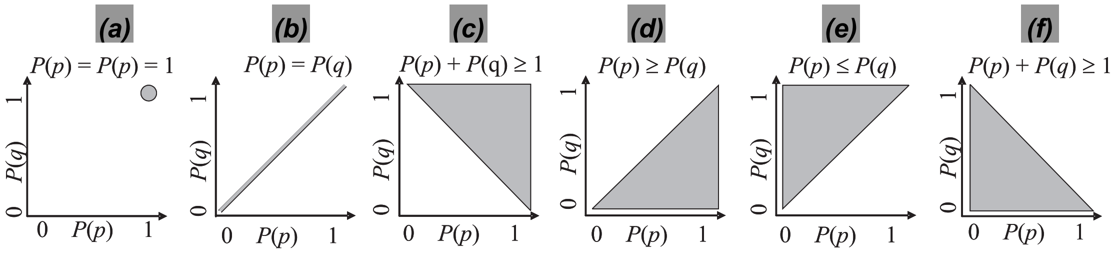


Figure 13. Possibility spaces for $P(p)$ and $P(q)$ of different connectors. The area that is darkened in each graph represents the parameter region for which the quantitative preconditions for the possible truth of a connector are fulfilled. In contrast, the remaining white areas show the parameter regions in which a connector is false *a priori*, without turning over any card. (a) conjunction, (b) equivalence, (c) adjunction, (d) replication, (e) implication, (f) exclusion (cf. Table 26).

Only four connectors (C_2 to C_5) can be used in a selection task to vary the probabilities of $P(p)$ and $P(q)$ to some extent independently. These four connectors only have one forbidden cell, leaving the degree of freedom needed to vary the probabilities of the cells. These connectors should be called ‘Bayesian connectors’: The implication, the replication, the adjunction, and the exclusion. Here we present them in models with fixed marginals, $P(p_{res} \mid M_D) = P(p_{res} \mid M_I)$ and $P(q_{res} \mid M_D) = P(q_{res} \mid M_I)$. The new corresponding basic models are shown in Table 28 to Table 30 while their possibility spaces are shown in Figure 13c, d, e, f.

Table 28

Basic Model of the Replication (P ← Q; Necessary Condition) With Fixed Marginals

(a) M_D	*q*	*non-q*	marg.
p	q	$p - q$	p
non-p	0	$1 - p$	$1 - p$
marg.	q	$(1 – q)$	1

(b) M_I	*q*	*non-q*	marg.
p	$p\,q$	$p\,(1 - q)$	p
non-p	$(1 - p)\,q$	$(1 - p)\,(1 - q)$	$1 - p$
marg.	q	$1 - q$	1

Note: In the cells of M_D and M_I the following notation is used: $p := P(p)$, $q := P(q)$

Table 29

Basic Model of the Adjunctive Connector (P OR Q; P ∨ Q; Inclusive Disjunction) With Fixed Marginals

(a) M_D	*q*	*non-q*	marg.
p	$p + q - 1$	$1 - q$	p
non-p	$1 - p$	0	$1 - p$
marg.	q	$(1 - q)$	1

(b) M_I	*q*	*non-q*	marg.
p	$p\,q$	$p\,(1 - q)$	p
non-p	$(1 - p)\,q$	$(1 - p)\,(1 - q)$	$1 - p$
marg.	q	$1 - q$	1

Note: In the cells of M_D and M_I the following notation is used: $p := P(p)$, $q := P(q)$

Table 30

Basic Model of the Exclusion Connector (P NAND Q; P ↑ Q) With Fixed Marginals

(a) M_D	*q*	*non-q*	marg.
p	*0*	*p*	*p*
non-p	*q*	1 - *p* - *q*	1 - *p*
marg.	*q*	(1 - *q*)	1

(b) M_I	*q*	*non-q*	marg.
p	*p q*	*p* (1 - *q*)	*p*
non-p	(1 - *p*) *q*	(1 - *p*) (1 - *q*)	1 - *p*
marg.	*q*	1 - *q*	1

Note: In the cells of M_D and M_I the following notation is used: $p := P(p)$, $q := P(q)$

The model with fixed marginals for the implication (if *p* then always *q*; '$p \to q$') has been presented before (cf. Table 21b, p. 118). Also this model has general quantitative constraints. There cannot be less *q* cases than *p* cases, because otherwise the deterministic hypothesis '$p \to q$' would be false: $P(p) \leq P(q)$ (Figure 13e).

In Table 28 the model for the inverse implication ('replication' or 'necessary condition') is presented. This connector claims that *p* is a necessary precondition for *q* to occur, but *p* may occur without eliciting *q*. Logically, we are still concerned with a deterministic connector. In contrast to the implication, here $P(non\text{-}p \wedge q \mid M_D)$ is set to zero in the dependence model. Correspondingly, the quantitative precondition for the validity of the model is reversed: $P(p) \geq P(q)$ (see Figure 13d).

In Table 29 the adjunctive Bayesian model (inclusive OR) with fixed marginals is presented. For example, we may want to test the following hypothesis in regard of a population of laboratory animals: "It is always the case that an animal is contaminated with the toxic substance *p* or with the toxic substance *q* or with both toxic substances." The dependence model represents the truth of the inclusive OR, instead of the independence model. But as before, the adjunctive dependence model (M_D) does not only differ and from the independence model M_I in regard of the falsifying case $non\text{-}p \wedge non\text{-}q$ (there should be no such cases, if M_D is true), but there should also be a difference, for instance, between $P(p \wedge q \mid M_D)$ and $P(p \wedge q \mid M_I)$. This also results in positive information gain values for the *p* and *q* cards, which cannot falsify the rule. The quantitative preconditions of the adjunction, $P(p) + P(q) \geq 1$, are also depicted in Figure 13c. If it is true that all members of a population have *p* or *q* or both, there needs to be at least as many *p* and *q* cases as to cover the whole population. For example, if $P(p) = .8$, $P(q) = .7$ then the hypothesis may be true, but it remains possible that the hypothesis is also false, for instance if all *q* instances occur together with *p*.

Finally, Table 30 spells out the basic model with which to test the hypothesis that *p* and *q* never occur together (logical exclusion, ‘NOR’, ‘not both’). This Bayesian model presupposes that the probabilities for *p* and for *q* do not add up to one: $P(p) + P(q) \leq 1$ (see Figure 13f). Otherwise, there would be cases in which both, *p* and *q*, would necessarily occur together.

In conclusion, there are four ‘Bayesian connectors’ of which we may vary the probability of $P(p)$ and $P(q)$ independently, all presupposing different quantitative constraints.

7.2 Experiment 4 to 7 – Towards A Bayesian Logic

In the following four experiments all four ‘Bayesian connectors’ will be tested. Since the Bayesian approach to the WST has not been previously extended to other connectors, these investigations will for the first time not only assess the Bayesian predictions for testing an implication but also those for testing a replication, an adjunction and an exclusion.

The four individual experiments on a different ‘Bayesian connectors’ will be tested together, likewise using similar context stories for all connectors. Hence, the four experiments may be regarded to be one. Nonetheless, for each connector the probabilities will be varied in four steps, resulting in 16 different tasks. It will be suitable to analyse the four connectors separately as four different experiments.

MSTs (cf. previous Chapters) are used to meet all preconditions of the model with fixed marginals. In each experiment, one probability condition should violate the quantitative preconditions of the model. In contrast to previous experiments on conditionals (e.g., Oaksford et al., 1999[42]) here the aim is to check whether subjects realised whether this precondition has been violated, allowing them to drop out of the selection procedure if this is the case.

Method

Design, Task Order and Participants

Design and Task Order. In all four experiments, each of which is concerned with a different connector (adjunction, implication, replication, and exclusion), four

[42] Previous frequency manipulations with conditionals have been discussed and criticised in depth in Chapter 3.2, 5 and 6 (cf. particularly pp. 70 f., 83 f.)

probability conditions are investigated (*P*(*p*) and *P*(*q*) are low-low, low-high, high-low, high-high). This results in a total of 16 conditions.

Each participant received a booklet containing four tasks. The tasks were arranged so that no participant got the same connector or the same frequency condition twice. A nested Latin square design was used in order to show each connector and each frequency condition equally often in all four positions. This resulted in 16 used different task orders (connector = number, frequency condition = letter; the first two orders were: 1A, 2C, 3B, 4D; 3C, 1B, 4D, 2A; etc.). Hence, this results in a between-subject design for each connector.

Participants. Eighty students from the University of Göttingen took part in the experiment (79 % female, 21 % male; mean age 22 years). The vast majority of the students studied psychology (85 %). The participants were randomly assigned to each of the 16 used series of tasks.

Predictions

The models for the adjunction, the replication and the exclusion (Table 28 to Table 30) have been modelled along the same lines as the implication (cf. pp. 62 f.).

In Table 31 the resulting expected information gain values (*EIG*) for the *p*, *non-p*, *q* and *non-q* cards are given for the four connectors and the four used probability conditions. In each of the four experiments (the four connectors), a low-low, a low-high, a high-low and a high-high probability condition was used for *P*(*p*) and *P*(*q*). The exact values for *P*(*p*) and *P*(*q*) are given in the table. They vary from connector to connector, so that for each connector the quantitative preconditions for testing that connector (cf. Figure 13c, d, e, f) were only violated in one out of four conditions (‘violation of constraint’).

Falsificationism does not predict frequency effects, but always the selection of the two cards which enables location of the forbidden case in the truth table of each connector, for instance, *p* and *non-q* in the case of the implication or *non-p* and *non-q* cases in the case of the adjunction (inclusive or).

Table 31

The 16 Conditions of Experiment 5 to 8 and the Model Behaviour (EIG Values)

		(A) Low-low conditions	(B) Low-high conditions	(C) High-low conditions	(D) High-high conditions
Exp. 5 adjunction	$P(p) \vee P(q)$	.20 ∨ .15	.20 ∨ .85	.85 ∨ .20	.85 ∨ .80
	EIG of *p* versus *non-p*	Violation of constraint	$p > \boldsymbol{\neg p}$ .28 > .08	$\boldsymbol{\neg p} > p$ .61 > .03	$\boldsymbol{\neg p} > p$ .11 > .00
	EIG of *q* versus *non-q*	Violation of constraint	$\boldsymbol{\neg q} > q$ .61 > .03	$q > \boldsymbol{\neg q}$ .28 > .08	$\boldsymbol{\neg q} > q$ .08 > .00
Exp. 6 replication	$P(p) \leftarrow P(q)$	.20 ← .15	.20 ← .85	.85 ← .20	.85 ← .80
	EIG of *p* versus *non-p*	$p > \boldsymbol{\neg p}$ .28 > .08	Violation of constraint	$\boldsymbol{\neg p} > p$ .11 > .00	$\boldsymbol{\neg p} > p$ .61 > .03
	EIG of *q* versus *non-q*	$\boldsymbol{q} > \neg q$ .61 > .03	Violation of constraint	$\boldsymbol{q} > \neg q$ .08 > .00	$\neg q > \boldsymbol{q}$ .28 > .08
Exp. 7 implication	$P(p) \rightarrow P(q)$	.15 → .20	.15 → .80	.80 → .15	.80 → .85
	EIG of *p* versus *non-p*	$\boldsymbol{p} > \neg p$ .61 > .03	$\boldsymbol{p} > \neg p$ .11 > .00	Violation of constraint	$\neg p > \boldsymbol{p}$ .28 > .08
	EIG of *q* versus *non-q*	$q > \boldsymbol{\neg q}$ .28 > .08	$\boldsymbol{\neg q} > q$ .08 > .02	Violation of constraint	$\boldsymbol{\neg q} > q$ .61 > .03
Exp. 8 exclusion	$P(p) \uparrow P(q)$	.15 ↑.20	.15 ↑ .80	.80 ↑.15	.80 ↑.85
	EIG of *p* versus *non-p*	$\boldsymbol{p} > \neg p$ .11 > .00	$\boldsymbol{p} > \neg p$ .61 > .03	$\neg p > \boldsymbol{p}$ .28 > .08	Violation of constraint
	EIG of *q* versus *non-q*	$\boldsymbol{q} > \neg q$ .08 > .00	$\neg q > \boldsymbol{q}$ .28 > .08	$\boldsymbol{q} > \neg q$ .61 > .03	Violation of constraint

Note: For each connector the used frequencies are shown first. '.20 ∨ .15' represents an *p* OR *q* hypothesis with *P*(*p*) = .20 and *P*(*q*) = .15. Then the *EIG*(*p*) and *EIG*(*non-p*), and the *EIG*(*q*) and *EIG*(*non-q*) values are listed in their relative order. It was shown where the quantitative preconditions for the truth of the hypothesis were violated ('violation of constraint'). The falsificationist predictions are indicated by using bold print.

The Bayesian approach predicts frequency effects. For each dependent variable a variation from the falsificationist prediction is predicted in one out of three conditions that do not violate the preconditions. The frequency effects are only predicted if fixed marginals are assumed.

Example: For the adjunction (inclusive or) falsificationism generally predicts *non-p* selections and *non-q* selections. The Bayesian approach predicts an increase of *p* selections only if there are few *p* cases and many *q* cases. Apart from the used mathematical formalism, this prediction can also be made understandable intuitively. If the independence hypothesis is true, a selection of the seldom *p* card presumably leads to one of the many *q* case on the reverse side. But if the OR hypothesis is true, a *p* card selection would presumably lead to a *non-q* card on the reverse side, since the *q* cards are 'needed' to cover the number of remaining *non-p* cases, in order to secure a *p* or a *q* on at least one side of the cards. Thus, a *p* selection (via Bayes theorem) provides probabilistic information about whether M_D or M_I is true.

In contrast, the potentially falsifying *non-p* card is rendered less informative than in other conditions. If the OR hypothesis were true, the reverse side of a *non-p* card needs to be a *q* card anyway, but if the independence hypothesis is true there will also be a high probability of finding a *q* case on the reverse side. Analogous arguments can be constructed also for the other connectors.

Materials and Procedure of the Experiments

Each participant received a booklet containing four task pages. The order of the tasks was described previously. 16 MSTs were used (4 connectors/experiments × 4 probability conditions). For all MSTs, intentionally similar materials and instructions were used. The participants were told to ask the experimenter if they had problems understanding the task. A general instruction page (cf. pp. 99 f.) was used. It was emphasised that there would be a number of tasks of which each would be concerned with a new situation and a new hypothesis. The tasks should be treated completely independently of each other.

The instructions were in German. The MSTs all bore the headline "Research with Animals on Environmental Toxins". The instruction commenced "You are a scientist investigating the effect of environmental toxins in rats. Measurements have been taken to ascertain whether the concentration of a particular substance in their liver is above a particular threshold."

In the next paragraph the hypothesis was formulated. "This time you aim to test the hypothesis that in a new group of rats (which has been exposed to specific environmental conditions) the following law holds". Depending on the experiment (the connector tested) the rules were formulated:

- *Adjunction*: "*For each rat at least one of the two substances, Editozidum (substance E) OR Karbozidum (substance K), is above the threshold.* (Annotation: It is of no importance to you which of the two substances are above the threshold; you are only interested to know, whether for all rats at least one of the two substances is above threshold.)"[43]
- *Replication*: "*Lithozidum (substance L) is always above threshold, IN CASE Pharozidum (substance P) is above threshold.* (Annotation: Additionally, Lithozidum may occur also without Pharozidum.)"[44]
- *Implication*: "*If Naritozidum (substance N) is above threshold, THEN Teritozidum (substance T) is always above threshold.* (Annotation: Additionally, Teritozidum may occur also without Naritozidum.)“[45]

[43] „*Bei allen Ratten liegt jeweils bei wenigstens einem von zwei Stoffen, Editozidum (Stoff E) ODER Karbozidum (Stoff K), die Messung über dem Schwellenwert.* (Erläuterung: Es ist Ihnen unwichtig, welcher der beiden Stoffe über dem Schwellenwert liegt, Sie wollen nur wissen, ob bei allen Ratten wenigstens einer der beiden Stoffe über dem Wert liegt.)“

[44] „*Lithozidum (Stoff L) liegt dann immer über seinem Schwellwert, FALLS Pharozidum (Stoff P) über seinem Schwellenwert liegt.* (Erläuterung: Außerdem kann Lithozidum auch ohne Pharozidum auftreten).“

- *Exclusion*: "*For NO rat there are simultaneously both substances, Oxalozidum (substance O) and Spirozidum (substance S), above their thresholds*. (Annotation: Individually the substances are allowed to be above their threshold.)"[46]

The subsequent instructions were identical in all 16 conditions and only varied in two respects.

Firstly, the names of the substances differed according to the tested connector. In the following example we will use 'substance X' and 'substance Y' as placeholders.

Secondly, the frequencies within the displayed cards of course varied according to the frequency conditions. Table 31 shows the used frequencies in the four conditions (low-low, low-high, high-low and high-high condition). For the different connectors (⊙) we used two variants for each of the four probability conditions (cf. Table 31). In the following instructions we use '.20 ⊙ .85' ($P(p) = .20$, $P(q) = .85$) as an example. For the eight possible frequency patterns a particular pattern was defined which excludes any optical match between the X and Y card sides (for details see Table 32).

Table 32
Coordinates of the Seldom Cards in the Displays of the X or Non-X Sides and of the Y or Non-Y Sides

$P(X)$ ⊙ $P(Y)$	Coordinates of seldom *X* or *non-X* cards	Coordinates of seldom *Y* or *non-Y* cards
('0.15 ⊙ 0.20')	1,6; 2,2; 2,9	1,4; 1,10; 2,1; 2,6
('0.20 ⊙ 0.15')	1,6; 2;1; 2,5; 2,8	1,2; 1,9; 2,4
('0.15 ⊙ 0.80')	1,2; 2.5, 2.9	1,4; 1.8; 1.10; 2.3
('0.20 ⊙ 0.85')	1,2; 1,4; 2,8, 2,10	1,9; 2,1; 2,5
('0.80 ⊙ 0.15')	1,2; 1,7; 2,5; 2,10	1,3; 1,9; 2,7
('0.85 ⊙ 0.20')	1,3; 1,10; 2,7	1,4; 1,6; 2,1; 2;9
('0.80 ⊙ 0.85')	1,3; 1,8; 1,10; 2,1	1,6; 2,4; 2;9
('0.85 ⊙ 0.80')	1,2; 1;7; 2;5	1,3; 1,6; 2,1; 2;7

Note: The positions in the 2 rows × 10 columns matrix are shown. For instance, '1,7' represents a position in the first row and in the seventh column of the matrix.

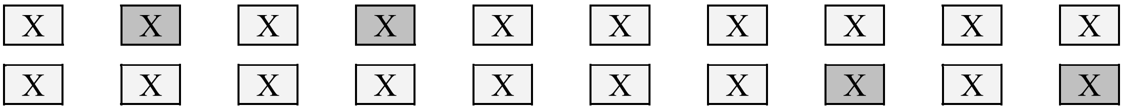
Since WSTs have been found to be very sensitive to instructions, here I also present a complete translation of the remaining instructions (the original material can be obtained from the author):

[45] „*Wenn Naritozidum (Stoff N) über seinem Schwellenwert liegt, DANN liegt immer auch Teritozidum (Stoff T) über seinem Schwellenwert*. (Annotation: Außerdem kann Teritozidum auch ohne Naritozidum auftreten.)"

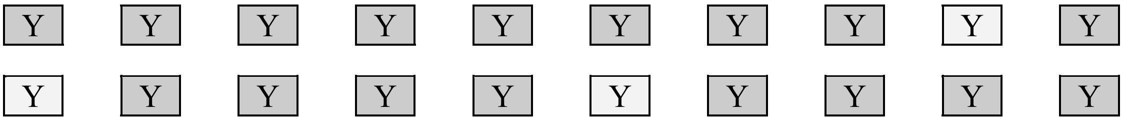
[46] "*Es liegen bei KEINER Rate gleichzeitig BEIDE Stoffe, Oxalozidum (substance O) and Spirozidum über ihren Schwellenwerten*. (Erläuterung: Einzeln dürfen die Stoffe über dem Schwellenwert liegen.)"

"With regard to your hypothesis, you think that both cases are equally possible: The hypothesis may either be true or not true (and both substances occur completely independently from each other).

There is a file on the results of the investigations with the rates. This file is kept by a medical technician. For each of the 20 individuals from the investigated group of rats there is a card. The cards are labelled on two sides. On one side it states whether substance X was above threshold or not and on the other side, whether substance Y was above threshold or not. If a substance X or Y is above threshold, the substance is marked dark (▨), if it is below threshold it is marked pale (□). The technician first displays all 20 cards with the information on substance X facing upwards. You now see how many of the rats in regard of *substance X* were above (dark) or below (bright) the threshold:



As you also aim to see the reverse sides of the cards, with the results for substance Y, the technician quickly puts all cards together and displays them with the other side facing upwards. In this process the cards get *mixed up completely*. *Hence, the new order of cards does not correspond to the previous one*. You now see how many of the rats are above (dark) or below (bright) the threshold in regard of substance Y:



From this you know the frequencies of both card sides, but you do not know their relation. Would you turn over additional cards separately in order to test you hypothesis?

□ Yes, I would turn over single cards. (If you tick this box, please continue the task).

□ No, based on the information about the frequencies I know already that my hypothesis cannot be true anyway.

If you have ticked the 'yes' box, please continue here:

The technician is not willing to turn many more cards separately. He only allows you to turn over one single card with which you may – in a spot check – test the truth or falsity of your hypothesis.

Please select in both continuations of the story one card by encircling it. (Please answer both questions completely independent of each other.)

Given a scenario in which the technician redisplayed the cards with information about substance X, which card would you select to turn over?

Given a scenario in which the technician displayed the cards with the information about substance Y instead, which card would you select to turn over?

If you had to choose between these two possible continuations again, which of the two cases would you select? (Please encircle the preferred choice a second time.) Afterwards you may turn over the page."

Results

Order of Task Presentation

Before analysing the results in detail, it should be checked whether the sequence of task presentation had any major impact on the results. Although the used Latin square design (see p. 146) ensured that each combination of experiment (the four connectors) and probability condition (low-low; low-high, high-low, high-high) was presented equally often in all four positions, here a split-half tests should investigate whether this position had a significant impact.

Firstly, there was no significant connection between the portion of correct answers in regard of understanding the quantitative preconditions of the task and the order of presentation: $r_\varphi = 0.04$ (Pearson test: $\chi^2_{(1,\, n\, =\, 320)} = .07, p = .40$).

Secondly, it should be tested whether the task order had an impact on the selections expected on Bayesian grounds. This test only applies to conditions in which it was normatively correct to continue the task and select cards, and to participants who actually did this. Two tests were conducted to ascertain whether the number of correct Bayesian selections improved from the first half of the tasks to the second. For the task order there was neither a significant correlation for the *p* versus *non-p* selections ($r_\varphi = -.11$, Pearson test: $\chi^2_{(1,\, n\, =\, 194)} = 2.41, p = .12$) nor for the *q* versus *non-q* selections ($r_\varphi = .06$, Pearson test: $\chi^2_{(1,\, n\, =\, 197)} = .90, p = .34$).

Therefore, the task order could be neglected in the further analyses of the four experiments.

Results of Experiment 4 – Adjunction

Quantitative Preconditions. Firstly, the sensitivity of the participants to the quantitative precondition of testing an adjunction (inclusive OR) is investigated.

Table 33
Sensitivity to Quantitative Preconditions of the Adjunction

		Empirical continuation		Total
		No	Yes	
Normative	No	75 % (15)	25 % (5)	20
continuation	Yes	8 % (5)	92 % (55)	60
Total		20	60	80

Note: Normative and empirical answers to the question whether further cards need to be turned over.

Table 33 shows the relation between the normative and the empirical answers to the question whether the given (quantitative) information had violated the preconditions, that is, whether further cards need to be selected to test the rule or not. Only the low-low condition (Condition A: ‘.20 ∨ .15’) the ‘or’ hypothesis is false a priori, because even if *p* and *q* in an optimal way would never occur together, $P(p) + P(q) = .35$, they cannot cover 100 % of the cases. Hence, only in this condition, it is normatively correct to answer that no further selections are needed; in the other conditions participants normatively had to continue the selection task in order to find out what is true.

The results of Table 33 show that the participants were sensitive to these quantitative preconditions (Pearson test: $\chi^2_{(1,\, n = 80)} = 35.36$, $p < .001$; $r_\varphi = .67$).

Main selection tasks. For the three remaining frequency conditions (conditions B, C and D) and for the 55 participants who had correctly continued the selection procedure, Table 34 and Table 35 show the results for the *p* versus *non-p* and the *q* versus *non-q* selections.

Since both questions were formulated in a way that stressed the independence of both questions (see Method), their results are presented separately. A substantial portion of the participants selected the cases that were logically correct (*non-p* and *non-q*). The logically correct cases are also correct from a Bayesian perspective in two of three conditions. However, the predicted differences between the conditions all took the route expected by the Bayesian approach.

Table 34
Percentage and Number of P versus Non-P Card Selections

	(B)	(C)	(D)
	Low-high condition	High-low condition	High-high condition
	.20 ∨ .85	.85 ∨ .20	.85 ∨ .80
P	37 % 7	20 % 4	0 % 0
Non-P	63 % 12	80 % 16	100 % 16
n	19	20	16

Table 35
Percentage and Number of Q versus Non-Q Card Selections

	Low-high condition	High-low condition	High-high condition
	.20 ∨ .85	.85 ∨ .20	.85 ∨ .80
Q	10 % 2	40 % 8	0 % 0
Non-Q	90 % 17	60 % 12	100 % 16
n	19	20	16

As predicted for the *q* versus *non-q* choice both main comparisons, that between low-high condition and high-low condition and that between high-low condition and high-high condition, proved to be statistically significant (B vs. C: $\chi^2_{(n = 39)} = 4.44$, $p < .05$, $r_\varphi = .34$; C vs. D: exact Fisher test $n = 36$, $p < .01$, $r_\varphi = -.48$). As was also expected, no significant difference between the low-high condition and the high-high condition was exhibited (B vs. D: *e*xact Fisher test, $n = 35$, $p = .49$).

Likewise for the *p* versus *non-p* selection, the difference between the low-high probability condition and the high-high condition was highly significant (B vs. D: *e*xact Fisher test, $n = 35$, $p < .01$, $r_\varphi = -.46$). Similarly, expectations were confirmed in that there was no difference between the high-low condition and the high-high condition (C vs. D: *e*xact Fisher test $n = 36$, $p = .11$). Nevertheless, the postulated change between the low-high condition and the high-low condition did not become significant (B vs. C: $\chi^2_{(1,\ n\ =\ 39)} = 1.36$, $p_{\text{one-tailed}} = .12$; $r_\varphi = -.18$). However, on the whole the expected frequency effect between the high-low and the high-high conditions on the one hand, and the low-high condition on the other was reliable (B vs. C D: exact Fisher test, $n = 55$, $p < .05$, $r_\varphi = -.31$).

Additional selection task. Now the additional dependent variable is analysed, which asked which of the two previous selection options would be preferred (either the one resulting from the *p* versus *non-p* choice or that resulting from the *q* versus *non-q* choice). The analysis has to be based on the selections in the two previous tasks. For all patterns of preceding results, it was required to determine whether the actual choice accorded with the Bayesian predictions, which can be derived from the *EIG* values of the previously selected cards shown in Table 31 (p. 148). The index variable got the value one, if the Bayesian predictions were met; it was set zero, if no answer was given; and it was set to minus one, if the selections differed from the one predicted by a Bayesian approach. Of the 55 participants in the three remaining conditions of the adjunctions, 9 participants did not answer, 31 answered according to the Bayesian predictions and 15 contrary to the Bayesian predictions (Table 36). There was a significant effect of the participants who answered this question to choose the option predicted by the Bayesian model (one-dim. $\chi^2_{(1,\ n\ =\ 46)} = 5.57$, $p_{\text{one-tailed}} < .01$).

Table 36
Number of Bayesian Answers in the Choice Between *P* vs. *Non-P* and *Q* vs. *Non-Q*

Type of answer	Low-high condition		High-low condition		High-high condition		Overall	
(-1) Non-Bayesian	21 %	4	40 %	8	19 %	3	27 %	15
(0) No answer	16 %	3	15 %	3	19 %	3	16 %	9
(1) Bayesian	63 %	12	45 %	9	62 %	10	57 %	31

Results of Experiment 5 – Replication

Quantitative Preconditions. For the replication hypothesis, the first question, asking whether one needs to check any cards to test the hypothesis, descriptively showed tendencies in the predicted direction. But the resulting correlation was comparatively low and statistically the effect became only marginally significant (Pearson test: $\chi^2_{(1,\ n\ =\ 80)}$= 2.50, $p_{\text{one-tailed}} = .06$; $r_{\varphi} = .18$). Some written comments made by the participants suggest that the used formulation of the replication, '*p*, in case of *q*' was often interpreted differently, as either a biconditional or a conditional hypothesis. However, even here there were significantly

Table 37
Sensitivity to Quantitative Preconditions of the Replication

		Empirical continuation		Total
		No	Yes	
Normative	No	55 % (11)	45 % (9)	20
continuation	Yes	35 % (21)	65 % (39)	60
Total		32	48	80

Note: Normative and empirical answers to the question whether further cards need to be turned over.

more correct than false selections (50 vs. 30; one-dimensional $\chi^2_{(1,\ n\ =\ 80)} = 5.00$, $p < .05$).

Main selection tasks. For the remaining 39 participants who correctly proceeded to select cards, the results are given in Table 38 and Table 39. Descriptively the results for all predicted differences accorded with the predicted direction again.

Table 38
Percentage and Number of P versus Non-P Card Selections

	(A)	(C)	(D)
	Low-low condition	High-low condition	High-high condition
	.20 ← .15	.85 ← .20	.85 ← .80
P	69 % 11	23 % 3	22 % 2
Non-P	31 % 5	77 % 10	78 % 7
n	16	13	9

Table 39
Percentage and Number of Q versus Non-Q Card Selections

	Low-low condition	High-low condition	High-high condition
	.20 ← .15	.85 ← .20	.85 ← .80
Q	69 % 11	92 % 12	50 % 5
Non-Q	31 % 5	8 % 1	50 % 5
n	16	13	10

For the *p* versus *non-p* comparisons the predicted difference between the low-low condition and the low-high condition was significant (A vs. C: $\chi^2_{(n = 29)} = 5.99$, $p < .05$, $r_\varphi = -.46$). Likewise, this was the case for the expected difference between the low-low condition and the high-high condition (A vs. D: $\chi^2_{(n = 25)} = 5.00$, $p < .05$, $r_\varphi = -.45$). As also expected, there were no differences between the low-low and the high-low conditions (C vs. D: exact Fisher test, $n = 25$, $p = 1.00$).

For the *q* versus *non-q* comparisons only one of the two predicted differences became significant, that between the high-low condition and the high-high condition (C vs. D: exact Fisher test, $n = 23$, $p_{one\text{-}tailed} < .05$, $r_\varphi = -.48$); the difference between the low-low condition and the high-high condition failed to become significant (A vs. D: *e*xact Fisher test, $n = 26$, $p = .58$, $r_\varphi = -.18$). As expected, there was not significant difference between the low-low and the high-low condition (A vs. C: $\chi^2_{(n = 29)} = 2.43$, $p = .12$).

Additional selection task. In the choice between their previous *p* vs. *non-p* selection and their previous *q* versus *non-q* selection more participants descriptively selected cards that were predicted by the Bayesian approach, but here the contrast did not reach significance (one-dimensional $\chi^2_{(1,\, n\,=\,31)} = 1.58, p = .20$).

Table 40
Number of Bayesian Answers in the Choice Between *P* vs. *Non-P* and *Q* vs. *Non-Q*

Type of answer	Low-low condition		High-low condition		High-high condition		Overall	
(-1) Non-Bayesian	31 %	5	46 %	6	10 %	1	31 %	12
(0) No answer	19 %	3	8 %	1	40 %	4	20 %	8
(1) Bayesian	50 %	8	46 %	6	50 %	5	49 %	19

Results of Experiment 6 – Implication

Quantitative Preconditions. The majority of participants were sensitive to the quantitative preconditions for testing an implication (Table 41).

Table 41
Sensitivity to Quantitative Preconditions of the Implication

		Empirical continuation		Total
		No	Yes	
Normative	No	75 % (15)	25 % (5)	20
continuation	Yes	10 % (6)	90 % (54)	60
Total		21	59	80

Note: Normative and empirical answers to the question whether further cards need to be turned over.

For the most part they refrained from testing the hypothesis, if its quantitative preconditions were violated, and continued to test the hypothesis, if the precondition was not violated (Pearson test: $\chi^2_{(1,\, n\,=\,80)} = 32.74, p < .001, r_\varphi = .64$).

Main selection tasks. Table 42 and Table 43 show the results for the *p* versus *non-p* selections and the *q* versus *non-q* selections in the three remaining probability conditions. As with the previous connectors here too all four predicted differences descriptively corresponded to the predicted direction.

Statistically, the results for the *p* versus *non-p* comparison were not significant (A vs. B: exact Fisher test, $n = 33, p = .61, r_\varphi = -.15$; A vs. D: exact Fisher test, $n = 34, p = .10, p_{one\text{-}tailed} = .09, p_{two\text{-}tailed} = .10, r_\varphi = -.31$; in the latter case the alternative test procedure of a likelihood ratio chi-square would be significant). As predicted, there was no significant difference between the low-high condition and the high-high condition (B vs. D: exact Fisher test, $n = 37, p = .45$).

Table 42

Percentage and Number of P versus Non-P Card Selections

	(A) Low-low condition .15 → .20	(B) Low-high condition .15 → .80	(D) High-high condition .80 → .85
P	93 % 14	83 % 15	68 % 13
Non-P	7% 1	17 % 3	32 % 6
n	15	18	19

Table 43

Percentage and Number of Q versus Non-Q Card Selections

	Low-low condition .15 → .20	Low-high condition .15 → .80	High-high condition .80 → .85
Q	73 % 11	44 % 8	35 % 7
Non-Q	27 % 4	55 % 10	65 % 13
n	15	18	20

For the *q* versus *non-q* selections both predicted differences were significant: There was a difference between the low-low condition and the low-high condition (A vs. B: $\chi^2_{(n = 33)} = 2.80$, $p_{one\text{-}tailed} < .05$, $r_\varphi = .29$) and one between the low-low condition and the high-high condition (A vs. D: $\chi^2_{(n = 35)} = 5.04$, $p < .05$, $r_\varphi = -.38$). As also expected, there was no reliable difference between the low-high condition and the high-high condition (B vs. D: $\chi^2_{(n = 38)} = .35$, $p = .55$, $r_\varphi = -.10$).

Additional selection task. The results for the final choice between the two previous selections are presented in Table 44. Overall there were significantly more selections which were in accordance with the Bayesian predictions than with the alternative answers (one-dimensional $\chi^2_{(1,\, n = 45)} = 21.75$, $p < .001$).

Table 44

Number of Bayesian Answers in the Choice Between *P* vs. *Non-P* and *Q* vs. *Non-Q*

Type of answer	Low-low condition	Low-high condition	High-high condition	Overall
(-1) Non-Bayesian	33 % 5	11 % 2	20 % 4	21 % 11
(0) No answer	20 % 3	6 % 1	20 % 4	15 % 8
(1) Bayesian	47 % 7	83 % 15	60 % 12	64 % 34

Results of Experiment 7 – Exclusion

Quantitative Preconditions. The majority of participants refused a further selection of cards if the quantitative preconditions for the validity of the exclusion were violated and went on to the selection procedure, if it was correct to do so (Pearson test: $\chi^2_{(1,\ n = 80)} = 34.19$, $p <$ 0.001; $r_\varphi = .65$).

Table 45
Sensitivity to Quantitative Preconditions of the Exclusion

		Empirical continuation		Total
		No	Yes	
Normative	No	80 % (16)	20 % (4)	20
continuation	Yes	12 % (7)	98 % (53)	60
Total		23	57	80

Note: Normative and empirical answers to the question whether further cards need to be turned over.

Main selection tasks. The result for the *p* versus *non-p* and the *q* versus *non-q* selections are presented in Table 46 and Table 47. Again, the predicted differences in all cases corresponded to the directions of the descriptively found differences.

Table 46
Percentage and Number of P versus Non-P Card Selections

	(A) Low-low condition .15 ↑ .20	(B) Low-high condition .15 ↑ .80	(C) High-low condition .80 ↑ .15
P	100 % 19	88 % 15	60 % 9
Non-P	0 % 0	12 % 1	40 % 6
n	19	16	15

Table 47
Percentage and Number of Q versus Non-Q Card Selections

	Low-low condition .15 ↑ .20	Low-high condition .15 ↑ .80	High-low condition .80 ↑ .15
Q	100 % 19	59 % 10	86 % 12
Non-Q	0 % 0	41 % 7	14 % 2
n	19	17	14

With respect to the *p* versus *non-p* selections both predicted differences were (highly) significant, the low-low versus the high-low condition (A vs. C: exact Fisher test, $n = 34$, $p < .01$, $r_\varphi = -.52$) and the low-high versus the high-low condition (B vs. C: exact Fisher test, $n = 31$, $p < .05$, $r_\varphi = -.40$). In contrast, as also predicted, the low-low versus the low-high condition did not differ reliably (A vs. B: exact Fisher test, $n = 35$, $p = .46$).

With respect to the *q* versus *non-q* selections the postulated difference between the low-low and the low-high condition was significant (A vs. B: exact Fisher test, $n = 36$, $p < .01$, $r_\varphi = -.52$), but the one between the low-high and the high-low condition was not (B vs. C: exact Fisher test, $n = 31$, $p = .13$, $p_{\text{one-tailed}} = .11$ $r_\varphi = .29$). However, if the low-high condition is tested against the low-low and the high-low condition together, the predicted difference remains significant (B vs. AC: exact Fisher test, $n = 50$: $p < .01$, $r_\varphi = .43$). As expected, there was no difference between the low-low and the high-low conditions (A vs. C: exact Fisher test, $n = 33$, $p = .17$, $r_\varphi = -.30$).

Additional selection task. Table 48 shows the findings for the final selection task. Overall the number of selections predicted by the Bayesian model (conditional on the previous selections) is larger than the number of alternative other selections (one-dimensional $\chi^2 = 5.23$, p < .05).

Table 48
Number of Bayesian Answers in the Choice Between *P* vs. *Non-P* and *Q* vs. *Non-Q*

Type of answer	Low-low condition	Low-high condition	High-high condition	Overall
(-1) Non-Bayesian	37 % 7	18 % 3	29 % 4	28 % 14
(0) No answer	5 % 1	23 % 4	14 % 2	14 % 7
(1) Bayesian	58 % 11	59 % 10	57 % 8	58 % 29

Discussion

Bayesian Logic and Further Factors?

Overview. The obtained results of the four experiments provide evidence for the effects predicted on the basis of Bayesian logic of hypothesis testing, which I developed in analogy to the Bayesian approach to testing conditionals (v. Sydow, 2002; Hattori, 2002; Oaksford & Chater, 2003). The predicted differences could have been partly more pronounced, and a relevant part of the answers corresponds to answers which would have to be predicted on falsificationist grounds. However, qualitatively all predicted differences in all four experiments were in accordance with the predicted direction (24 comparisons, 6 per experiment) and most of them became significant. There may well be two processes involved in solutions of the WST (cf. e. g. Evans, 1995), perhaps one activating a falsificatory solution and one activating a Bayesian solution.

In any case, the Bayesian approach appears not only to be limited to conditionals but – to some extent at least – to be also applicable to the other 'Bayesian connectors'. This is a novel finding and not trivial at all. Firstly, even for testing implication hypotheses the Bayesian predictions, as we have seen, are highly controversial (e. g. Oberauer et al., 2004; Chapters 3.2 to 6). Secondly, it would have been possible that the effect of frequency information is limited to the implication, since this effect has not been tested before for other connectors.

Detailed discussion. Now the three dependent variables are discussed separately, also considering the found deviations.

(a) Quantitative Preconditions. Participants were generally sensitive to the quantitative preconditions for testing the truth of the different connectors. Participants mostly refrained from testing further cards if the quantitative information made clear that the different preconditions for the tested Bayesian connectors were violated and they correctly continued for the selection task if this was not the case. Only in the case of the replication rule were some deviations found which can plausibly be explained by the used formulation of the rule (p, in case *q*) which some participants may have misinterpreted as biconditional.

(b) Main selection tasks. Most of the predicted differences in respect of the *p* versus *non-p* and *q* versus *non-q* card choices have between confirmed. The predicted frequency effects obviously play a substantial role for all 'Bayesian connectors'.

However, some aspects of the results have been problematic. Firstly, the frequency effects were only partially confirmed for the implication rule. This will be discussed in detail in small print. In any case, here too all effects followed the predicted direction and if analysed with the analogous cases from the replication rule (understood as reversed implication), all comparisons become significant.

For the implication rule, both comparisons regarding the '*p* versus *non-p'* choice did not become significant, while only the two '*q* versus *non-q'* comparisons did. Likewise, for the replication connector one of the two analogous '*q* versus *non-q*' comparisons did not reach significance, whereas the '*p* versus *non-p'* comparisons did.

Although all the results for these connectors followed the predicted direction and the results are much more confirmative than many previous results testing the Bayesian approach (e. g. Oberauer et al., 2004; cf. Chapter 5 and 6 for details), they are not fully coherent with the Sydow model, but rather with the Oaksford model, which only predicts *q* versus *non-q* effects. In Experiment 3 the Oaksford model and the Sydow model has been dissociated for the first time. But here the marginal probabilities were fixed and hence the normative predictions had to be based on the Sydow model.

One explanation for the partial deviation found for the implication rule may be that the *non-p* cases have not been mentally represented by the participants. An analogous case may be made for the replication. Mental model theory (Johnson-Laird & Byrne, 1991, 2002) may be right in advocating the importance of incomplete representations in testing descriptive hypothesis. But the incomplete model would not be a logical mental model in the sense of mental model theory. Mental model theory cannot account for the frequency effects found and hence, cannot explain the differences of the present experiment to the other fully confirmative Bayesian findings (Experiment 1, 2, 3, cf. Experiment 8, 9). Instead, this contrast, in my view suggests incomplete Bayesian or causal models. For instance, the instruction of the previous Experiment 3 induced a clear causal direction of viruses and symptoms, and, even more important, also an alternative cause. (Cf. the causal model interpretation of Experiment 3, pp. 130 f.) In contrast, in the present experiment the two toxins *p* and *q* were not causally ordered and no clear alternative cause had been specified. Also in the current experiment another hidden cause needs to be assumed, since there are more *q* toxins than *p* toxins ($P(p) > P(q)$). However, the causal interpretation is here relatively unclear and since an alternative cause has not explicitly mentioned, the *non-p* and *non-q* cases may not always have become mentally represented.

However, the results of Experiment 2 make clear that a *causal* representation is obviously not a precondition for complete representation. We have obtained fully confirmative data also with abstract material in Experiment 2 (cf. also v. Sydow, 2002). But in this experiment the alternative *non-p* cases were indeed formulated positively. Hence, it is premature to speculate about the exact conditions under which a normative Bayesian mental model is represented incompletely. But the one found deviation suggests that there perhaps may be the need to develop something like a ‘Bayesian mental model theory’.

Another plausible possibility is that these partial deviations were unsystematic, if they are seen in the context of the other positive results (cf. Experiment 1, 2, 3, cf. Experiment 8, and 9).

In any case, a closer analysis reveals that even the currently discussed problematic results are not that bad (the *p* versus *non-p* comparisons for the implication rule and the analogous *q* versus *non-q* comparisons for the replication rule). Firstly, all expected differences went in the expected direction. Secondly, if one combines the analogous data of the implication and replication rule, the difference also for these problematic difference became significant ($\chi^2_{(1,\, n = 91)} = 5.24, p < .05$; $r_\varphi = .28$).

Hence, if the analogous results for the implication and replication are taken together, thereby increasing *N*, they do support the predicted differences, also in regard of the more problematic comparisons.

Secondly, many logical-falsificationist answers were found as well. This is the case although falsificationism cannot account for any of the found frequency effects. It has to be conceded that both, the postulated Bayesian strategy, but also the falsificationist strategy, substantially contributed to the results.

Over all connectors and frequency conditions the Bayesian and the falsificationist strategy could explain an about equal number of found selections (72 % vs. 74 %). None of the two strategies had a larger impact (Pearson $\chi^2_{(1,\, n = 784)} = 0.53, p = .47, r_\varphi = 0.03$).

For each connector two out of three conditions (and one out of three comparisons) led to the same predictions from a Bayesian and a falsificationist perspective. Only one condition (or two out of three comparisons) dissociated between these norms.

If we look only at these conditions both strategies seem to explain about 50 % of the selections (48 % vs. 52 %; assuming 20 cases for each condition: one-dimensional $\chi^2_{(1,\, n = 160)} = 0.32$, $p = .57$). Also in this perspective both strategies seem to explain an equal portion of the data.

If we would have included the selection tendency correction for the Bayesian model, introduced by Hattori (2002, and Oaksford and Chater, 2003), the Bayesian strategy would look better. Additionally, the logical answers may have benefited from the presentation of the contrast of several logical connectors.

In any case, the novel effects predicted by Bayesian logic were at least about as strong as those predicted by falsificationism.

Additionally, there seems to be a difference between the four tested connectors. Investigating the dissociating conditions only, the Bayesian effects seem to be stronger for the implication and replication connectors and weaker for the adjunction and exclusion connectors.

This difference may perhaps appear reasonable from a causal model perspective (e. g. Waldmann, 1996; Pearl, 1988, cf. 2000), since the adjunction (the inclusive or) and the exclusion are normally not represented in standard causal graphs (cf. Discussion of Chapter 6). But this causal explanation is presumably not needed here.

If we not look at the single conditions, but if we compare the effect sizes, we come to the reversed result. The average effect size for the critical comparisons of the adjunction and the exclusion is r_φ = .40. For the implication and replication it is only r_φ =.34. (If the problematic comparisons discussed earlier would be excluded from analysis, both groups of connectors would have exactly the same average effect size for the critical comparison, r_φ = .40). Hence, the current results cannot substantiate a claim that the four investigated connectors have been treated along different lines.

(c) Additional selection task. The choice between the two cards the participants had chosen in the previous selection tasks, largely confirmed the Bayesian predictions. The found frequency effects concern the selections between all sorts of different previous selections. In 11 out of 12 conditions there were more Bayesian answers than non-Bayesian ones. Of the four global statistical tests, only the one for replication did not become significant. Hence, at least for three connectors the predicted frequency effects were found in this additional task.

(d) Conclusion. In all four experiments, evidence was found in favour of the Bayesian predictions, which were here expounded for the first time for all four connectors. Although not all predicted differences became significant, most did and descriptively all comparisons went in the predicted direction. No comparison became significant which was not predicted to do so. Although there was also a substantial portion of answers which correspond to the falsificationist norm of hypothesis testing, the results provide first evidence in favour of subjects' sensitivity for the role of quantitative information and for a Bayesian logic of information selection. Frequency information had a clear systematic impact on all four investigated connectors.

Other Theories of the WST

Other non-Bayesian psychological theories of the WST cannot explain the found frequency effects in Experiment 4 to 7, which were postulated and tested for the first time here.

Mental logic theory (ML theory, e. g., Braine, 1978; Rips, 1994; O'Brien, 1995) cannot account for any frequency effects in the WST, since the postulated incomplete set of mental rules does not depend on frequencies (cf. pp. 10 f, 113 f., 172 f.)

Mental model theory (MM theory; e. g., Johnson-Laird & Byrne, 1991; see pp. 11, 114, 128, 173, 272), which is based on falsificationism, is coherent with the substantial number of falsificationist answers, but it cannot account for the found frequency effects.

Firstly, the found sensitivity of participants to quantitative preconditions of testing different connectors – also explicable by falsificationism – cannot be suitably be explained by MM theory, since it is a central psychological tenet of that theory that people use a minimum number of models with a single representative of a logical class (e. g. Johnson-Laird & Byrne, 1991, 36).[47]

Secondly, here we are concerned with a purely probabilistic manipulation, not with a manipulation of the general plausibility to find a counterexample in a certain situation (cf. Green & Larking, 1995; Love & Kessler, 1995; Green 1998). MM theory (e.g., Johnson-Laird & Byrne, 1991, 2002) has not made any predictions how purely probabilistic manipulations of $P(p)$ and $P(q)$ may be connected to fleshing out a model. Hence, MM theory cannot account for any of the frequency effects found in Part II.

Moreover, the most plausible proposal, about how one may nonetheless connect these probabilities, leads to predictions, which would run contrary to the empirical findings (cf. Section 5.7, p. 114). From a MM perspective it should be plausible that a high number of *non-p* and *non-q* cards may increase the probability that these cases are fleshed out mentally, leading to a full representation of the logical connector.[48]

Thirdly, the incomplete models of different connectors, postulated by MM theory, do not contribute to explain the results. For instance, for the adjunction MM theory would need to postulate a fleshed out model in all conditions (e. g. Johnson-Laird & Byrne, 1991, p. 44 f.), since the hypothesis was not formulated as open to either mean an as “p or q” (leaving room for the inclusive or exclusive interpretation) but using a “*p* or *q or both*” formulation.

[47] On probabilistic mental model accounts, cf. footnote 34.

[48] If one were to think of other more refined mathematical relations between $P(p)$ and $P(q)$, and *P*(*counterexample*). But, first I cannot think of a reasonable extension which would fit all the data. Let us consider the conditional. One possibility seems to be to use $P(p \ \& \ non\text{-}q \mid M_l)$ as an estimate for *P*(counterexample). But in this case there would be no frequency effects between the used low-low and high-high condition. Actually, the probability of turning a card, *U*, would be $U(non\text{-}q)_{\text{low-low}} = U(non\text{-}q)_{\text{high-high}} > \text{U}(non\text{-}q)_{\text{low-high}}$ which does not correspond to our findings. Secondly, even this mentioned relatively simple formalisation is based on an independence model, and the calculation of the Cartesian product $P(p) \times P(non\text{-}q)$. But such a calculation presupposes a complex (probabilistic) representation of *non-q* cases in order to determine whether *non-q* cases are to be represented. This seems to be at odds with the central idea of MM theory that a connector is only represented by a few single instances of logical classes.
(Although MM theory has assumed that models are built before one actually comes to select any cards, *P*(counterexample) may also be formalised by the probability of thinking of a counterexample in the moment of checking the *non-q* card, $P((p \mid non\text{-}q) \mid \text{M}_\text{l})$. Also such a measure seems to be very different from MM theory and rather reminiscent of a probabilistic-falsificationist approach, like that of Kirby (1994). But in any case, I want to mention that this formalisation, in the case of the conditional would lead to the prediction $U(non\text{-}q)_{\text{low-low}} = U(non\text{-}q)_{\text{low-high}} < U(non\text{-}q)_{\text{low-high}}$, which is in contrast to the findings of $U(non\text{-}q)_{\text{low-low}} < U(non\text{-}q)_{\text{low-high}} = U(non\text{-}q)_{\text{low-high}}$.)

Finally, MM theory does not provide any explanation of the frequency effects found, if people were asked which of the two previously selected cards, they would prefer to select.

For all these reasons, mental model theory cannot explain the frequency effects found in Experiment 4 to 7.

Pragmatic reasoning schema theory (pp. 13 f.) and *social contract theory* (pp. 14 f.) were formulated primarily for deontic or social contexts. Moreover, both have only been formulated for conditionals. Nonetheless, both theories assumed that standard descriptive WSTs generally yield unsystematic selections patterns or *p* & *q* selection patterns (for details cf. Discussion of Experiment 2, p. 115). The confirmation of Bayesian frequency effects for a broader set of connectors provides evidence for a much more systematic Bayesian account of testing descriptive WSTs.

Relevance theory of the WST (see pp. 17 f., cf. 71 f., 177 f., 281 f.; Sperber, Cara & Girotto, 1995; Sperber & Girotto, 2002, 2003) is framed in terms of cognitive effects and effort that do not appear to have any immediate application to the purely probabilistic manipulations in the current experiments (cf. Oaksford & Chater 1995a, 1996, Oaksford et al., 1997, 455). Additionally, relevance theory cannot explain the sensitivity of the participants to the quantitative preconditions of the connectors. Moreover, Relevance theory itself has never been applied to the investigated additional connectors.

Oaksford and Chater (1995a) have proposed that their Bayesian approach provide a more formal relevance measure than postulated by relevance theory. I think this is correct and that this could also be said regarding the extensions proposed in this work. But relevance theory, abstract as it is and without its own explicit account of frequency effects, is not committed to predicting current results and equally well may have predicted opposed results. Hence, relevance theory clearly cannot account for the current findings.

The *matching bias account* (Evans 1972, 1989; cf. Oaksford & Chater, 1994), likewise, cannot explain the found frequency effects for the different connectors (cf. particularly the generalisation of Evans, Legrenzi, Girotto, 1999). The effects cannot be only due to matching selections with cases mentioned in the rule, since atomic propositions were used which were exclusively formulated positively. However, a Bayesian approach and a matching bias approach need not to be mutually exclusive (cf. Yama, 2001, Evans, 2002, Yama, 2002).

If we analysed the four experiments together the matching bias hypothesis can also be assessed using the current evidence. Evaluating the results for all connectors and all conditions together, the material was designed so that a falsificationist approach as well as a Bayesian approach (despite their differences) predicted 50 % selections of positive instances (*p* or *q*), matching with the cases in the rules, and 50 % negative selections (*non-p* or *non-q*). Actually 392 selections were made on the whole (the first two selection questions; all four connectors). Empirically, 55 % of these selections were positive and 45 % negative. Although matching bias seems not to play a less important explanatory role, than the Bayesian approach and the falsificationist approach, the current data indeed shows a ‘residual matching bias effect’.

Overall, the observed frequency effects confirmed that the postulated Bayesian logic of hypothesis testing contributes substantially to the results. The found frequency effects and the sensitivity to quantitative preconditions of different connectors cannot be explained by any other theory of the WST.

8 General Discussion of Part II – Flexible Bayesian Logic Theory of the WST

Part II has been concerned with the testing of descriptive rules and has investigated the proposal of a knowledge-based Bayesian logic of the WST.

In contrast, the traditional yardstick to evaluate selections in the Wason selection task (Wason, 1966, 1968) has been the logical-falsificationist norm for testing a hypothesis (Popper, 1934/1994, 1972, 1996), which remained highly influential till today (e. g., Johnson-Laird & Byrne, 1991; Almor & Sloman, 1996; Moshman & Geil, 1998; Ahn & Graham, 1999; Osman & Laming, 2001; Feeney & Handley, 2000; Stanovich & West, 2000; Handley, Feeney & Harper, 2002; Johnson-Laird & Byrne, 2002).

In Part II the aim was to improve, both theoretically and empirically, on the Bayesian approach of hypothesis testing in the WST developed by Oaksford and Chater (1994, 1998a, cf. 2003). These authors have provided a refined mathematical approach to testing conditionals, which they claim to provide a universally valid account of testing descriptive conditionals. However, most previous experiments failed to provide positive results for this approach (e.g., Oaksford et al., 1999; Oberauer, et al. 1999, 2004).

By building on this mathematically refined approach, the aim was to extend it theoretically by advocating a knowledge-based approach. This approach made it possible to achieve largely positive empirical results. Moreover, it allowed differentiation between different fundamental models of conditionals. It was concluded that the knowledge-based approach is an indispensable extension, which solves important problems of the Bayesian approach. Moreover, a Bayesian logic of hypothesis testing has been proposed here that extends the Bayesian approach to other logical connectors.

In the following summary, I will describe the chain of thought and the main results of Part II. I refrain from discussing the results of other authors, since they have been discussed in detail before (pp. 70 f., 83 f.). After the summary, I will recapitulate why previous non-Bayesian theories of the WST cannot explain the findings of Part II.

8.1 Summary of the Line of Argument and the Results of Part II

Chapter 3 provided the philosophical basis for abandoning falsificationism and also for advocating a knowledge-based modification of the Bayesian account of hypothesis testing in the WST. Hume's problem and Hempel's problem of induction were examined and a proposal for a solution of both problems was outlined. Actually, any Bayesian account, which normatively opposes falsificationism, needs to assume that there is a positive solution to these problems.

In the first part of the third Chapter, 3.1, Hume's problem of induction was investigated. This fundamental problem has largely been ignored by Bayesians, who have been more concerned with Hempel's problem. Falsificationism claims to provide a negative solution to Hume's fundamental problem by confining correct hypothesis testing exclusively to falsifications and discarding any kind of confirmation. In contrast, it has been argued here that not only naïve inductionism but also falsificationism falls pray to Hume's problem of induction. Even Hume himself suggested that the problem to warrant inferences from the past to the future in a rational way (the problem of prediction and of rational empirical science) not only affects confirmative evidence, but also disconfirmatory evidence and even falsificatory evidence. No evidence could reasonably be applied to future instances or analogous instances, without making additional assumptions for instance about the constancy of nature. Even if something has been falsified yesterday, it can be true today. With regard to prediction, falsificationism does not lead any further than confirmation (despite some other asymmetries). For predictions, constancy assumptions have to be made. In Section 3.1, I advocated that these assumptions may be rationally based on knowledge, which is itself based on induction. It is argued that the apparently self-referential use of inductive premises in inductive reasoning need not to be completely circular, as long as the evidence for these additional premises is obtained independently from the inference in question. Nonetheless, such an approach transcends a purely bottom-up idea of induction. From this perspective, any inference from observed to unobserved cases is and has to be knowledge-based. Although knowledge-based inductive inferences are fallible, it is claimed that they are rational as long as the conclusions follow probabilistically from their premises, which are likewise only valid probabilistically. Since knowledge-based induction is neither a process of pure induction nor a process of deduction, here it has been called 'synduction'. According to this synductive proposal for solving Hume's problem of induction, rational knowledge acquisition necessarily needs to combine bottom-up with top-down processes and external-to-internal processes with internal-to-external processes.

In Section 3.2, the standard Bayesian resolution of Hempel's paradox of the ravens is discussed. It is reasoned that it is false to assume that the paradox can be resolved by falsificationism. The standard Bayesian solution to this paradox is interpreted as a synductive approach of using additional knowledge about probabilities, which goes beyond formal logic and falsificationism. Furthermore, it is argued that the standard Bayesian solution (e.g., Alexander, 1958) is essentially incomplete, as long as it ignores structural preconditions. The advocated knowledge-based Bayesian solution of the paradox takes these structural preconditions into account and emphasises synduction with regard to the assumed constancy assumptions as well. Here it was shown that this extended knowledge-based or synductive Bayesian account actually implies more than one resolution of the paradox of the ravens.

The following chapters, Chapter 4 to Chapter 7, were more directly concerned with the Bayesian account of the WST. Other non-Bayesian theories of the WST have been dealt with in Chapter 2. My proposals in the WST debate directly build on the technically refined Bayesian account of Oaksford and Chater.

In Chapter 4 the Bayesian model of Oaksford and Chater (1994, 1998a) was introduced. Additionally, objections to this model were considered (e. g. Evans & Over, 1996; Laming, 1996; Green & Over, 1998; Klauer, 1999), and it was shown that a theoretical objection, which has been formulated most forcefully by Laming (1996), has not yet been ruled out. Laming argued that the universal account of testing conditionals advocated by Oaksford and Chater (1994) is not warranted, since instead one may also construct quite different basic models. Oaksford and Chater's universal model of testing a conditional had been criticised as being merely the result of post hoc data fitting. Although Oaksford and Chater replied to many other objections (e.g., Oaksford & Chater, 1996, 1998b; Oaksford, 1998; Chater & Oaksford, 1999a) they continued to advocate a universal model account with regard to the assumed fundamental constancy assumptions. (This is also true for Oaksford and Chater, 2003.) Hence, I concluded that Laming's basic theoretical objection has not yet been ruled out. Furthermore, the empirical situation of the Bayesian account was reviewed. Although some frequency effects have indeed been found by previous research, it has become apparent that most results were rather negative concerning exact predictions of any specific universal Bayesian model.

In Chapter 5 a knowledge-based Bayesian account was pursued (cf. v. Sydow, 2002) as an alternative to a universal account of testing conditionals (Oaksford and Chater 1994, 1998a; also Hattori, 2002; Oaksford & Chater, 2003; Oaksford & Wakefield, 2003; cf. similarly also Evans, Handley & Over, 2003; Over & Evans, 2003; Over, 2004). The knowledge-based account was aimed at ruling out the objections of Laming (1996; cf. Evans & Over, 1996) both theoretically and empirically by actively introducing the preconditions of the assumed model in the instruction. In Chapter 5 a model with fixed marginal probabilites was investigated, which was first fully elaborated by von Sydow (2002) and about the same time by Hattori (2002) and Oaksford and Chater (2003). Whereas the latter authors advocated this model universally, here it is only advocated for specific preconditions. This may be seen as a synthesis of the more informal demands for more flexible approaches (Evans and Over, 1996; Green, Over & Pyne, 1997; Green & Over, 1998) with the mathematically refined and most discussed approach of Oaksford et al. (1998a, 1999, 2002, 2003a, 2003b; Hattori, 2002; Oberauer, 1999, 2004). In Chapter 5 it is argued in detail that previous results for this model – if it is taken to be universal – were mostly negative (e. g. Oaksford et al. 1999; Oberauer, et al. 1999, 2004). Moreover, all seemingly positive results of previous research mentioned by Oaksford and Chater (2003) were reviewed. Most of these were not explicitly designed to induce the specific constraints of this particular model. By discussing the technical intricacies of these experiments, it was shown that there have been no clearly positive results for the model with fixed marginals so far. Moreover, the partially positive effects found might all be explained by other factors as well, which do not necessitate a Bayesian approach (cf. Osman & Laming, 2001). Oaksford and Wakefield (2003) have argued that the predominance of negative results may be due to the absence of a sequential sampling method. They used such a method in one recent experiment with positive results. As Oberauer et al. (2004) pointed out this experiment was also clearly problematic for reasons I have discussed in detail. Moreover, Oberauer et al. obtained negative results in their own sequential test. In contrast, I have advocated that if the preconditions of the model with fixed marginals are induced more clearly, it should be possible to obtain positive results (cf. v. Sydow, 2005). Chapter 5 was aimed at replicating the positive findings of von Sydow (2002), using other materials and ruling out some possible objections. A many cards selection task (MST) was used in order to introduce the structural preconditions of the model exactly.

In Experiment 1, the Sydow model was tested with raven material, in order to connect the raven paradox debate with the WST debate. The results fully confirmed the Sydow model. This also corroborates the Bayesian resolution of the raven paradox for the first time by a selection task using ravens material.

In Experiment 2, letter-number material from the original task (Wason, 1996) was used in order to show that the predicted frequency effects can also be obtained with this abstract non-causal material.

In contrast to previous results, the obtained positive results corroborated the advocated knowledge-based Bayesian approach. Hence, a successive presentation is neither necessary nor sufficient to achieve positive results. The Sydow model was also tested later, in Experiment 3, 6, 8, and 9. On the whole, *q* versus *non-q* effects were found in all six experiments and *p* versus *non-p* effects were found in at least five out of six experiments. This provides the first clear and reliable confirmation of the model with fixed marginals, based on the exact introduction of the preconditions of the model. It was discussed that these results cannot be explained by any non-Bayesian account of the WST (e. g. pp. 104 f.).

In Chapter 6 the advocated knowledge-based account was tested more directly, by voluntarily introducing *different* structural Bayesian models of testing a conditional in the same context for the first time. All models concern the test of the truth or falsity of a deterministic conditional in a selection task. Apart from the Sydow model (v. Sydow, 2002; Oaksford & Chater, 2003), the Oaksford model (Oaksford and Chater, 1994, 1998), which has actually been given up by its authors (Oaksford & Chater, 2003), and the Laming model (Laming, 1996) were tested. The results corroborated the predicted different frequency effects of at least for the Sydow and the Oaksford model. Additionally, subjects were shown to be able to estimate further aspects of the Bayesian models correctly. The obtained results can only be explained by a knowledge-based approach. Although my modelling approach clearly builds on the work of Oaksford and Chater, it goes beyond their universal approach (Oaksford & Chater, 1998a, 2003). Likewise, the results are problematic for the universal claim that the truth or falsity of a conditional only affects $P(q \mid p)$ (cf. Evans, Handley &

Over, 2003; Over & Evans, 2003; Over, 2004). It was argued that the non-Bayesian theories of the WST cannot explain the findings of Chapter 6.

In Chapter 7 'a Bayesian logic of hypothesis testing' was proposed which extended the concept of a Bayesian test of conditionals to other logical connectors. Previous Bayesian proposals on the WST were limited to the conditional. The quantitative preconditions of the 16 connectors of propositional logic were analysed, showing that only four connectors are true 'Bayesian connectors', for which $P(p)$ and $P(q)$ can be varied independently. In the experiments on these four connectors, participants were sensitive to the quantitative preconditions for testing these connectors. Moreover, despite the presence of many falsificationist answers, the selection tasks provided first evidence for the postulated influence of frequency effects predicted by a more general Bayesian logic of the WST. Again it was argued that the obtained frequency effects cannot be explained by any other theory of the WST.

8.2 Non-Bayesian Theories of the WST

The other theories of the WST, mental logic theory, mental model theory, pragmatic reasoning schema theory, social contract theory and relevance theory, cannot account for the findings obtained in Part II.

Since these theories have been outlined before (pp. 3 f.) and mostly discussed in previous chapters (pp. 112, 127, 172), these discussions will here only be summarised (cf. also the General Discussion of Part III).

Mental Logic Theory of the WST (ML Theory)

ML theory (e. g., Braine, 1978; Rips, 1994; O'Brien, 1995, see pp. 10 f., 113 f.) has tried to explain the first anomaly of the WST (p. 8), the normal lack of *non-q* selections in standard descriptive WSTs by an incomplete inventory of mental reasoning rules (e.g., Rips, 1994). Since it is claimed that, mentally, no direct Modus tollens rule is available (opposed to a mental Modus ponens rule) the logically correct *non-q* selections require a complicated line of argument which leads to errors. Other advocates of ML theory came to refuse to apply this theory to the WST (O'Brien, 1995).

In any case, ML theory has neither predicted any frequency effects nor can it explain them. According to ML theory, the difficulty in solving a WST should not depend on frequencies since the availability of particular mental rules has been defined independently of frequencies. Therefore, ML theory cannot explain any of the frequency effects observed in Experiment 1 to 7 (cf. Experiments 8 and 9 in Part III). Moreover, ML theory is unable to explain the differences between the Bayesian models (Chapter 6) and it is difficult for ML theory to provide an explanation for the participants' sensitivity to quantitative preconditions of connectors (Chapter 7).

Mental Model Theory of the WST (MM Theory)

MM theory (see pp. 11 f., 114 f., 128 f., 163 f., 272 f.; e. g., Johnson-Laird and Byrne, 1991, 1995, 2002; cf. Johnson-Laird, 1983) postulates that people do use a minimum of instances (at most the four logical cases) to test a conditional or another dyadic logical connector (cf. Johnson-Laird & Byrne, 1991, 36).

MM theory explains deviations from the falsificationist norm in standard WSTs (normally *p* or *p* & *q* selections instead of *p* & *non-q* selections) by incomplete or false representations of the logical cases. MM theory has claimed that the absence of correct *non-q* selections is normally based on an incomplete model of a conditional (or biconditional) which has not been fleshed out. Johnson-Laird and Byrne (1991, 80) wrote: "In short, the model theory predicts that people will select the card falsifying the consequent whenever the models are fleshed out with explicit representations of that card". They claimed that any "experimental manipulation that lead reasoners to flesh out their models of conditional, and, in particular, to construct an explicit model of an appropriate counterexample, should enhance performance in the selection task“ (Johnson-Laird & Byrne, 1995, 346).

Some authors have proposed that a generally high subjective probability to find a counterexample in a given scenario should lead to a fleshing-out of a model of the conditional (cf. Green & Larking, 1995; Love & Kessler, 1995; Green 1998). They directly manipulated the assumption to find counterexamples. But this needs to be distinguished from a purely probabilistic manipulation of *P*(*p*) and *P*(*q*).

Mental model theory has problems to explain the results of the experiments of Part II for the following reasons:

(a) MM theory has not provided any predictions as to how purely probabilistic manipulations should be related to the probability of finding a counterexample.[49]

[49] On probabilistic mental model accounts cf. footnote 34.

[50] For similar reasons also other complicated formalisations cannot explain the current results (cf. pp. 114 f., 163 f. for details).

Hence, MM theory cannot explain any of the frequency effects obtained in Experiment 1 to 7 (cf. also Experiment 8 and 9 in the next part).

(b) Even if one allowed for a post hoc extension of MM theory, the most plausible extension would lead to predictions that run contrary to the frequency effects found. In my view, the most plausible link between the probabilities $P(p)$ and $P(q)$ and a fleshing out of the mental model of a conditional, should be – from the perspective of mental model theory – to assume that the incomplete '*p* & *q*' model had a higher probability of being fleshed out if many not represented *non-p* and *non-q* cases were visible. However, this would lead to predictions of more *non-q* selections in the low probability condition and would be in contradiction with the reliable *q* versus *non-q* effects found here.

Also more complicated formalisations of the assumed link between the probabilities ($P(p)$ and $P(q)$) and the general expectation to find a counterexample are, in my opinion, not viable for logical MM theory. One may try to base the expectation to find a counterexample directly on the probability for this case, *p and non-q,* given the hypothesis is false (M_I:): $P(p \ \& \ non\text{-}q \mid M_I)$. But even this relatively direct estimate, would be absurd from a MM perspective, since it presupposes a quantitative representation of many *non-q* cases, in order to establish whether a much simpler model is to be fleshed out. Moreover, also the involved multiplication of the two probabilities $P(p)$ and $P(q)$ to calculate $P(p \ \& \ non\text{-}q \mid M_I)$ seems to require more cognitive recourses than the model to be fleshed out itself. In my opinion, this suggests that any such hypothetical extension of MM theory would not be in the spirit of MM theory. But even if this were not the case, the above link between probabilities and fleshing-out of a model would for most of the present experiments lead to the same predictions for the low or high probability conditions. Hence, in any case also this more complicated estimate cannot explain the findings.[50] Finally, such a modification would clearly go beyond current mental model theory, and cannot be considered in a serious way until it has been explicitly formulated.

(c) The finding of *q* selections and the finding of *non-p* selections for a conditional can only be explained by MM theory by a biconditional understanding. Since the former selection, is found in low probability conditions and the other in the high probability conditions, MM theory would have to assume that biconditional interpretations have dominated all our selections.

Firstly, this is implausible, since in all experiments it was salient that $P(p) < P(q)$. Additionally, in some experiments, for instance Experiment 3 an alternative cause was explicitly mentioned, which was aimed at preventing a biconditional interpretation. Moreover, the additional frequency estimations in Experiment 3a have shown clearly that for the Sydow model – which was mostly investigated here – almost all participants interpreted the conditional correctly (also if they had to assume that the hypothesis were true). Furthermore, Experiment 6 corroborated that subjects are generally sensitive to quantitative preconditions of different logical connectors.

Secondly, even if all experiments were based on a biconditional interpretation, this would not solve the problem of fleshing-out (see the two previous arguments *a* and *b*).

(d) MM theory clearly cannot explain the different selection patterns found for the Sydow and Oaksford model in Experiment 3. MM theory would need to postulate a biconditional interpretation for the Sydow model (and, as we have seen, this is absurd), but it would need to postulate a conditional interpretation for the Oaksford model. Apart from the problem discussed before, there is no reason why one should assume different mental models for these two Bayesian models with the same causal graph.

Even more clearly MM theory cannot account for the confirmed additional implications of the Sydow and the Oakford model, since the resulting difference between the fixity or the increase of $P(q_{res})$ has no counterpart in mental model theory (see Discussion of Experiment 3, pp. 128 f.).

(e) MM theory cannot explain the frequency effects found for the four different connectors in Experiment 4 to 7 and no natural explanation arises from MM theory for the found sensitivity to quantitative preconditions of connectors (pp. 163 f.).

(f) All in all, without denying some merits of mental model theory, it has predicted neither the results obtained here nor can explain them in a post hoc way.

Social Contract Theory (SC Theory)

The SC theory of the WST (see pp. 14 f., 115 f., 264 f.) abandoned any normative approach and instead postulated a domain-specific cheater detection module as a specific unsystematic adaptation to a specific selection pressure. The 'Darwinian algorithm' of cheater detection should account for the content effects of clear-cut logical or illogical selection patterns found in social contracts (Cosmides, 1989; Cosmides & Tooby, 1992, Gigerenzer & Hug, 1992; cf. Cummins, 2000; Fiddick, Cosmides, & Tooby, 2000; Fiddick, 2004). Although SC theory mainly investigated what I have called 'prescriptive conditionals' (cf. General Discussion of Part III), it has not been silent on descriptive tasks.

Advocates of this theory have argued that no "thematic rule that was not a social contract had ever produced a content effect that was both robust and replicable" (Cosmides & Tooby, 1992, p. 183; cf. Cosmides, 1989, pp. 200). Based on this claim, proponents of SC theory used descriptive tasks as control conditions, assuming – and, for their part, finding – that descriptive tasks always lead to a diffuse mix made out of *p* selections, *p* & *q* selection and some further selection patterns (Cosmides, 1989; Cosmides & Tooby, 1992; Gigerenzer & Hug, 1992).

Hence, two aspects of the results obtained in Part II contradict these additional claims of SC theory. Firstly, the systematic and highly reliable *p* versus *non-p* and *q* versus *non-q* frequency effects found in at least five experiments refute the assumption that there are no robust and replicable 'content effects' in descriptive tasks. Secondly, the frequency effects and additional findings were predicted by a normative approach and were highly systematic. This is at odds with their claim that the selections in descriptive tasks are completely irrational. Instead, they seem to a substantial degree to be based on a flexible Bayesian logic. SC theory, of course, can neither account for frequency effects, nor for the different effects of different Bayesian models nor for the quantitative sensitivity of participants for the constraints of different connectors.

Additionally, I also want to point out that the two following experiments in Part III are also concerned with frequency effects. It will be shown that frequency effects can also be obtained for social contracts – at least if they are understood as descriptive tasks (see Experiment 8 and Experiment 9 for details). Opposed to the predictions of SC theory this shows that *non-q* selections in social contracts do not always refer to the activation of a cheater detection module.

Pragmatic Reasoning Schema Theory (PRS Theory)

PRS theory of the WST (pp. 13 f., 116 f., 270 f.) has proposed domain-specific schemas for particular recurring experiences of life. An obligation schema and a permission schema has been elaborated (Cheng & Holyoak, 1985; Cheng, Holyoak, Nisbett, & Oliver, 1986; Politzer & Nguyen-Xuan, 1992; Holyoak & Cheng, 1995a, 1995b). PRS theory exclusively concentrated on these schemas, and hence this theory seems inapplicable to the testing of descriptive hypotheses. But this is not entirely correct, since PRS theory made two claims for descriptive WSTs as well.

Firstly, Cheng and Holyoak (1985, 396) postulated that an "arbitrary rule, being unrelated to typical life experiences, will not reliably evoke any reasoning schema". This comment presumably referred to the unsystematic selection patterns in the abstract letter-number task (Wason, 1966). Hence, the systematic findings obtained with this material (Experiment 2) seens to be problematic for this claim. However, in my opinion, one may also here be concerned with a 'hypothesis testing schema', actually, with a particular hypothesis-testing schema corresponding to the Sydow model.

Secondly, and more importantly, Cheng and Holyoak (1989, 306) indeed briefly suggested that there may also be schemas for causality and covariance. Without elaborating this point, they claimed that in selection tasks both should always lead to *p* and *q* patterns (the predominant pattern for such tasks in the literature at that time). Although Cheng (1997, Novick, L. R., & Cheng, P. W., 2004) has made decisive contributions to the related field of causal learning, Cheng and Holyoak, to my knowledge, have never altered their predictions concerning descriptive WSTs. Hence, in regard to this prediction PRS theory is at odds with all frequency effects found in the Experiments 1 to 9.

Matching Bias Heuristics

The matching bias heuristics, first postulated by Evans (1972; 1989; cf. Oaksford & Chater, 1994; Evans, 2002) was not explicitly tested here, since we were only concerned with positive conditionals, without varying the use of negations. Matching bias cannot explain effects of pure probability manipulation and hence cannot explain any of the frequency effects in Part II (and Part III). However, a Bayesian approach and a matching bias approach need not to be mutually exclusive (cf. Yama, 2001).

Moreover, the high base-rates of *p* and *q* selections in some experiments may well refer to a residual matching bias. When testing the four different Bayesian connectors it was possible to investigate this question (cf. p. 165). This analysis used symmetries in the predictions for the connectors. A small matching bias was found to superimpose the other effects. Although Evans, Legrenzi, and Girotto (1999) have extended matching bias to some other connectors, apart from conditionals, they have not shown before that this concept also may be extended to the set of connectors investigated here. But it should be noted that the found effect only warrants a 'residual matching bias', superimposing other effects with higher effect sizes. In this view, the matching bias heuristics do not seem to be the main explanation of the *p* and *q* selections (often found in testing conditionals), as originally claimed, but only one additional factor superimposing the Bayesian and logical explanations.

Relevance Theory

Relevance theory (see pp. 17 f., 71 f., 116 f., 281 f.; Sperber, Cara & Girotto, 1995; Sperber & Girotto, 2002, 2003; cf. also Sperber & Wilson, 1986; cf. Evans, 1994) has advocated that those cards are selected in a WST which appear to be relevant. However, this would also be advocated for instance by any Bayesian approach. The crucial question is, of course, what does it mean to be relevant? Relevance theory provides a quite general answer to this question. Cards are relevant, if they are connected with a high 'cognitive effect' and a low 'cognitive effort'. In my view, such an account is close to a tautology, as long as 'cognitive effect' and 'cognitive effort' are not defined more precisely (cf. Discussion Part III).

Actually, Oaksford and Chater (1995a) have proposed that their information gain approach provides a more formal relevance measure and they argued that the results of Sperber et al. (1995) may have been due to implicitly manipulating $P(p)$ and $P(q)$ (cf. pp. 71 f.). In my view, this may explain some of their original results, but I think not all of them. Relevance theory and the Bayesian approach may well be complementary (cf. Oaksford & Chater, 1996; Almor & Sloman 1996; Hardman, 1998). More recently Sperber and Girotto (2002, 2003) made an important contribution in regard of what I call focusing (without linking this to deontic logic; cf. Part III).

However, relevance theory, framed in terms of cognitive effect and effort, does not have any immediate predictions for purely probabilistic manipulations (cf. e. g. Oaksford et al. 1997) which would be independent from a Bayesian information gain approach. For the manipulations of the present experiments, relevance theory may equally well have 'predicted' this or opposed results and hence it cannot be said to explain any of the *q* versus *non-q* and *p* versus *non-p* frequency effects found here. Moreover, relevance theory cannot explain the sensitivity to quantitative preconditions (Experiment 4 to 7) or the sensitivity to different Bayesian models and their different additional quantitative implications (Experiment 3).

Hence, relevance theory cannot explain the findings of Part II. Moreover, although relevance theory is formulated so general, that it is difficult to think of any finding that would strictly contradict relevance theory (auxiliary hypotheses may completely alter the predictions of this theory), the obtained systematic results, predicted on a rational basis of a knowledge-based Bayesian logic, can be said to be at odds with any relevance theory which neglects the postulated explanatory factors.

Conclusion

The obtained results of Part II provide positive support for a knowledge-based Bayesian logic of hypothesis testing in the WST. Largely positive results were achieved; concerning frequency effects (Experiment 1 to 8, cf. Experiment 9 and 10) in particular, but also regarding other dependent variables (Experiment 3 and Experiments 4 to 7). These positive findings are in contrast to the problematic and incoherent empirical findings of previous studies.

However, also some deviations from the Bayesian predictions were found, for instance, the Laming model in Experiment 3 could not been corroborated. It was suggested that this may be due to the fact that the corresponding causal model is more complicated than the other models tested. Additionally, falsificationist answers still had a relevant influence and a (small) residual matching bias was also confirmed. More research is needed to check to what extent the corroborated knowledge-based Bayesian account can be generalised.

However, the current results support the Bayesian approach more clearly than any other previous study. They provide evidence for a knowledge-based Bayesian logic of the WST and favour this account over all other psychological theories of the WST.

Part III Towards a Flexible Deontic Logic of Testing Prescriptive Rules

After having investigated central problems of selecting information for testing descriptive rules (e.g., laws of nature) in Part II, we now turn to the testing of prescriptive rules (e.g., moral commandments or laws of a legal system). In Part II, a flexible Bayesian logic (FBL) theory of testing descriptive rules has been proposed theoretically and tested empirically. In the present part, a *flexible deontic logic* (FDL) theory of checking prescripts will be proposed and tested.[51] Both theories on the one hand refer to rational and – in a broad sense – even logical norms of reasoning; on the other hand, they refer to the use of domain-specific knowledge.

The advocated FBL theory on testing descriptive rules seems to provide an explanation for the first anomaly of the WST, the dominance of seemingly irrational *p* & *q* selection patterns in standard (descriptive) WSTs (cf. Section 1.3). However, FBL theory does not account for the second anomaly, according to which particular content effects elicit clear-cut selections patterns in obligations and permissions (either 'correct' *p* & *non-q* or 'incorrect' *non-p* and *q* patterns). These content effects have led to the development of domain-specific theories of the WST (for details see 2.3, 2.4). In the current part, FDL theory will be exposed in order to explain also the second anomaly of the WST. FDL theory is presented as an alternative to the proposed domain-specific theories but also to domain-general theories such as mental model theory. FDL theory is domain-specific but nonetheless advocates a more rational and systematic explanation of deontic content effects.

Research on *content effects in the WST* that was linked to obligations, permissions and social contracts has been outlined in Chapter 1.1. Before presenting FDL theory, only a few main aspects of this debate will be recalled.

Research in the 1970s and 1980s showed that not all thematic rules elicit logical facilitation effects, but only rules with a specific content. Certain thematic rules, such as "If I eat haddock, I drink gin" (Manktelow & Evans, 1979) did not enhance the

[51] Aspects of this Part II have been published in von Sydow (2005a), von Sydow, Hagmayer, Metzner, & Waldmann (2005), von Sydow & Hagmayer (2006) and von Sydow (submitted).

performance in testing conditionals, while rules like “If a person is drinking beer, then the person must be of full age” did (Johnson-Laird, Legrenzi & Legrenzi, 1972; Griggs & Cox, 1982; but cf. also Wason & Shapiro 1971). Pragmatic reasoning schema theory and evolutionary social contract theory advocated domain-specific explanations for these findings.

According to Cheng and Holyoak’s (1985, 1989; Holyoak & Cheng, 1995b) *pragmatic reasoning schema theory* (PRS theory, for details cf. 2.3), content effects are due to specific reasoning schemas. Advocates of this theory have argued that these schemas do not always enhance the selection of logical patterns for conditionals; they may in principle also trigger illogical ones. For the social realm, PRS theory proposed specific production rules only for a permission schema and for an obligation schema (Cheng & Holyoak, 1985; Cheng, Holyoak, Nisbett, & Oliver, 1986; cf. Politzer & Nguyen-Xuan, 1992; Holyoak & Cheng, 1995).

Cosmides and colleagues (Cosmides, 1989; Cosmides & Tooby, 1992; Gigerenzer & Hug, 1992; cf. Cummins, 2000; Fiddick, Cosmides, & Tooby, 2000) proposed an evolutionary *social contract theory* (cf. 2.4), advocating a domain-specific, innate and modular mechanism of cheater detection.

FDL theory will build on the insights of PRS theory and SC theory but will go beyond these approaches by advocating a system of deontic logic and a mechanism of a flexible rational focus based on the goals of cheater and cooperator detection. Differences and similarities to former approaches will be discussed in detail in Section 9.4.

Outline of Part III. The following part consists of a theoretical chapter on FDL theory, three chapters which are mainly empirical, and a chapter with a detailed general discussion.

Chapter 9 expounds the proposal of FDL theory theoretically. FDL theory postulates that the standard testing of prescriptive rules combines deontic logic with goal based focus effects. FDL theory is presented as a synthesis of converging lines of research, integrating aspects both of domain-general and domain-specific approaches. FDL theory is claimed to explain content effects which have bedevilled the WST debate (and hence the rationality debate) for long.

In the Chapters 10 to 12, the three main predictions of FDL theory will be elaborated and investigated in five experiments. In Chapter 10, different standard test strategies for either descriptive or prescriptive rules are investigated by analysing the differential effects of frequency information. Here a refined classification of frequency effects for prescriptive rules is elaborated. In Chapter 11, the advocated deontic logic of the WST is examined, distinguishing four types of deontic conditionals. In Chapter 12, the postulated flexible focus mechanism is investigated in two experiments on the goals of cheater and cooperator detection. Also novel double focus effects are tested, allowing us to dissociate FDL theory from other theories of the WST.

Chapter 13 concludes with a general discussion of the results of Part III. FDL theory is evaluated and some aspects of FDL theory, which need further investigation, are mentioned. Moreover, all other relevant theories of the WST are discussed in detail in the light obtained results.

9 Flexible Deontic Logic Theory (FDL-Theory)

In this chapter, FDL theory is proposed as a rational but domain-specific account of testing prescriptive rules.

Firstly, it is argued that the testing of conditionals has different meanings in descriptive and prescriptive tasks. The normal test strategies in these tasks should lead to dissociation in the treatment of frequency information. Secondly, deontic logic is advocated as a rational basis for testing prescriptive rules. Thirdly, it is shown how the goal of the task, which is partly determined by domain-specific knowledge, may flexibly but rationally, determine the focus on particular cells of an ought table. Finally, FDL theory is presented as a theoretical and empirical synthesis of converging lines of research (cf. v. Sydow, 2005, v. Sydow et. al., 2005, 2006).

9.1 The Meaning of Testing Rules about Is or Ought

Although content effects have been observed particularly for prescriptive rules, only few researchers concerned with the WST have recognized the important *normative* differences between testing prescriptive and descriptive rules.

The distinction of ‘is’ and ‘ought’ is fundamental. In philosophy, for example, the distinction of ‘is’ and ‘ought’ is older than Aristotelian formal logic and throughout history has been mirrored by basic philosophical dichotomies like ontology and ethics, or theoretical and practical philosophy. Despite all the differences between philosophical schools the distinction of a prescriptive and a descriptive realm is part of a *philosophia perennis:* it has been made by almost all great philosophers, including Plato, Aristotle, Augustine, Hume, Kant and by many modern analytic philosophers (e.g., Moore, von Wright).

In accordance, descriptive and prescriptive rules have a fundamentally different meaning. Descriptive rules describe states of the world (i.e., facts) and can therefore be true or false. In contrast, prescriptive rules state what *should* be done or omitted; they often state what is right or wrong.

Prescriptive rules are central for moral, social and religious regulation systems in groups or societies. Among the earliest written documents of human thought and behaviour are legal or religious codices (e.g., The Code of Hammurabi or The Books of Moses). Moreover, languages have words or grammatical structures to express prescriptive concerns. There are psychological (even emotional), social and legal regulatory systems which deal with the violation or the fulfilling of ethical, social or legal rules. Although their underlying unity has mostly been ignored (Cheng & Holyoak, 1985) and although there may be indeed for instance differences in the emotional reactions to violations of different kinds of prescriptive rules (e.g., Beller & Bender, 2004; Fiddick, 2004), prescriptive rules are here understood quite generally. They could be moral, social, legal or religious rules, regulations or precepts. The rules can be explicit (e.g., codified laws, contracts, commands, implications) or implicit (e.g., ethical or moral principles, social conventions, codes of behaviour); they may be concerned with the common good of a society, with the good of two or more parties or with the individual good. Hence, also prudential rules in principle may be understood as prescriptive rules – as long as they state how one *ought* to behave. In principle, even the rules of games or technical standards often are to be understood as prescriptive rules. Interestingly, Piaget (1932) actually commenced his historic essay on the moral development of children discussing their rules when playing with marbles.

Corresponding to the differences between descriptive and prescriptive rules and to those between theoretical and practical reason, fundamental differences also exist in testing these rules. The testing of descriptive rules is normally connected to questions of explanation or prediction and to the truth or falsity of a rule, whereas the testing of prescriptive rules is connected to questions of value, power and ethics and to the violation or compliance with a rule or regulation.

Three aspects of the dichotomy of descriptive and prescriptive rules can be distinguished:

Firstly, prescriptive rules cannot directly be falsified by conflicting evidence. In the context of the WST, this was pointed out by Manktelow and Over (1991, 1995). If many people violate a law, like 'If one is drunk, one is forbidden to drive a car on public roads', it is still illegal to do so.

Secondly, and more generally, what is true is not necessarily right and what is false is not necessarily wrong (and vice versa). To represent this distinction, here deontic ought tables are introduced in order to complement descriptive truth tables. Ought tables specify what is right or wrong, allowed or forbidden. A truth table has to be distinguished from an ought table. Not only falsification but also a confirmation of a descriptive rule has no (direct) implications for the corresponding prescriptive proposition. If literally nobody offers one's seat on a crowded bus to an elderly person, this does of course not entail that it is morally or legally forbidden to do so. Vice versa, prescriptive propositions do also not (directly) imply the truth of a corresponding descriptive proposition. The moral and legal forbiddance of murder does not entail the empirical law of nature that no murder takes place at all.

Thirdly, and of most importance here, according to FDL theory these two types of rules are normally tested in different ways (cf. Oaksford & Chater, 1994; even made a similar distinction, but did not elaborate it, cf. Chapter 10). Descriptive conditionals should be tested according to the norms of Bayesian reasoning (Oaksford & Chater, 1994, 1998, 2003; v. Sydow, 2004; cf. e.g.: Evans & Over, 1996; Green, Over, & Pyne, 1997; Klauer, 1999). Not only the logical form but also frequency information is to be taken into account to optimize expected information gain. This has been elaborated in detail in Part II. The Bayesian testing of conditionals determines the important cells of a truth table mainly *a posteriori*, based on frequency information and resulting expected information gain.[52] In contrast, the testing of prescriptive rules is typically concerned with an *a priori* focus on specific cells of an ought table. The typical interpretation of a deontic WST is that the tester should "find out those" or "select those" who either act in accordance or in discordance with that rule.

Frequency information does play an important role for the testing of descriptive rules (cf. Part II). Given a clear and a priori detection goal in testing prescriptive rules,

[52] Also tests of descriptive rules may restrict relevant cells (cf. Chapter 6; v. Sydow, 2004).

there should be not the same kind of frequency effects. Frequency effects should be absent for prescriptive rules.

But this bold statement needs to be confined in two respects: Firstly, there may be some kinds of rational frequency effects also for prescriptive rules. For instance, there may be deontic *within focus frequency effects* (cf. Manktelow, Sutherland & Over, 1995), which should affect only the cards that refer to the focused case. Moreover, frequency information should be irrelevant for *standard* prescriptive WSTs only and there may be tasks in which probability determines which cells are to be focused in the first place. These different kinds of probability effects, which we will be dissociated from the standard frequency effects in descriptive tasks, are discussed in detail in Chapter 10.

Secondly, the dichotomy of the testing of descriptive and prescriptive rules does not need to be bound to the stronger claim that there are no factors or mechanisms common to both types of testing (cf. Manktelow & Fairley, 2000, Sperber & Girotto, 2002, 2003). Besides a common logical basis, another level of analyses may show that factors, like causal knowledge (e.g., Waldmann, 1996; Krynski & Tenenbaum, 2004) or decision theoretic considerations (e.g., Manktelow & Over, 1992, 1995) may well play a role in both realms.

However, fundamental difference in the normal testing of prescriptive rules and of descriptive rules is postulated: In Bayesian testing of descriptive rules the selection patterns result *a posteriori* from their different information gain values. In contrast, the testing of prescriptive rules normally *a priori* focuses on particular cells.

The distinction may explain major phenomena in the WST debate. It would be consistent with the Bayesian approach for testing descriptive WSTs (cf. Part II), and it would explain why clear-cut selection patterns are more often found for prescriptive rules. Moreover, this distinction may explain why Cosmides (1989, Exp. 1, 2; cf. Gigerenzer & Hug, 1992, Exp. 1) found a difference between conditions in which an anthropologist tested for the truth or falsity of a social rule and cheater detection conditions in which participants were concerned with detecting cheaters. She only found clear-cut selection patterns in the cheater detection conditions. This difference does not need to point to a cheater detection module and to an inability to test descriptive rules, but it may point to rational ways to solve two quite different kinds of tasks. (In Chapter 10 and in the General Discussion of this part former results and other theories are discussed in more detail.)

In conclusion, it is advocated that in testing prescriptive rules we are normally quite literally concerned with a *selection* task: The standard instruction of testing prescriptive rules has been to 'select all cheaters' (particularly in the tasks of Cosmides, 1989, and Cosmides & Tooby, 1992). A construction of a similar task for the testing of descriptive rules would result in the rather unusual instruction to 'collect all cases where the conjunction *p* and *non-q* occurred' (cf. Wason, 1968; Sperber & Girotto, 2002, 2003). Although the ability to focus on different cells of a truth-table or an ought-table may well be based on a *general* rational process, found as well in the descriptive as the prescriptive realm, there are essential differences in how descriptive and prescriptive rules are to be tested in standard contexts. Here we will investigate for the first time the postulated difference of these two task types with respect to potential frequency effects.

9.2 Deontic Logic Explains Deontic Content Effects

FDL theory claims that domain-specific content effects, which are the basis for the illogical domain-specific approaches, can be explained and systematized based on deontic logic.

Deontic logic for long time has analyzed logical relations between prescriptive propositions, such as obligations, prohibitions and permissions. The term deontic is derived from ancient Greek *dein*, meaning, roughly, 'to oblige' or 'ought'. Hence, deontic logic is a logic of ought sentences or more generally the logic of prescriptive sentences. First steps of a deontic logic date back to Aristotle, Leibniz, Bentham or Husserl. Based on modern logic Ernst Mally in *Grundgesetze des Sollens* (1926) has first developed an axiomatic system of deontic logic. But it turned out that his axiomatization entailed that a proposition (*p*) is true if and only if it is obligatory (*O*), or in symbols: Ox $\Leftrightarrow x$. This is, of course, counterintuitive. Only the improved axiomatic system of the Finnish philosopher Georg Henrik von Wright, published 1951 in *Mind*, led to a broad acceptance of deontic logic. Nonetheless, there have been some fruitful analogies to (alethic) modal logic, the logic of possibility and necessity. But it has become widely accepted that deontic logic is not identical with (alethic) modal logic (cf. Hilpinen, 1981; cf. here Chapter 13). Also today deontic logic is much more controversially discussed than propositional logic, but it is obvious that it has many insights to offer (see e.g., Hilpinen, 1970, 1981; Nortmann, 1989; von Wright, 1981, 1994; cf. Iwin, 1975).

It is remarkable that the psychological WST debate and the philosophical or logical debate on deontic logic remained isolated for so long. Although the phenomena connected to deontic reasoning have received much attention in the WST debate, deontic logic has been only exceptionally mentioned (Gigerenzer & Hug, 1992, sic; Lowe, 1993; Manktelow & Over, 1995; and see particularly the conference paper of Beller, 2001.) For the connected but different context of deontic syllogistic reasoning Buciarelli and Johnson-Laird (2005) lately elaborated a full proposal based on deontic logic (on deontic conditional *inference,* cf. Quelhas & Byrne, 2003, and Johnson-Laird & Beller, 2003).

Building on insights of deontic logic, it is proposed here that the logical structure of a prescriptive rule is represented by an ought table. One reason for distinguishing ‘is tables’ from ‘ought tables’ is that this allows us to distinguish representations and claims about the is level and the ought level. If a claim is true or false on the is level, it need not to be right or wrong on the ought level and vice versa. Ought tables can be understood as complete representations, at least on the computational level, for the prescriptive aspects of a deontic sentence or a deontic situation. The cells of an ought table represent states of affairs or actions which can either be right or wrong, allowed or forbidden.

A *universal obligation*, for example, like “Thou shalt love thy neighbour as thyself” implies that it is always right to love one’s neighbour and always wrong to hate her/him.

In this work we will be concerned with two-valued binary 2 × 2-ought tables, based on the ought values ‘allowed’ or ‘forbidden’. In principle one may construct 16 such ought tables analogous to the 16 basic truth tables for binary two-valued propositional logic. 2 × 2-ought tables are a complete and structured representation of all possible allowed or forbidden combinations of states of the affaires (or actions), p and q.

More specifically, here we are concentrating on four binary deontic connectives with one forbidden cell. It is proposed here that these four deontic connectives can be represented by four different types of deontic conditionals: conditional obligation, conditional prohibition, conditional permission and conditional permission to refrain.

As mentioned, Bucciarelli and Johnson-Laird, 2005, made a proposal in the context of deontic syllogistic reasoning, which made use exactly of these four conditionals. However, they did not applying this to the WST (See p. 275 for details).

With regard to the WST debate, this transcends earlier proposals in not using explicit negations and modals to formulate for instance a prohibition (cf. particularly Beller, 2001). Obligations and prohibitions are here treated as conditionals with the same basic deontic status. This deontic semantics of four different types of conditionals also goes beyond the traditional interpretation of conditionals as material implications (even in the very liberal version of Johnson-Laird & Byrne, 2002).

A *conditional obligation* "if *p* then one ought to do *q*" asserts that it is wrong (and not false) if *p & non-q* happens. Assume that a tribal rule says, "if you are a bachelor, you must bring fish to the medicine man," then it is forbidden to be a bachelor and not to bring fish – even though there might be actual cases of bachelors who don't bring fish (cf. Table 49).[53]

Table 49
Ought Table of a Conditional Obligation

	Brings fish (*q*)	Does not bring fish (*non-q*)
Bachelor (*p*)	Allowed	Forbidden
Husband (*non-p*)	Allowed	Allowed

A *conditional prohibition*, such as "if you are a bachelor, you are forbidden from going to the bath house", prohibits that one is a bachelor (*p*) and one goes to the bathhouse (*q*). Without additional knowledge all other cases are assumed to be allowed (cf. Table 50).

Table 50
Ought Table of a Conditional Prohibition

	Goes to the bath house (*q*)	Does not go to the bath house (*non-q*)
Bachelor (*p*)	Forbidden	Allowed
Husband (*non-p*)	Allowed	Allowed

A *conditional permission,* "if you are a bachelor, you are allowed to eat the aphrodisiac Cassava root" states that the case to be a bachelor (*p*) and to eat cassava root (*q*) is allowed. Furthermore, this conditional implicitly implies that for husbands (*non-p*) it is forbidden to eat cassava root – otherwise there would be no need for any allowance (cf. Table 51).

Table 51
Ought Table of a Conditional Permission

	Eats Cassava root (*q*)	Does not eat Cassava root (*non-q*)
Bachelor (*p*)	Allowed	Allowed
Husband (*non-p*)	Forbidden	Allowed

[53] There has been some confusion about the terminology of obligation or permission schemas. Cf. Cheng & Holyoak (1985, 1995), Manktelow & Over (1990).

A *conditional permission to refrain*, such as “if you are a bachelor, you may refrain from hunting the dangerous Karogi oxen” determines the allowance to be a bachelor and not to take part in the communal activity of hunting the dangerous Karogi oxen. Implicitly this conditional also conveys the prescription that non-bachelors, namely married man, are obliged to take part in hunting Karogi oxen (cf. Table 52).

Table 52
Ought Table of a Conditional Permission to Refrain

	Hunts Karogi oxen (*q*)	Does not hunt Karogi oxen (*non-q*)
Bachelor (*p*)	Allowed	Allowed
Husband (*non-p*)	Allowed	Forbidden

Whereas conditionals are formalized by propositional logic only by one truth table, the table of the material implication (cf. Table 1), we saw that a system of four different formalizations with one forbidden cell can be derived on the basis of deontic logic. This will be highly relevant for the WST debate. The representation in ought tables provides a more general systematics of different types of deontic conditionals, than the specific production rules assumed by pragmatic reasoning schema theory.

Ought tables and the proposed semantics of four types of deontic conditionals allow us to draw different conclusions from the same representational basis: Let us assume the following deontic scenario in a Thai Buddhist temple district, containing a temple and a garden. Let us assume, the deontic status of the following combinations of acts would be known to us: One is forbidden from entering the temple, wearing shoes, $p \wedge q$. One is allowed to enter a temple barefoot, $p \wedge \neg q$. Finally, one is also allowed to enter the garden with, $\neg p \wedge q$, or without shoes, $\neg p \wedge \neg q$. From the resulting ought table all four proposed deontic conditionals can be derived. For instance, we may derive the conditional prohibition “if one is wearing shoes, one is forbidden from entering the temple” and the reversed conditional prohibition “if one is entering the temple, one is forbidden from wearing shoes.” If the temple district is only made of the temple and the garden (assuming a 2 × 2 ought table), and one has to enter the district anyway, the following deontic conditionals can also be derived: the conditional obligation “if one is wearing shoes, one is obliged to go into the garden” and even the permission to refrain “if one is barefoot, one is allowed to refrain from going into that garden”.

In a WST, the construction of an ought table is partly based on the deontic sentence to be tested. There are syntactical cues for such constructions, like the modal verb ‘may’ indicating a permission rule. However, the word ‘may’ can also indicate a possibility instead of a permission. Moreover, other formulations from the prescriptive realm, like ‘is allowed to’, can replace a ‘may’. Hence, human deontic logic seems to account for the semantics of words like “is forbidden”, “is obliged” or “is allowed” in constructing ought tables.

However, in order to classify a sentence as a deontic proposition, even these deontic words are not always needed. They are needed neither to indicate a prescriptive meaning of a sentence nor to indicate a *particular* ought table. Let us imagine that two parties agree on the sentence, “If I give you $20, you give me your watch”. Although the sentence does not contain any syntactic cues for being prescriptive, the semantics of the sentence and the pragmatics of a situation in which two parties agree upon this sentence normally imply a resulting prescriptive status. Moreover, the semantics and pragmatics of the if-then clause allows an interpretation not as a conditional obligation, but as a biconditional obligation, with duties for both sides (*quid pro quo*) and two forbidden cells in an ought table (Manktelow & Over, 1991; Johnson-Laird & Byrne, 1991, 1992, 1995; Politzer & Nguyen-Xuan, 1992; Oaksford & Chater, 1994; Almor & Sloman, 2000; Beller & Spada, 2003; cf. also 9.3 or Chapter 13).

For deontic conditionals, syntactic, semantic and pragmatic knowledge conveys which cases are forbidden and which are allowed. Here it has been claimed that the four mentioned kinds of conditionals normally induce the four described ought tables, each with a different forbidden cell. Of course, this can only be assumed if no additional premises, hidden in the semantics or pragmatics of the situation, come into play. As discussed, this may lead to changes in the values of the 2 x 2 ought table.

Semantic or pragmatic prior knowledge may add premises to the syntactically stated premises. If we take semantic knowledge into account the descriptive conditional “if York is in Paris, then he visits the Louvre” is correctly interpreted in a biconditional sense. Firstly, ‘York’ is here interpreted as the name of a person and not a name of a city. This interpretation is based on semantic knowledge and the personal pronoun ‘he’ in the subordinate clause. Secondly, and more importantly, geographic knowledge provides the additional premise that “if he is in the Louvre, he is in Paris”. Hence, the additional use of semantic knowledge results in a biconditional interpretation of the conditional. This usage of additional semantic premises is presumably valid for the construction of both of deontic rules and o descriptive rules (cf. the double source approach of Beller, 1997, 2003, Beller & Spada, 1998, 2003).

But additional knowledge may also extend a 2 × 2 ought table. If the idea of prior knowledge is applied to the above *conditional permission to refrain*, "if you are a bachelor, you may refrain from hunting the dangerous Karogi oxen", it may not always follow from this sentence that husbands are obliged to take part in that hunt. For instance, if it were also known that the weak and wounded are permitted to refrain anyway from participating, one would exclude those husbands from this duty who belong to that group of weak and wounded. A more complex ought table, with more than one forbidden cell, would need to be constructed.

Thus, the above description of the four conditional deontic connectives is meant as a basic interpretation, which is only claimed to hold for a 2 × 2 ought table and thus for dichotomous and exhaustive classes (like bachelor vs. husband) which are internally homogeneous in regard of allowed and forbidden cases.

Compared to the main theories of the WST, such a deontic logic combines two apparently opposed ideas. On the one hand, it preserves the concept of a logical core, emphasized traditionally by domain-general theories of the WST. On the other hand, it follows pragmatic reasoning schema theory in assuming that there are different schemas of prescriptive conditionals. FDL theory can account for the empirical difference between obligations and permissions by differences in the corresponding ought tables. Here a more general semantics of conditionals has been outlined, referring to all four possible cases with one forbidden cell in an ought table. Unlike pragmatic reasoning schema theory, the answers in the WST are claimed to be based on the more general concept of ought tables and the rational systematics of deontic logic.

9.3 Pragmatic Cell Focus on Cheater or Cooperator Detection

FDL theory claims that deontic logic alone is not sufficient to account for selection patterns in prescriptive WSTs. Deontic logic itself does not provide a norm according to which cards are to be selected if one should check a prescriptive rule. A further essential component of FDL theory postulates that based on one's goals (or based on corresponding pragmatic contexts) people will rationally focus on different cells of an ought table. In the view of FDL theory the so-called 'cheater detection algorithm' is, nothing but one specific focus on the forbidden cells of an ought table (cf. Oaksford & Chater, 1994; Chater & Oaksford, 1996; Love & Kessler, 1995; Liberman & Klar, 1996; Sperber, 2002, 2003). Here it is argued that the 'cheater detection algorithm' is to be explained by the more general phenomenon of focusing systematically on different cells of an ought table. The idea of focus effects (Oaksford & Chater, 1994; Sperber, 2002, 2003) is combined with the deontic logic of testing conditionals and an account suggesting that focusing is part of the normal checking of prescriptive rules.

The flexible focus in checking deontic rules is essential for the checking of prescriptive rules and is often connected with social sanction systems for punishment or gratification, which both appear to be important in ensuring a just division of the benefits and burdens of cooperation (cf. Rawls, 1999/1971; Fehr & Fischbacher, 2003, 2004).

In deontic contexts, checking prescriptive rules typically involves searching either for individuals who have violated the rule or for individuals who have complied with the rule. Hence, checking prescriptive rules does not involve testing cells which are informative for the truth or falsity of these rules (v. Sydow, 2004; cf. Part II), but to *select* – according to ones goals – those acts or those persons which are either following or violating a contractual, social, moral or religious rule.

For example, if generosity is a prescribed norm of a society, then stingy persons should be punished and generous people should be rewarded. Thus, depending on whether punishments or rewards constitute the current pragmatic goal, different cases should be searched for.

The cell focus of testing prescriptive rules is flexible and goal dependent. In some contexts, the focus may indeed be on cheaters, but in others it may be on cooperators. It is claimed that these focus effects are not arbitrary, but they work on the basis of underlying ought tables. A cooperator focus will normally be concerned particularly with those allowed cells of an ought table which are associated with (honourable) adherence to a rule: if one is concerned with a conditional prohibition this is the *p & non-q* cell and but for a conditional obligation it is the *p & q* cell.[54]

[54] More generally, the focus for rule following will be on the cell which is opposed to the forbidden cell, understood as the possible alternative action to a violation of the rule. Hence, FDL theory predicts an asymmetry of using fixed preconditions or optional actions as antecedent or consequent of a rule. For example,

According to the pragmatic aspects of the synthesis of FDL theory the focus on cheater or on cooperator detection is not only dependent on the explicit goal formulated in the instruction, but depends as well on implicit influences on the goal instruction. The setting of the task, the role or perspective of the tester and even the connotation of the formulation of the rule itself may all contribute to the construction of a subjective focus. If there is contradicting implicit and explicit information, subjects may either use the two foci simultaneously (if both are salient) or only the more salient one. However, if the scenario of a WST is formulated openly enough, the induced goals may be varied freely to elicit both cell foci. This should for example be the case if the goals of punishment and gratification in a social sanction system are both regarded to be equally plausible. Let us think of an obligation rule, like "If someone is a bachelor, then he must abduct a virgin from a hostile tribe", or a permission rule, like "if someone is a bachelor, then he is forbidden from fleeing from a lost battle". (It is assumed that we do not need to share the moral of some barbarian rules to understand its deontic logic.) Again, we imagine ourselves to be members of a council of elders, who have to check whether the rules of the tribe have been violated or followed. Depending on whether our goal is to honour those who followed the rule or to punish those who violated the rule the focus of selection should be on cooperator or cheater detection cells (Table 6, 7).

Table 53
Conditional Obligation Rule (with a Normal Precondition Action Order)

(a) a Cheater Focus (Full Circle)

	Does abduct virgin (*q*)	Does not abduct virgin (*non-q*)
Bachelor (*p*)	Allowed	Forbidden
Husband (*non-p*)	Allowed	Allowed

(b) Cooperator Focus (Dotted Circle)

	Does abduct virgin (*q*)	Does not abduct virgin (*non-q*)
Bachelor (*p*)	Allowed	Forbidden
Husband (*non-p*)	Allowed	Allowed

in the obligation rule a cooperator focus with a fixed antecedent will lead to *p* & *q* cooperator selections, but a fixed consequent to *non-p* & *non-q* selections. This will be addressed in more detail elsewhere. Furthermore, additional knowledge about which of the allowed cells is laudable may play a role. In this case our predictions would be more readily formulated based on a three- or more-valued deontic logic, based on values like 'allowed', 'forbidden', and 'obligatory' or 'honourable'. A discussion of such a more complex approach lies outside the scope of this article.

Table 54

Conditional Prohibition Rule (with a Normal Precondition Action Order)

(a) Cheater Focus (Full Circle)

	Flees from battle (*q*)	Does not flee from battle (*non-q*)
Bachelor (*p*)	Forbidden	Allowed
Husband (*non-p*)	Allowed	Allowed

(b) Cooperator Focus (Dotted Circle)

	Flees from battle (*q*)	Does not flee from battle (*non-q*)
Bachelor (*p*)	Forbidden	Allowed
Husband (*non-p*)	Allowed	Allowed

A further prediction of combining deontic logic with the focus idea is that if one ought to check for both, for cheaters and for cooperators, this may result in what I call ‘double focus effects’. Testing for instance an obligation rule, this may cause a novel dominant *p* & *q* & *non-q* selection pattern. It will be shown that double focus effects can dissociate FDL theory from mental model theory. (More detailed predictions and a more detailed discussion of double focus effects is provided in Section 12.3.)

Although perspective effects may perhaps also be interpreted as some kind of ‘focus effects’, such focus effects were explicitly only concerned with different cases of cheating (Gigerenzer & Hug, 1992; cf. e.g., Beller & Spada, 2003). In contrast, FDL theory predicts more general focus effects not only based on the goal of cheater detection but also on the goal of cooperator detection.

The predicted effects would be particularly critical for the ‘cheater-detection’ approach (Cosmides, 1989; Cosmides & Tooby, 1992; Gigerenzer & Hug, 1992), but would also challenge other theories of the WST, which do neither predict symmetrical focus effects nor different ought tables of conditionals (cf. General Discussion of Part III).

Before we come to test the different aspects of FDL theory empirically (Chapters 10 to 12), FDL theory will be presented as a synthesis of ideas.

9.4 FDL Theory as a Synthesis of Ideas

The outlined flexible deontic logic theory of testing prescriptive rules in the WST is, of course, not a *creatio ex nihilo*, but a proposal to integrate some aspects of other theories and some findings of the intensive research tradition on the WST. Although FDL theory differs from all other theories of the WST, its debts both to domain-specific and domain-general accounts of the WST shall be made explicit.

Domain-Specific Aspects of the Synthesis

Is-Ought Distinction

The postulated distinction of testing descriptive or prescriptive rules has been inspired particularly by the domain-specific accounts of pragmatic reasoning schema theory and of social contract theory (Cheng & Holyoak, 1985; Cosmides, 1989; Cosmides & Tooby, 1992; Holyoak & Cheng, 1995). Although FDL theory opposes the modular and irrational aspects of these theories (cf. pp. 264 f.), and although these theories provide no positive account for descriptive rules, it does owe much to the domain-specific treatment of deontic or social rules.

Additionally, other approaches have contributed to make an is-ought distinction explicit. The assumed is-ought distinction was first applied to the WST by Manktelow and Over (1990) – without distinguishing different test strategies for the two realms. Moreover, the full is-ought distinction, as assumed here, was made possible by the development of an elaborated Bayesian account of testing descriptive rules in the WST (Oaksford & Chater, 1994, 1998, 2003; von Sydow, 2004; cf. also Evans & Over, 1996; Green, Over, & Pyne, 1997; cf. Part II).

Finally, the postulated contrast of two different dominant test strategies with a different effect on frequency information may be seen to be derivable from the analysis of Oaksford and Chater (1994), which was indeed partly also concerned with prescriptive rules. However, firstly, Oaksford and Chater (1994, 1998a) mainly concentrated on descriptive tasks. Secondly, they did not explicitly derive nor test the different predictions for prescriptive and descriptive rules concerning frequency effects, which, in my view, may also be derived from their analysis. Thirdly, they did not propose a deontic logic and did not distinguish different foci.

Here the postulated different predictions for frequency effects in checking either descriptive or prescriptive rules will be investigated for the first time (Chapter 10).

Domain-Specific Mechanisms

FDL theory also integrates particular mechanisms originally proposed by domain-specific theories into a more coherent theory.

Firstly, pragmatic reasoning schema theory has distinguished two schemas of deontic conditionals from a material implication. FDL theory continues this schema based tradition of deontic reasoning, but systematizes this idea based on ought tables of a deontic logic, leading to a complete system of four types of deontic conditionals with one forbidden cell. If one understands the postulated deontic logic of the WST as an extension of PRS theory, one would have to add production rules like “If the precondition is fulfilled than the action is forbidden” to that theory.

Secondly, although FDL theory turns against a cheater detection approach (Cosmides, 1989; Gigerenzer & Hug, 1992) the postulated flexible focus mechanism may be seen as a generalization of the cheater detection mechanism, generalising the focus idea to cases of cooperator detection and cases of a double focus (and combining it with deontic logic).

Thirdly, FDL theory is a domain-specific theory in assuming that the goals of cooperator or cheater detection (or other goals) only elicit the postulated foci in a WST, if the goals are understood to be plausible and applicable in a given situation. If we want to predict which situations render goals plausible, we need to take the individual and the collective learning history into account, including evolutionary, historical, social, biographical and even ethical considerations. Only in situations in which it is equally plausible to apply both the cheater and the cooperator detection goal to the WST does FDL theory predict the described selection patterns. In contrast to both domain-specific theories, I would not advocate either an evolutionary background or a schema theoretic background – but both. I do agree with domain-specific approaches at least in so far as I see a need for further research on the domain-specific contexts to determine the conditions under which certain goals and certain ought tables are elicited in the first place – but only as a precondition to apply a more general rational and flexible deontic logic.

Rational and Systematic Aspects of the Synthesis

More like domain-general theories, FDL theory defends a systematic and comparatively rational approach of testing deontic rules, based on deontic logic and a general focus mechanism.

Deontic Logic

Particularly, Manktelow and Over (1995) referred to aspects of a deontic logic; but they have only discussed perspective effects (permissions and obligations) and not the full system of all four deontic conditionals with only one forbidden cell. In this sense, the deontic logic account advocated here is more general.

Lately, Bucciarelli and Johnson-Laird (2005) seem to have broken with traditional mental model theory (Johnson-Laird & Byrne, 1991, 1992, 1995, 2002), which due to the principle of truth cannot account for prohibitions (cf. General Discussion), and have elaborated a complete system of deontic conditionals for syllogistic reasoning, including prohibitions (cf. Chapter 13, pp. 275 f.). Although the current proposal was derived independently from their approach (v. Sydow, 2005a; cf. v. Sydow, et al., 2005), their paper has to be acknowledged as an earlier proposal of a fully equivalent deontic logic of the four different types of conditionals. However, Bucciarelli and Johnson-Laird (2005) did not apply this theory to the WST. Moreover, FDL theory applies the four types of deontic conditionals to the WST, without relying on any specific mental model assumption. Unlike mental model theory, FDL theory combines deontic logic with a flexible focus mechanism, which would replace or complement an explanation based on incomplete representations (mental models). Moreover, the postulated focus mechanism leads to different predictions concerning selection patterns than the concept of fleshing out a mental model (cf. Chapter 10, 12, and, particularly, Chapter 13 and the General Discussion of Part III).

Even earlier, also Beller proposed in two conference papers (2001, 2003) a reasoning account explicitly based on deontic logic and provided evidence for the deontic logic of all four mentioned types of conditionals. In regard of the WST, only Beller (2001) may be said to have provided a first test of the four deontic conditionals (cf. Experiment 10). However, Beller (2001) did not formulate the four deontic conditionals explicitly on equal footing (without negations and without modals), and, more importantly, he did not combine them with the idea of *different* foci.

Focus Effects

FDL theory for the first time combines deontic logic with focus effects. A flexible focus concept was earlier proposed in the context of relevance theory (Sperber, 2002, 2003; cf. Oaksford & Chater, 1994; Love & Kessler, 1995; see also the discussion of Experiment 11). Relevance theory postulates that attention may be drown to different specific cases. FDL theory interprets this mechanism as the standard mechanism for testing prescriptive rules. Unlike relevance theory, FDL theory combines a flexible focus with a deontic logic of testing prescriptive rules. In this respect, the FDL theory of the WST can be regarded as a synthesis of a full deontic logic of conditionals (cf. Bucciarelli & Johnson-Laird, 2005; see p. 275) based on ought tables with the focus concept first elaborated in relevance theory (Sperber & Girotto, 2002, 2003). Unlike all other theories of the WST, FDL theory combines these two mechanisms.

The perspective effects introduced by Gigerenzer and Hug (1992) might perhaps also be seen as a kind of focus effects. However, the perspective effects of Gigerenzer et al. were exclusively concerned with shifts *within* different cheater cases, not *between* cheater and cooperator cases. Moreover, their account did elaborate on the link between deontic logic and such focus effects.

In conclusion, several aspects of FDL theory have been discussed before in various contexts, but FDL theory provides a coherent framework and it is aimed at integrating ideas from both the converging domain-specific and the domain-general research traditions.

In the following five experiments the three claims of FDL theory are tested successively, firstly, the different effect of frequency information on descriptive or prescriptive rules, secondly, deontic logic, and, thirdly, the combination of deontic logic and focus effects (climaxing in an experiment on double focus effects).

10 First Aspect of FDL Theory: Frequency Effects in Descriptive Versus Prescriptive Rules

The first claim of FDL theory to be investigated is the postulated difference between testing the truth or falsity of a descriptive rule and checking for violations of a prescriptive rule.

Before describing and discussing the two experiments (Section 10.2 and 10.3) a further introduction on the postulated test strategies will be given (10.1, cf. Section 9.1).

10.1 On Frequency Effects in Prescriptive and Descriptive rules

I have outlined the difference in the meaning of descriptive hypotheses, concerning for instance laws of nature, and in the meaning of prescriptive rules, concerning for instance laws in a legal system, in that the latter cannot directly be falsified or confirmed by empirical evidence (cf. Section 9.1, Manktelow & Over, 1991, 1995). Additionally, FDL theory has postulated different test strategies depending on whether a WST is understood as a descriptive or as a prescriptive task. Standard descriptive tasks are claimed to be tested along Bayesian lines (cf. Part III). In contrast standard prescriptive tasks are predicted to be tested according to a deontic logic and an a priori focus (for instance, on the forbidden cases of an ought table). If this account is correct, normally frequency information should affect descriptive WSTs but not prescriptive WSTs.

For example, consider the following imagined prescriptive rule of a foreign tribe: “If one rides a horse, one must wear one or more feathers”. A guard who should check particularly this rule should punish rule violations. With this goal in mind, the guard should (a priori) only be interested in events of ‘horse riding’ (*p*) and of ‘wearing no feather’ (*non-q*), because the combination of these cases (*p & non-q*) is to be punished. Alternatively, let us think of an anthropologist who aims to check the truth or falsity of the hypothesis that “If one rides a horse, one wears one or more feathers”.

Provided that horse riding and the wearing of feathers are both rare, a Bayesian analysis (together with a few additional assumptions) would predict that the cases mentioned are the most informative cases. If we assume rarity, selection patterns like *p* or *p* & *q* or perhaps *p* & *q* & *non-q* are to be expected, but no *p & non-q* patterns. Only if the atomic propositions, mentioned in the hypothesis, are frequent (most tribesmen ride a horse and most wear more than one feather), the selections of other cases may become informative (Oaksford & Chater, 1994, 2003; v. Sydow, 2004; for details see Part III).

This difference of testing prescriptive and descriptive rules may explain three known phenomena in the WST literature. Firstly, the account on testing descriptive rules would be coherent with the well-established predominance of 'confirmatory' *p* & *q* patterns in such rules (Wason, 1966; Johnson-Laird & Wason, 1970; etc.). Secondly, if a cheater goal is given this account on checking prescriptive rules may explain the 'facilitation' effects for obligation (and permission) schemas (e.g., Cheng & Holyoak, 1985). Thirdly, this distinction would be coherent with found contrast of clear-cut selection patterns in social contracts, and less clear-cut patterns in testing descriptive rules (e.g., Cosmides, 1989; Gigerenzer & Hug, 1992; Sperber & Girotto, 2002, 2003). For instance, Cosmides (1989, Exp. 1, 2, cf. Gigerenzer, Exp. 1) found clear-cut *p* & *non-q* selection patterns in a social contract task and a reduced number of *p* & *non-q* selections in a task were an anthropologist should test the truth or falsity of the same rule. This could be interpreted in the light of the postulated differential use of frequency information in prescriptive and descriptive tasks.

However, the postulated differential effect of frequency sensitivity has not yet been tested directly. This should be done in the following two experiments.

In these experiments, MSTs will be used again with a forced choice option between *p* versus *non-p* and *q* versus *non-q* selections. The advantages of a forced-choice instruction have been discussed earlier (see Section 5.3). Here an additional reason for these *particular* forced-choice options is to exclude rational frequency effects, which may occur in prescriptive tasks too (cf. Kirby, 1994, Humberstone, 1994): Even if it would be true that participant a priori focus on a particular cell in testing a prescriptive rule, it would still be rational to consider frequency information in regard of rather choosing between the cards which relate to this focus. For instance, a clear cheater goal should elicit cards selections relating to the *p* and *non-q* cell of the ought table of an obligation. Nonetheless, there may be rational frequency effects for *p* versus *non-q* cards depending on $P(\neg q \mid p)$ and $P(p \mid \neg q)$. Such effects will here be called *within-focus frequency effects*. In contrast, the Bayesian analysis is many interested in other frequency effects, like *q* versus the *non-q* frequency effects. In order to exclude within-focus frequency effects, the forced choice was here between *p* versus *non-p* and *q* versus *non-q*.

The design and the general predictions for both experiments are illustrated in Table 55. In the descriptive and the prescriptive conditions, conditionals are tested with a single forbidden (false) *p & non-q* cell. The descriptive conditional as well as the prescriptive conditional will be social rules. The descriptive social rule should confirm the frequency effects found in Part II for testing descriptive rules. The predictions are based on the Sydow model (see Chapter 5). For the low probability conditions (‘0.10 → 0.15’) the following relation between the expected information gain values of the selections result: *EIG*(*p*) > *EIG*(*non-p*), *EIG*(*q*) > *EIG*(*non-q*). For the high probability conditions (‘0.85 → 0.90’), an increase of *non-p* and *non-q* selections is predicted. In contrast, if the rules used are understood as prescriptive obligations (with an a priori cheater focus), there should be a generally high level of *p* and *non-q* answers without any impact of the frequency manipulation.

Table 55

General Predictions for the Card Selections in Experiment 8 and Experiment 9 *in the High versus the Low Probability Conditions*

		Low probability conditions	High probability conditions
Descriptive task conditions	*p* vs. *non-p*	*p*	*non-p*
	q vs. *non-q*	*q*	*non-q*
Prescriptive task conditions	*p* vs. *non*-p	*p*	*p*
	q vs. *non-q*	*non-q*	*non-q*

Note. Comparing the low and high probability conditions in each row it becomes clear whether frequency effects are predicted or not.

In Experiment 8 a comparatively weak manipulation was used, which led only to some of the predicted differences between testing descriptive and prescriptive tasks. It will be argued that not all rational uses of frequency information were removed for the prescriptive WSTs. The pattern remains consistent with the distinction of different strategies for checking prescriptive rules and for testing descriptive rules. A refined classification of different kinds of frequency effects, consistent with such a distinction, will be given at the end of Experiment 8 and in the introduction to Experiment 9.

In Experiment 9, a stronger manipulation was used and all remaining rational uses of frequency information in prescriptive tasks, which would not relate to the tested dichotomy, were excluded.

10.2 Experiment 8 – Descriptive versus Prescriptive Rules I

Method Experiment 8

Design and Participants

The experiment had a 2 (prescriptive vs. descriptive rule) × 2 (low probability vs. high probability) between-subjects design.

91 participants originally took part in the experiment. The participants got a little present and could win additional prizes.

80 participants from the University of Göttingen (56 % female, 44 % male, mean age: 23 years) finished the task formally correctly. Most of the participants were students: most studied psychology (22 %), law (22 %) and economics (20 %). The participants were randomly assigned to the four conditions.

Eleven participants were excluded, because they did not comply with the formal instructions or because they knew the task. (Most of them selected more than the two cards, to which the choice was explicitly limited, opting for the pattern *p p q,* for instance.) These participants were replaced by other subjects.

Procedure and Materials

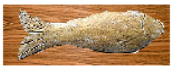
The participants were asked to fill in a questionnaire on the campus. The experimenter checked that each participant solved the task separately. The deontic MSTs were carried out as paper and pencil tasks and the instructions were in German.

The prescriptive and the descriptive conditions varied in the beginning and at the end of the instructions, the middle section was identical.

In the prescriptive conditions it was stated that the participants should imagine they were members of a council of elders of a tribe, who had a *police function*, and whose task it was to *punish* those, *who had violated the tribal law* (italic text was highlighted in the original instruction).

In the descriptive conditions, participants should imagine that they were an *ethnologist investigating a tribe*'s social rules.

In all conditions, a tribal law was presented. The descriptive rule read, "*If someone is a Bachelor, then he brings fish to the medicine man*" and the descriptive rule only the modal verb 'must' was added: "*If someone is a Bachelor, then he must bring Fish to the medicine man.*" In the descriptive conditions, the alternative hypothesis was clarified: in the alternative case marital status and acts of bringing fish only appear together in a completely random way. This was done since FBL theory asserts that a hypothesis is tested against some alternative hypothesis.

In all conditions, it was then explained that there is a two-sided wooden panel for each male member of the tribe. On one side of each panel, it said whether the person is a bachelor (followed by:) or husband ( ). Then they were told, that on the other side of the same panel it was registered, whether the tribe member brought fish () or not ( ). The instruction continued that a chief of that tribe shows the participant all front sides of all panels of the tribe (the number of bachelors and husbands in that tribe). After mixing the panels, he then showed all the back sides of the cards (the number of fish and non-fish panels).

There was a high and a low frequency condition in both task types. In all conditions, 20 panels were considered, and hence 20 front sides and 20 back sides were shown (Sydow model). The probabilities for the antecedent ('bachelor') and for the consequent ('brings fish') were '0.10 → 0.15' in the low probability conditions and '0.85 → 0.90' in the high probability conditions.[55] It was emphasised that the mixing of the panels between the displays of the front side and the back side of the cards, caused that the mapping of the single panel sides is still unknown.

Then the participants had to imagine that the chief of the tribe allows them to turn over only two cards separately. The goals were repeated, depending on the condition: in the prescriptive conditions the goal to punish the violation of the law, in the descriptive condition the goal to test the truth or falsity of the hypothesis.

The participants were then required to tick those cards, which they would turn over to fulfil their task. They could tick one card to be turned over out of the first display (a particular bachelor or a particular husband) and another out of the second display (a particular fish or a particular non-fish).

[55] The seldom cards were always shown in the same position weither as front sides or as back sides of the panels (cf. Experiment 1).

Results Experiment 8

The results are displayed in Table 56 and they are graphically illustrated by Figure 14.

Table 56
Card Selections for the Low and High Probability Conditions in Descriptive and Prescriptive Tasks, Experiment 8 (N = 80)

Card selected	(a) Descriptive tasks				(b) Prescriptive tasks			
	Low prob. condition		High prob. condition		Low prob. condition		High prob. condition	
P	90 %	18	45 %	9	85 %	17	75 %	15
¬P	10 %	2	55 %	11	15 %	3	25 %	5
Q	65 %	13	25 %	5	60 %	12	25 %	5
¬Q	35 %	7	75 %	15	40 %	8	75 %	15
n		20		20		20		20

Note. Percentage and number of participants selecting each card. Predicted answers in darkened cells.

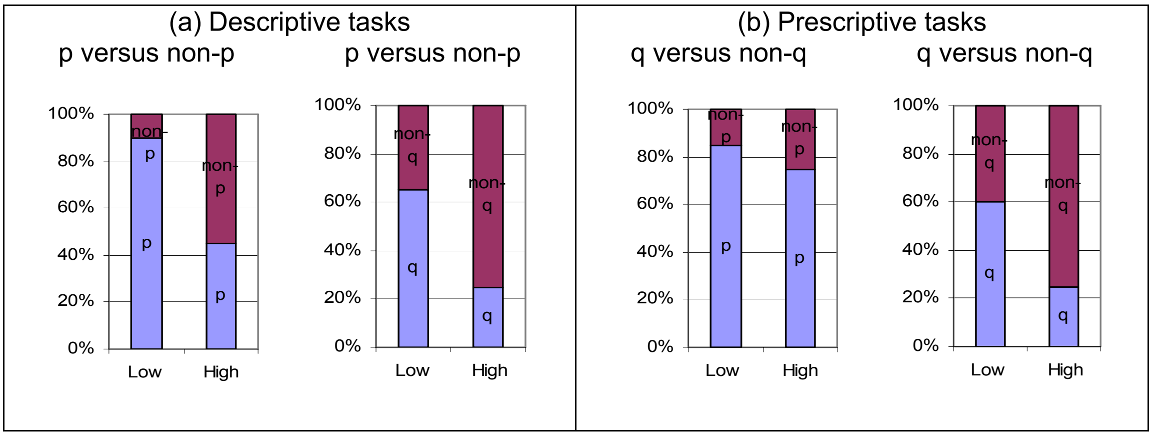


Figure 14. Bar graphs of the percentages of the *p* versus *non-p* selections and *q* versus *non-q* selections in the low and high probability conditions of the two task types.

If comparing the low and the high probability condition, *non-q* and *non-p* selections were predicted to increase in the descriptive tasks. As predicted the portion of *non-q* selections against *q* selections significantly increased in the high probability condition (Pearson $\chi^2_{(1)} = 6.46$, $p < .05$, $r_\varphi = .40$). Likewise, for the more problematic *p* versus *non-p* alternative the predicted increase in the proportion of *non-p* selections was found (Pearson $\chi^2_{(1)} = 9.23$, $p < .01$, $r_\varphi = .48$).

For the prescriptive tasks a high level of *p* and *non-q* selections were predicted independently of the frequency conditions. As expected, there was no significant difference between the conditions for the *p* versus *non-p* choice (exact Fisher test: $p = .70$, $r_\varphi = .12$). But contrary to the predictions there was a higher number of *non-q* selections in the high probability condition than in the low probability condition (Pearson $\chi^2_{(1)} = 5.01$, $p < .05$, $r_\varphi = .35$).

Discussion of Experiment 8

The results of Experiment 8 only provided partial support for my original predictions.

Positively, the results confirmed the postulated role of frequency information for descriptive hypothesis, even if the hypothesis is a social rule. In the low versus high frequency conditions, the predicted *p* versus *non-p* and *q* versus *non-q* frequency effects were found based on introducing the preconditions for the Sydow (cf. Chapter 5). For descriptive tests of social rules, this is a novel finding predicted by FBL theory, which goes beyond our previous results. Additionally, as predicted the *p* versus *non-p* frequency effect disappeared in the prescriptive task conditions, a constantly high level of *p* selections was found in both probability conditions.

However, against the predictions another frequency effect, the *q* versus *non-q* effect did not disappear. In the corresponding low probability condition, many *q* cooperator selections were found instead of the 'correct' *non-q* cheater selections.

It will be argued in this discussion that these mixed results can be explained in a way which is coherent with the postulated distinction of different standard tests for descriptive and prescriptive tasks.

Can the remaining frequency effects in checking prescriptive rules be explained in a way that is coherent with the postulated dichotomy of testing prescriptive and descriptive rules differently? Are there rational reasons for frequency effects also for prescriptive rules which have not been considered here so far, but which do not contradict our assumption that prescriptive tasks normally have an *a priori* focus?

Within focus frequency effects. As argued in the introduction to this experiment FDL theory would be coherent with frequency effects, if they would concern only the cards referring to the same focus (here: *p* versus *non-q* effects). However, this cannot explain the found *q* versus *non-q* effect (cf. instruction of Experiment 9).

Are there other rational frequency effects plausible also in prescriptive tasks? The comments of the participants certainly pointed out that they had obviously tried to make rational use of the salient frequency information provided. Many participants in their written or oral comments explicitly provided justifications for their selections. In my view there are three, more or less, rational justification strategies regarding why participants made use of the frequency information provided also in the probability condition of testing the prescript “If someone is a bachelor, then he must bring fish to the medicine man” and ‘falsely’ selected *q* (‘brought fish’) instead of the predicted *non-q* (‘did not bring fish’).

Task reinterpretation strategy. If the advocated Bayesian approach is correct, frequency information is of particular importance for testing descriptive rules. Hence, in a laboratory experiment frequency information may become itself a cue for interpreting the task as a hypothesis-testing task. For example, a participant in a prescriptive task condition wrote that a *p* and *q* selection appears rational to him, because this can prove with a high probability that all bachelors actually bring fish. This reinterpretation strategy is false in the light of the formally given instructions, but it would be in accordance with Grice’s (1975) communicative cooperative principle (see below). Provided this reinterpretation of the task, the found *p* and *q* selections would be coherent with FBL theory. Nonetheless, since *p* versus *non-p* frequency effects did disappear, there needs to be an additional explanation, which exclusively explains the presence of *q* versus *non-q* frequency effects and the disappearance of *p* versus *non-p* effects.

Sorting Strategy. Two participants who selected *p* and *q* in the low-frequency prescriptive task condition wrote in their comments that they aimed to select all those who *followed* the rule, in order able to punish the rest later on. This strategy indeed should not affect the *p* versus *non-p* selections, but only the *q* versus *non-q* selections. In this case, the *p* versus *non-p* effect should indeed disappear, since *p* (‘Bachelor’) is relevant in both frequency conditions for sorting the bachelors in cheaters and cooperators. In contrast, for the choice between *q* (‘brought fish’) and *non-q* (‘did not bring fish’) it would be rational in the low probability condition to check only for the few *q* cases (‘brought fish’) and then to punish the larger number of remaining non-cooperators. A sorting strategy is irrational only if we assume that the selection process terminates with the end of the task and if the task does not allow identifying all cooperators. In contrast, the mentioned participants seem to have understood this task as a beginning of a complete (and just) check to find all violators of the rule.

Goal reinterpretation strategy. Some participants who selected *p* and *q* cases in the low probability condition neither questioned the prescriptive task character of the task nor adopted a sorting strategy. Instead, they questioned whether the goal of the task was indeed cheater detection. For instance, a participant argued that since more fish had been brought than there were bachelors, there should in any case be enough fish. Hence, he argued, there was no need to punish anybody. But why should goal reinterpretation lead to a cooperator detection goal in the low frequency condition? If only a low percentage follows a rule, a cooperator detection goal generally becomes more efficient to distinguish followers and violators of a rule, since fewer subjects have to be singled out. Conversely, a cheater detection strategy becomes more efficient if a high percentage follows the rule. Using some assumptions about the reasonable application of the checked social rule as additional premises, this answer becomes fully reasonable. But do we actually know anything about the numbers of rule followers and rule violators based on our probability manipulation? Not directly, but the following estimation is plausible: If one randomly matches 'bachelor' cases and 'brought fish' cases in the low probability condition '$0.10 \xrightarrow{Oblig.} 0.15$', a high percentage of bachelors would cheat. In contrast, in the high probability condition '$0.85 \xrightarrow{Oblig.} 0.90$' a high portion of bachelors would follow the rule anyway. In conclusion, the goal reinterpretation strategy provides another more or less rational explanation why some participants in the low probability condition may have adopted a *p* & *q* cooperator-detection strategy instead of the predicted *p* & *non-q* cheater detection strategy. In order to make use of the salient frequency information, some participants seem to have adopted the goal that corresponds to the most efficient strategy of sanctioning. The resulting kind of *q* versus *non-q* frequency effect for prescriptive tasks differs from the frequency effects predicted for descriptive tasks. Hence, retrospectively, our data would be coherent with such an interpretation.

The comments of the participants and the results suggest that there was a tendency that participants actively interpreted the task in a way, in which all salient information provided can be utilised, even if this interpretation contradicts some aspects of the explicit instruction. This active (and subjective) construction of the task interpretation would be coherent with Grice's (1975) co-operative principle. A participant can assume that an instruction provides exactly the information the participant needs to know to solve the task correctly – and nothing more. Hence, a participant will reinterpret the task in a way in which one can make rational use of the frequency information provided as long as this only moderately violates the instructions. This constructive approach to task understanding would, for instance, also be in line with evidence that information inconsistent with a situation have a low probability of being encoded, remembered or retrieved (see, e.g., research in the wake of Bransford, Barclay, & Franks, 1972).

The three outlined misunderstandings and reinterpretations of the instruction are problematic in the light of the formal instructions. However, given that participants subjectively represented the task falsely in the described ways, they would have been rational and coherent with the distinction made here. Moreover, if the majority of participants in the low frequency condition adopted the outlined latter two interpretations, one would have predicted exactly the pattern of partially reduced frequency effects in the prescriptive task condition as actually found in Experiment 8.

In the following experiment, the aim was to corroborate this explanation by replicating Experiment 8, while excluding the outlined alternative interpretations of the task.

10.3 Experiment 9 – Descriptive versus Prescriptive Rules II

In Experiment 9 the postulated frequency effects for testing descriptive rules and the absence of frequency effects for prescriptive rules (with an a priori focus) should be tested, excluding the alternative task interpretations found in Experiment 8.

A Typology of Unintended Frequency Effects in Prescriptive Rules

In Experiment 8 it has become apparent that reduced frequency effects can also occur in prescriptive tasks. It has been elaborated that FDL theory is coherent with four kinds of frequency effects also in prescriptive tasks. Here a more systematic treatment of all possible four frequency effects in prescriptive MSTs should be provided. It will also be outlined, how these frequency effects for prescriptive tasks differ from those for descriptive tasks (cf. Discussion of Experiment 9) and how to exclude these prescriptive frequency effects in order to test the postulated dichotomy of descriptive and prescriptive tasks under ideal conditions.

(a) *Frequency effects within a focus*. As argued before, FDL theory is not at odds with *p* versus *non-q* frequency effects, if these selections both refer to a *p* & *non-q* focus. Frequency may of course also play a role if one had to decide between these two cards (see Kirby, 1994a, 1994b; Humberstone, 1994; Manktelow, Sutherland &

Over 1995). A Bayesian analysis of a deontic task with a focus on *p* & *non-q* should lead to the similar predictions as proponents of a probabilistically refined general falsificationist position have assumed (see Humberstone, 1994; Kirby, 1994a; cf. Green, Over & Pyne, 1997): One would have to predict more *p* selections in the low probability conditions and more *non-q* selections in high probability conditions. This position goes beyond traditional falsificationism, but falls short of a complete Bayesian approach. However, this within-focus probability effect needs to be distinguished from *q* versus *non-q* or *p* versus *non-p* frequency effects. Since possible within-focus frequency effects were intentionally excluded in Experiment 8, by not testing *p* versus *non-q* effects, no change is required for Experiment 9.

(b) *Task reinterpretation as descriptive task.* A prescriptive laboratory WST may be reinterpreted as a hypothesis-testing task, if frequency information is provided and this information has no other usage (cf. the Discussion of Experiment 8). If we assume that the task is (falsely) reinterpreted as a hypothesis-testing task, the use of frequency information becomes rational. In Experiment 9 such a possible re-interpretation should be prevented by putting additional emphasise on the hypothesis testing character or the cheater detection character of the corresponding tasks.

(c) *Frequency based 'sorting'*. There may be sorting effects even if a cheater detection goal in testing an obligation rule elicits an *a priori* focus on *p* & *non-q* cases and the participants understand that the ultimate goal is cheater detection. A low probability condition may lead to a cooperator detection goal – on a tactical level – in order to fulfil the cheater detection goal more efficiently; it may be more efficient to sort out the low number of cases of rule following first, in order to punish the remaining cheaters. This would lead to the found *q* versus *non-q* effects (and not to any *p* versus *non-p* effects) as found in the last experiment.

A sorting strategy is only rational if two preconditions are met: Firstly, further selections need to be allowed in the future. This precondition was formally not fulfilled. However, the goal of punishment may suggest that *all* possible rule violators need to be checked, ensuring that law is no respecter of persons. Hence, further selections are needed in any case – and a sorting strategy becomes plausible.

Secondly, if all bachelors are known, one can easily build the contrast set of all cheating bachelors. This precondition may be said to have been fulfilled in Experiment 8, because the cards of all bachelors were visible. Moreover, the context story is concerned with a little tribe, where it is plausible that a member of the council of elders knows all bachelors anyhow.

In Experiment 9 the possibility of a sorting interpretation will be excluded by stressing that there is no further sorting process; the punishment of violators will follow directly after checking the cards.

(d) *Frequency effects based on changed goals*. Also the final goal itself (either cheater or cooperator detection) may be influenced by frequency information. We have seen that the goal of cheater detection may gain plausibility if a rule is presumably often followed, whereas the goal of cooperator detection may gain plausibility if (ceteris paribus) the rule is generally violated (cf. Discussion of Experiment 8 for more details). Hence, frequency can be assumed to be a factor determining the choice of a focus in the first place. Although in Experiment 8 the goal was formally determined by the instruction before hand, the need to make rational use of the frequency information may have led participants to neglect these instructions. In Experiment 9 such a reinterpretation will be excluded by emphasising that participants should not aim to check for cooperators.

The outlined reasons for rational frequency effects for prescriptive tasks will be excluded in Experiment 9 in order to show the difference of an *a priori* focus and an *a posteriori* focus as purely as possible. FDL theory predicts for these tasks that frequency effects should only be found in the descriptive tasks, and should now *completely* disappear in the prescriptive tasks.

Method Experiment 9

Experiment 9, like Experiment 8, investigated whether frequency information has an influence on the card selections in testing prescriptive or descriptive rules. In contrast to Experiment 8 the confounding factors which may lead to (different) frequency effects also in prescriptive tasks should be excluded, to test the postulated contrast between a hypothesis testing task (descriptive rule) and a sanctioning task with a fixed focus (prescriptive rule) as purely as possible.

Design and Participants

Like Experiment 8, Experiment 9 had a 2 (prescriptive vs. descriptive rule) × 2 (low probability vs. high probability) between-subjects design.

Another 80 participants from the University of Göttingen (51 % male, 49 % female, mean age: 24 years) correctly finished the task. Again, the participants were predominantly students. The largest group of participants studied law (22 %) and economics (15 %). The participation was voluntarily (everyone got a little present). All participants were randomly assigned to the four conditions.

Twelve subjects were eliminated and replaced before analysing their dada, because they did not follow the formal instructions of the task (cf. Experiment 8).

Materials and Procedure

In the hypothesis testing conditions participants were asked to imagine being in the role of an ethnologist trying to test the truth or falsity of the descriptive rule "*If someone is a Bachelor, then he brings fish to the medicine man."* The instructions in these conditions were completely identical to Experiment 8, only a highlighted title "Testing of Hypotheses" was added to the instructions.

In the prescriptive task conditions, additional changes were made in order to exclude interpretations discussed for Experiment 8. Analogously to the hypothesis-testing task the headline "Punishment of Rule Violators" was added. Then the police function of the council of elders was explained as in the prescriptive task conditions of Experiment 8. The only item to be added in the first section was the explicit statement that a reward of an adherence to a law is not intended. Again the rule read "*If someone is a Bachelor, then he must bring Fish to the medicine man.*"

The middle section of the instruction was identical to Experiment 8. The wooden panels with 'bachelor', 'husband', 'fish' and 'no fish' inscriptions were explained. The two sides of the panels were subsequently displayed, separately showing both sides of the 20 cards. In the high probability conditions (0.10 → 0.15) and low probability conditions (0.85 → 0.90) the corresponding card frequencies were shown (cf. Experiment 8).

There were some modifications in the final section of the prescriptive conditions to emphasise the cheater detection character of the task: The changed instruction read as follows (translation from German): "Because the panels have been mixed (between the displays on either side), you do not know, whether someone has violated the law. Hence, no one can be punished. The chief of the tribe only allows you to turn over two panels separately, the goal being to *punish* found violators *directly*. You may turn over a panel separately from the display of back sides to see its front side. Additionally you may turn over a panel separately from the display of the front sides to see its back side. *You should stick to your goal, to* punish [additionally highlighted] *single persons who have violated the above law.* Please tick those cards, you would turn over, in order to fulfil your task as well as possible."

This clear formulation was used in order to contrast the focus strategy, dominant in deontic WSTs, against a descriptive hypothesis testing task in as ideal conditions as possible.

Results Experiment 9

The results are reported in

Table 57 and graphically illustrated by Figure 15. Descriptively it is apparent that all predicted differences for the descriptive task conditions went in the expected direction and that the predicted selections remained constantly dominant in the prescriptive task conditions.

Table 57

Card Selections for the Low and High Probability Conditions in Descriptive and Prescriptive Tasks, Experiment 9 (N = 80)

	Descriptive tasks				Prescriptive tasks			
Card selected	Low prop. Condition		High prop. condition		Low prop. condition		High prop. condition	
P	80 %	16	50 %	10	75 %	15	85 %	17
¬P	20 %	4	50 %	10	25 %	5	15 %	3
Q	75 %	15	35 %	7	25 %	5	20 %	4
¬Q	25 %	5	65 %	13	75 %	15	80 %	16
n		20		20		20		20

Note. Percentage and number of participants selecting each card. Predicted answers in darkened cells.

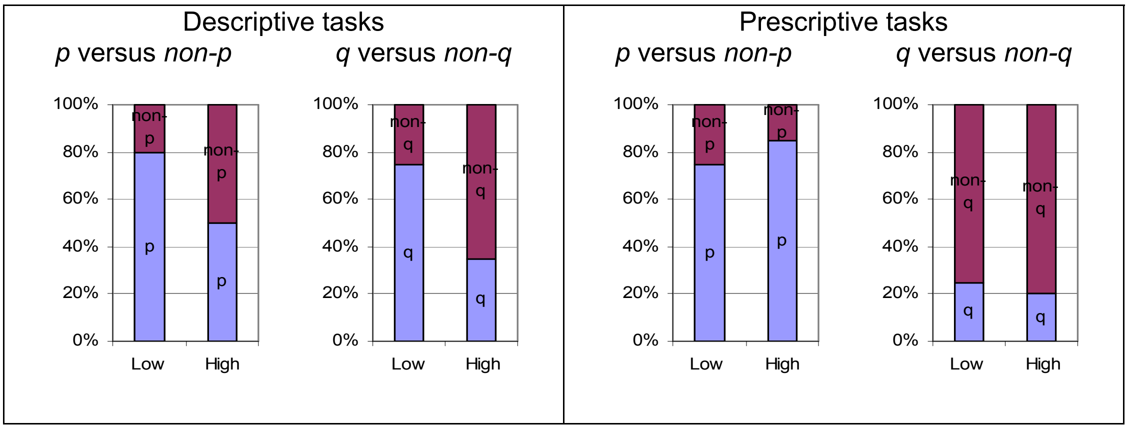


Figure 15. Bar graphs of the percentages of the *p* versus *non-p* selections and *q* versus *non-q* selections in the low and high probability conditions of the two task types.

The statistical analysis also fully confirmed the predictions. For the descriptive task conditions the *p* versus *non-p* selections showed a significantly higher number of *non-p* selections in the low probability condition than in the high probability condition

(Pearson test: $\chi^2_{(1)} = 3.96$, $p < .05$, $r_\varphi = .31$). Also the predicted increase of the portion of *non-q* selections (and the corresponding decrease of the portion of *q* selections) in the high relative to the low probability condition was corroborated (Pearson test: $\chi^2_{(1)} = 6.47$, $p_{one\text{-}tailed} < .01$, $r_\varphi = .40$.)

As predicted for the prescriptive task conditions now no difference was found between the low and high probability conditions both for the *p* versus *non-p* choice (exact Fisher test: $p = .70$, $r_\varphi = -.13$) and for the *q* versus *non-q* choice (exact Fisher test: $p = 1.00$, $r_\varphi = .06$). In conclusion, all four tests confirmed the predictions.

Discussion of Experiment 9 and General Discussion of Chapter 10

In Experiment 9, the frequency effects predicted for descriptive tasks were replicated for descriptive social rules and it was shown that the frequency effects vanished completely if the task was clearly formulated as a prescriptive task with an apriori focus.

In Experiment 8 the prescriptive formulation of the task only partially prevented frequency effects. This shows that we should not underestimate the intricacies of the relation between quantification and deontic tasks. Explanations have been proposed, of how the prescriptive task could have been interpreted by the participants in such a way as to rationally cause this particular kind of partial frequency effects. Experiment 9 confirmed that the frequency effects completely disappear, if these alternative interpretations of the task were excluded. If we take the results of the two experiments together, they corroborate the advocated idea of at least two different types of test strategies and their different sensitivity to manipulations of frequencies.

The corroboration of two test strategies in Experiment 9 may even seem to be trivial from a common sense view, but it is not at all trivial with regard to the WST debate. The obtained results are important for disentangling the two test strategies, which have not been distinguished in the WST debate in this way before. In deontic WSTs focus tasks were mostly used and contrasted to descriptive tasks (Cosmides, 1989, Cosmides & Tooby, 1992, Gigerenzer & Hug, 1992, Cheng & Holyoak, 1995) without understanding their completely different character (cf. Sperber & Girotto, 2002, 2003). As I advocate a Bayesian account for testing hypothesis (Part II) – unlike

Sperber et al. -, it here became possible to dissociate the two test strategies with regard to frequency effects for the first time – both in Experiment 8 and 9.

We will first discuss FDL theory and its relation to the distinguished kinds of frequency effects. Then the obtained results will be related to previous empirical results. Finally, it will be outlined that the findings are problematic for all other theories of the WST.

FDL Theory and Different Kinds of Frequency Effects

The found two dissociations between testing the truth of a rule in a descriptive task and checking an a priori focus, typical for a prescriptive task, supports the distinction postulated by FDL theory.

Although the results of Experiment 9 directly support the postulated different test strategies, we need to formulate a refined dichotomy of two test strategies in Experiment 8. Firstly, we have seen a particular kind of probability effects in prescriptive tasks in Experiment 8; secondly, I would also predict the absence of frequency effects under certain conditions for in descriptive tasks as well.

Concerning the latter point, I think that frequency effects may be absent, if descriptive tasks are formulated as categorisation tasks (cf. cf. Sperber & Girotto, 2002, 2003; van Duyne, 1974; Platt & Griggs, 1993; Liberman & Klar, 1996; cf. for another aspect v. Sydow, 2004). However, one may rebut that a categorisation task is not a proper hypothesis testing task concerned with the question of truth or falsity of a rule (cf. Oaksford & Chater, 1994, 1995, 2003; von Sydow, 2004), and, additionally, that a categorisation task can be seen as an a priori focus task as well (cf. the Discussion Experiment 3). Even if this would be the case there would still be two kinds of test strategies, which either lead or do not lead to frequency effects. The postulated differential effect of frequency information on these two strategies was tested and corroborated here for the first time. These two test strategies should at least normally be linked either to descriptive or prescriptive tasks.

Coming to the other point, as has been seen in Experiment 8 there may be frequency effects for prescriptive tasks as well. It has been argued that these frequency effects can be rational, at least under particular circumstances. However, most of these effects will in my opinion normally only become relevant, if salient frequency information presses participants to reinterpret the task in a way that allows them to make use of frequency information (see discussion of Experiment 9.) More important, it should be summarised that not only may there be rational frequency effects also for prescriptive tasks as well, but that these effects need to be distinguished from the frequency effects in our descriptive tasks (here the Sydow model, cf. 5.1):

(a) *Frequency effects based on reinterpretation*. If the prescriptive task is interpreted as a descriptive task (in order to make sense of the frequency information), the same frequency effects are, of course, predicted for both tasks. But in this case, it should be possible to provide independent evidence for such a reinterpretation, for instance, by asking the participants what they aimed to achieve by their selections.

(b) *Within-focus frequency effects.* FDL theory is coherent with *p* versus *non-q* frequency effects for prescriptive rules within a fixed focus, for instance a *p* & *non-q* focus. Within-focus frequency effects should only be effective in a forced choice between *p* and *non-q* selections, both referring to the same focus. In contrast, for instance *q* versus *non-q* frequency effects cannot be explained by within-focus effects.

(c) *Frequency effects based on sorting.* Frequency effects in prescriptive rules may be based on what I have called 'sorting' (cf. Discussion of Experiment 8). Sorting is, for instance, checking for cheaters by selecting all of the few cooperators, in order to punish the rest. However, in this case, *p* versus *non-p* choices should be unaffected by frequency manipulations, but *q* versus *non-q* choices should be affected. This was exactly the pattern found in Experiment 8.

(d) *Frequency effects based on changed goals.* As has been outlined in detail before, frequency may influence the goal of cheater or cooperator detection. Such a tendency may even have an effect against goals specified *a priori* in the instruction, if there are reasons to question these aspects of the instruction (cf. Discussion Experiment 8). For this type of frequency effects again only *q* versus *non-q* effects and no *p* versus *non-p* effects should be found, as it was actually the case in Experiment 8.

All four frequency effects possible in prescriptive task need to be delineated and investigated in future research using independent evidence for these strategies. As predicted, Experiment 9 showed that if the outlined interpretations are excluded, frequency effects are only efficient in the descriptive task and not in the prescriptive task. In the light of these results the partial suppression of frequency effects found Experiment 8 also matches exactly which the refined dichotomy.

Overall, in regard of the tested frequency effects the results support a refined dissociation between descriptive rules and prescriptive rules with an a priori focus.

Empirical Results and Frequency Effects in the Literature

How do these results relate to other empirical findings in the literature? It will be shown that the above tasks provide the first explicit test of two test strategies in the WST in regard of frequency effects.

I have argued earlier that the results of other authors may partly be due to the fact that they involuntarily may have tested the postulated contrast in sensitivity for frequencies of descriptive or prescriptive WSTs (e.g., Cosmides 1989, Exp. 1, 2; Gigerenzer, Exp. 1). For these experiments it seems plausible that the found increase of *p & q* selections in what I call 'descriptive tasks' may be due to the default rarity assumption. It is also plausible that in the used 'prescriptive' WSTs frequencies have played no role for the cheater detection task (since no reason pressed the participants to doubt that the focus was *a priori* fixed). But since the frequencies of the entities mentioned were not controlled or varied in these experiments, this cannot count as a direct test of the dichotomy advocated here.

I am not aware of any earlier experiments in the WST literature which directly introduced frequencies in social WSTs and at the same time varied the descriptive versus prescriptive status of the task.

Love and Kessler (1995, Exp. 1b) indeed tested both descriptive and prescriptive rules and at least intended to manipulate the probabilities of finding a violating case. However, in their experiment the frequency of cards was not varied. Instead, only the story was manipulated changing the plausibility of counterexamples. They were not interested in testing the dichotomy tested here. Moreover, Love and Kessler's (1995, Exp. 1b) instruction is in my view confounded with introducing a second cheater case.[56]

We may also have a look at experiments on frequency effects investigating them independently either for descriptive or prescriptive WSTs. There are of course many

[56] In their two deontic obligation conditions the social rule read 'if a man eats cassava root, then he must have a tattoo on his face'. The participants were instructed in both deontic conditions that only married man are allowed to have tattoos on their face. In the condition 'low probability of a counterexample' they added that bachelors can also easily put a tattoo on their face and only few of them do not manage to get it done. But in my view, bachelors who pretended to be husbands by putting a tattoo on their face may well be understood as cheaters, rendering also the *p & q* cell a possible cheater cell. Because of this confound the shown reduction in the number of *p & non-q* selection patterns (anyway still 43 % selected this pattern) cannot be attributed unequivocally to the kind of probability manipulation they used.

articles on frequency effects for descriptive WSTs and in Part II we have discussed in detail why many results were not as coherent with a Bayesian approach as the results obtained here. Frequency manipulation in deontic task has been investigated only in a few cases (Manktelow, Sutherland & Over 1995; Kirby, 1994, Exp. 4). But in the mentioned deontic WSTs the card frequencies were not properly varied. In my view, the experiments mentioned were actually concerned with utility effects. In any case, both the papers mentioned were only concerned with effects *within* the focused *p* and *non-q* cases. Hence, they did not predict, test or find probability effects in the sense as have been tested here.

In conclusion, the two experiments reported here are the first direct tests of the postulated difference between the sensitivity of descriptive and prescriptive tasks for frequency effects – they will not be the last ones.

How Do the Findings Relate to Other Theories of the WST?

The results are either in contradiction to previous other theories of the WST, or they point out their incompleteness.

(a) *Pragmatic reasoning schema theory* (PRS theory, see pp. 13, 176, 270; Cheng & Holyoak, 1985; Holyoak & Cheng, 1995) would have predicted the predominance of *p* and *non-q* selections in the prescriptive task conditions (obligation schema) found in Experiment 9, but cannot account for the frequency effects for the deontic task in Experiment 8. Moreover, the claims that PRS theory also made about descriptive tasks are inconsistent with the found frequency effects in the corresponding conditions (cf. General Discussion of Part II).

(b) Likewise, *social contract theory* (SC theory, see pp. 14, 264; Cosmides, 1989; Cosmides & Tooby, 1992; Gigerenzer & Hug, 1992) cannot explain any frequency effect in the WST (see pp. 175 f.). The results are at odds with the claim that an activation of cheater detection is needed, in order to achieve reliable deviations from standard selections. Moreover, although SC theory was mainly concerned with prescriptive rules, proponents of this theory have also claimed that no reliable and systematic content effects is to be found in testing descriptive tasks, instead predicting random *p* or *p & q* selections (see pp. 175 f.). Here the found pattern was not random but largely coherent with the Bayesian predictions. In any case, social contract theory is not able to account for present data.

(c) *Mental model theory of the WST* (MM theory, see pp. 11, 173, 272; Johnson-Laird & Byrne, 1991, 2002) has never given up the claim of being a completely domain-general theory, even when concerned with deontic WSTs (Johnson-Laird & Byrne, 1995). Hence, MM theory cannot distinguish between to kinds of test strategies. However, MM theory may try to explain the more logical selections in the predictive rule conditions by postulating that the instructions in this condition elicited a fleshing out the mental model. However, MM theory cannot account for any of the found frequency effects, which were here caused purely by probabilistic manipulations (cf. General Discussion of Part II for details, pp. 173 f.).

(d) *Mental logic theory* (cf. pp. 10 f.; Rips, 1994; O'Brien, 1995) does not predict any frequency effects (cf. pp 172 for details). Although mental logic theory in principle might be extended to deontic logic the two test strategies distinguished here, in my view, cannot not be harmonised with mental logic theory. In any case current mental logic theory can neither account for the found differences between the two types of tasks nor for any frequency effect in Experiment 8 and 9, based purely on probability manipulations.

(e) *Decision theoretic accounts of the WST* (Manktelow & Over, 1990, 1991, 1992, 1995; Evans, Over, & Manktelow, 1993; Over & Manktelow, 1993; Evans & Over, 1996) have not predicted any differences in testing prescriptive or descriptive rules. Manktelow and Over (1991, 1995) indeed were the first to point out the difference in the *result* of testing prescriptive and descriptive rules – only the latter can become falsified. FDL theory owes much to this distinction. However, Manktelow and Over (1991, 1995) did not predict different test strategies following this distinction (cf. Fairley, Manktelow, & Over, 1999). However, the investigated distinction is in my view not in principle at odds with an extended decision theoretic approach (cf. General Discussion of Part III).

(f) *Relevance theory of the WST* (Sperber, Cara & Girotto, 1995; Sperber & Girotto, 2002, 2003) might be understood in a quite general and almost tautological sense (cf. General Discussion of Part III, pp. 281). I would agree that frequency information changes the relevance of cards, but this is exactly the claim of any Bayesian account. In this sense, any theory of the WST (a Bayesian Account, a Mental Model account etc.) could be incorporated by relevance theory. Since this would be absurd, more specific predictions of relevance theory need to be considered. More specifically, relevance theory has predicted focus effects (Girotto, 2002, 2003).

But only FDL theory combines these focus effects with deontic logic (see Chapter 12). Moreover, relevance theory can only provide an unrefined and post hoc explanation for frequency effects in descriptive tasks (cf. Oaksford & Chater, 1995). Hence, in my view, relevance theory can neither account for any of the frequency effects found, nor for the found dissociation (cf. General Discussion of Part III).

Only Oaksford and Chater (1994) have pre-empted some aspects of the dichotomy of test strategies postulated and tested here. For prescriptive rules, they advocated a decision theoretic approach with weights for different cells and contrasted this to their Bayesian approach for descriptive tasks. Nevertheless, they have neither explicitly predicted a selective frequency effect, nor tested the dissociation supported here. Hence, I think their analysis is in principle coherent with the reported findings of the Experiments in this Chapter. I am not opposed to other implications of a decision theoretic account (for more details cf. General Discussion of Part III). However, also in regard of the following experiments, Oaksford and Chater (1994) neither proposed a deontic logic of different conditionals and nor combined this with a flexible focus based on cheater or cooperator detection.

In conclusion, no theory has explicitly predicted the found dissociation of testing prescriptive or descriptive tasks. Only the analysis by Oaksford and Chater in principle entails this prediction, but also in that case the found dichotomy has neither explicitly been predicted nor tested. Mental logic theory, mental model theory, pragmatic reasoning schema theory, and social contract theory are all inconsistent with the findings of this Chapter. Some other theories, like relevance theory and the decision theoretic approach may perhaps be *extended* to account for the findings.

11 Second Aspect of FDL Theory: The Deontic Logic of the WST

This chapter is concerned with the postulate of FDL theory that selections in a WST with prescriptive conditionals can be based on four types of deontic conditionals, based on deontic logic. Here the conditionals should be tested in a WST (cf. Beller, 2001, 2003; Bucciarelli & Johnson-Laird, 2005). To my knowledge, this is the first test of all four types of conditionals in a WST on equal footing (cf. Section 9.2). The predictions of other theories will be discussed later on.

Experiment 10 – Deontic Logic of the WST

The goal of this experiment was to elicit *four* different selection patterns corresponding to the four different forbidden cells in the ought tables of the respective deontic conditionals: conditional prohibition, conditional obligation, conditional permission, and conditional permission to refrain. In this experiment only we varied the ought table, not the focus. In all conditions a cheater detection focus should be induced.

Method of Experiment 10

Design and Participants

The experiment had a between-subjects design with four conditions, corresponding to the four types of conditionals, each with a different forbidden ought cell. Sixty-four students from the University of Göttingen (63 % female, 37 % male, mean age: 24 years) voluntarily took part in the experiment. Participants came from different departments; most of them have studied economics (23 %), arts (17 %), and law (17 %). As a reward, participants got a little present at the end of the experiment. The participants were randomly assigned to the four conditions.

Materials and Procedure

The deontic Wason selection tasks were carried out as paper and pencil tasks. The instructions were in German. In all conditions, participants were asked to imagine they were members of a council of elders, which had police functions. The council's purpose, they were told, was to punish those who violated the rules of the tribe. Then in each of the four conditions, one of the following four conditional tribal rules was presented (cf. Table 49 to Table 52):

- Conditional prohibition (translated from German): "If someone is a bachelor, he is forbidden from going to the bath house".
- Conditional obligation: "If someone is a bachelor, then each month he must bring fish to the medicine man".
- Conditional permission: "If someone is a bachelor, then he is allowed to eat the aphrodisiac Cassava root".
- Conditional permission to refrain: "If someone is a bachelor, he may refrain from taking part in hunting the dangerous Karogi oxen".

The rules were novel and unfamiliar to ensure that no prior experience directly with the rule was available.

The rest of the instruction was almost identical in all conditions. They only differed in the description of the cards, which corresponded to each rule. In all conditions four male members of the tribe were presented to the participants for possible checks. The tribesmen were represented by four cards. It was explained that one side of each card provided information about whether each tribesman was a bachelor or not, and that the other side provided information about whether he went into the bathhouse (or: brought fish to the medicine man; or: ate Cassava root; or: took part in hunting Karogi oxen) or not. In all conditions, only one side of each card was shown and the cards read as follows.

- Conditional prohibition condition: “Bachelor” (*p*), “Husband” (*non-p*), “Goes to the bath house” (*q*), and “Does not go to the bath house” (*non-q*).
- Conditional obligation condition: “Bachelor” (*p*), “Husband” (*non-p*), “Brings fish” (*q*), and “Does not bring fish” (*non-q*).
- Conditional permission condition: “Bachelor” (*p*), “Husband” (*non-p*), “Eats Cassava root” (*q*), and “Does not eat Cassava root” (*non-q*).
- Conditional permission to refrain condition: “Bachelor” (*p*), “Husband” (*non-p*), “Takes part in the hunt of Karogi oxen” (*q*), and “Does not take part in the hunt of Karogi oxen” (*non-q*).

Participants had to decide which card(s) are “really needed to be turned over to test whether the rule had been followed or had been violated”. They were requested to indicate all cards necessary to fulfil the given task.[57]

Results of Experiment 10

The Two Levels of Analysis

In this and the following experiments the results are presented on two levels of analysis, the level of card combinations and the level of single card selections. The level of card combinations allows the strongest test possible. On this level, the portion of participants can be tested producing exactly the pattern predicted by FDL theory. This tests whether participants exclusively and exhaustively select all cards predicted

[57] Although the final instruction did not explicitly formulate a cheater detection focus, it was assumed that this focus is generally elicited by the police function of the council, mentioned at the beginning of the task.

by FDL theory. [58] The standard level of single cards can be used as a weaker – but also relevant – test of whether the card selections are influenced in a way predicted by FDL theory. For instance, if we would find *q* selections instead of a predicted *p* & *q* pattern and, in another condition, *non-q* selections instead of a predicted *p* & *non-q* pattern, this would completely count against FDL theory on the level of card combinations, and only the level of single card combinations could reveal that such results would be partially confirmatory.

Selections of Card Combinations

Table 58 and Figure 16 descriptively show the number of participants selecting a *combination* of cards. The dominant selections for each type of deontic conditional corresponded to the four forbidden cells of their respective ought tables and hence to the predicted selection patterns:

Table 58
Selections of Card Combinations for Four Deontic Conditionals in Experiment 10 (N = 64)

Pattern selected	Cond. pro-hibition		Cond. ob-ligation		Cond. perm. to refrain		Cond. per-mission	
P, *Q*	81 %	13	6 %	1	12 %	2	12 %	2
P, ¬*Q*	0 %	0	56 %	9	12 %	2	0 %	0
¬*P*, ¬*Q*	0 %	0	0 %	0	69 %	11	6 %	1
¬*P*, *Q*	6 %	1	0 %	0	0 %	0	56 %	9
P	0 %	0	6 %	1	0 %	0	0 %	0
¬*P*	0 %	0	6 %	1	0 %	0	0 %	0
Q	0 %	0	0 %	0	0 %	0	18 %	3
¬*Q*	6 %	1	6 %	1	0 %	0	0 %	0
P, ¬*P*	0 %	0	0 %	0	0 %	0	6 %	1
Q, ¬*Q*	0 %	0	0 %	0	0 %	0	0 %	0
P, *Q*, ¬*Q*	6 %	1	12 %	2	0 %	0	0 %	0
P, ¬*P*, *Q*	0 %	0	0 %	0	0 %	0	0 %	0
P,¬*P*, ¬*Q*	0 %	0	0 %	0	6 %	1	0 %	0
¬*P*, *Q*, ¬*Q*	0 %	0	0 %	0	0 %	0	0 %	0
P, ¬*P*, *Q*, ¬*Q*	0 %	0	6 %	1	0 %	0	0 %	0
No card	0 %	0	0 %	0	0 %	0	0 %	0
n		16		16		16		16

Note. Percentage and number of participants selecting each card pattern. Predicted answers in darkened cells.

The most frequent selection pattern in the *conditional prohibition* condition was *p* & *q*, in the *conditional obligation* condition *p* & *non-q*, in the *conditional permission* condition *non-p* & *non-q* and in the *conditional permission to refrain* condition *non-p* & *q*.

For each relevant card pattern, it was tested whether the frequency of its selection was higher in the conditions, in which this corresponding pattern had been predicted, than in

[58] The card combination level was normally not discussed for analysing FBL theory. This had two reasons. First, strictly speaking FBL theory makes no predictions for successive selections of many cards. This is not the case for FDL theory. Secondly, in previous experiments separate forced-choice selections were used, instead of a free choice out of four cards.

all other conditions. For each of the four comparisons the selection patterns of the three contrasted conditions, which had all yielded similarly low results, were collapsed across conditions.

The *p & q* pattern was higher in the prohibition condition than in all the other three corresponding conditions taken together ($N = 64$, Fisher exact test: $p < .001$). Relative to the respective alternative conditions the *p & non-q* pattern was shown to be higher in the obligation condition (Fisher exact test: $p < .001$), the non-*p & non-q* pattern higher in the permission to refrain condition (Fisher exact test: $p < .001$) and the non-*p & q* pattern higher in the permission condition (Fisher exact test: $p < .001$). All tests were highly significant.

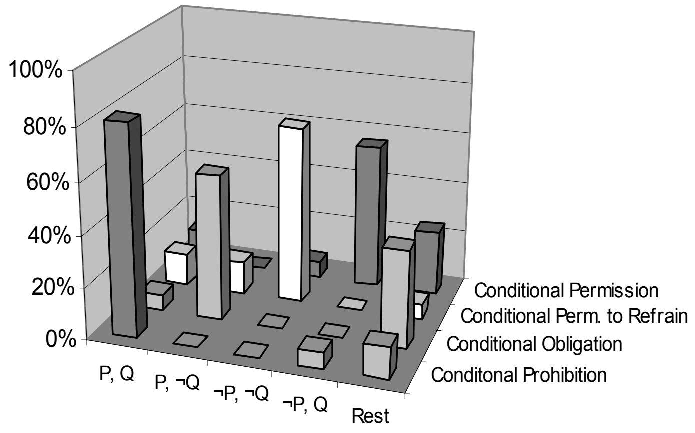


Figure 16. Bar graph of the percentage of participants selecting card combinations in checking the four types of conditionals in Experiment 10.

Selections of Specific Cards.

Table 59 shows the results of *single* card selections in the four WSTs. Regardless of whether we compare between or within conditions, the predicted selections were descriptively predominant.

To test the predictions statistically a hierarchical loglinear analysis for each card was carried out. To allow for an exact test of all predicted patterns the four conditions where grouped to form two factors. The newly coded 'antecedent' factor consists of the prohibition and obligation conditions versus the permission and refrain conditions. If FDL theory is correct, this 'antecedent' factor should determine whether participants tend to select *p* or *non-p* cards. Hence, for the loglinear analysis the first order interactions '*p* card × antecedent' and '*non-p* card × antecedent' are to be predicted. With respect to the *q* and *non-q* cards, no main effect or interaction effect of this factor is predicted.

Table 59
Selections of Single Cards for Four Deontic Conditionals in Experiment 10 (N = 64)

Card selected	Cond. prohibition	Cond. obligation	Cond. perm. to Refrain	Cond. permission
P	87 % 14	87 % 14	19 % 3	31 % 5
¬P	6 % 1	12 % 2	69 % 11	75 % 12
Q	93 % 15	25 % 4	87 % 14	12 % 2
¬Q	12 % 2	81 % 13	12 % 2	87 % 14
n	16	16	16	16

Note. Percentage and number of participants selecting each card. Predicted answers in darkened cells.

The factor 'consequent' contrasts the prohibition and permission to refrain conditions against the obligation and permission conditions. For this factor only first order interactions with selections of *q* cards and *non-q* cards are predicted (*q* card × consequent and in an opposed direction, a *non-q* card × consequent). Analyzing each card separately, no other main effect or interaction effect (including no second order effect of card selection × antecedent × consequent) is predicted.

The analysis was carried out using SPSS. For each card, we started with the saturated model with the three factors 'card selection (2) × antecedent (2) × consequent (2)'. For the *p-selection frequencies* hierarchical backward elimination ($p < .05$) in 5 steps led to a final model with the predicted generative class of antecedent × *p*-card. Likewise, the test that specific levels of effects ($1 \geq k \geq 3$) differ from zero only yielded significant results for the expected first order interaction level ($k = 2$, Pearson $\chi^2(3) = 25.65$, $p < .0001$), and not for the second order interaction level ($k = 3$, $\chi^2(1) = .59$) or the level of main effects ($k = 1$, $\chi^2(1) = .25$). For the saturated model we also estimated the eight single λ parameters and checked which of these parameters significantly differed from zero. Consistently the only parameter to become significant was the parameter 'antecedent × *p* card' (λ = -.77, z = -4.53, p < .001). It can be concluded that the *p*-card selections in the prohibition and obligation rule conditions were more frequent than in the refrain and permission conditions. As also expected, the results were affected by no other significant interaction or main effect parameter.

For the *non-p* card selections the iterative elimination, again in five steps, led to the predicted generating model class of 'antecedent × *non-p* card' (p < .05). Correspondingly, only the level of the first order interaction, containing the interaction 'antecedent × *non-p* card', was significant (k = 2, $\chi^2(3)$ = 26.09, p < .0001). Tested individually, only the parameter 'antecedent × *non-p* card' of the saturated model reliably differed from zero (λ = .82, z = 4.34, p < .001). Thus, whether a *non-p* card is selected or not is due to the predicted contrast of prohibition and obligation conditions versus refrain and permission conditions.

For *q* selections backward elimination confirmed the generating model class 'consequent × *q* card' (5 steps, p < .05). Only the predicted first order interaction level (k = 2, $\chi^2(3)$ = 33.98, p < .0001) significantly differed from zero, as did, more specifically, the predicted parameter 'consequent × *q* card' (λ = -.96, z = 4.82, p < .001). As expected, the *q*-card selection was dominant in the prohibition and refrain conditions, whereas in the obligation condition and the permission condition this card was normally not selected. No other effect was found.

Finally, backward elimination for the *non-q*-selections resulted in the predicted generating model class of consequent × *non-q*-card (5 steps, p < .05). Consistent with our expectations, only the level of first order interactions (k = 2, $\chi^2(3)$ = 33.11, p < .0001) and only the corresponding parameter of consequent × *non-q*-card (λ = -.91, z = 5.01, p < .001) reliably differed from zero.

Overall, there was a remarkable fit between the predictions of FDL theory and the empirical data. The results were all highly significant and there was no single significant deviation from our predictions.

Discussion of Experiment 10

The results of this experiment provide strong evidence that selection patterns are determined by deontic logic. The dominant selection patterns always corresponded to the forbidden cell of its respective ought table. Although the rules checked were all formulated in terms of a conditional, the differences between the conditions confirmed the four different kinds of card selections predicted. To the knowledge of the author this is the first experiment with deontic WSTs in which a parallel formulation of all four different types of deontic conditionals has led to a systematic variation of the cheater detection focus on all four cells of an ought table, including *p* and *q* patterns and *non-p* and *non-q* patterns. Only Beller, in an interesting paper, (2001) has provided similar results with WSTs; but in the reported experiment he did not at that time emphasize the equal footing of all four conditionals; instead, he worked with affirmative and negated obligations and permissions.

Although the results of Experiment 10 are indeed based on cheater detection (e.g., Cosmides, 1989), here cheater detection is not understood to be illogical. On the contrary, cheater detection is here understood to be based on the systematics of deontic logic. This deontic logic approach to deontic WSTs contradicts current domain-specific approaches, which have advocated specific schemas or Darwinian algorithms and not a general systematics of deontic logic for the prescriptive realm. Cosmides claimed explicitly that “all social contracts are either permission or obligation rules” (Cosmides 1989, 235). Combined with her claim that only social contracts lead to clear content effects this clearly contradicts the predictions of FDL theory that, for instance, a conditional permission to refrain rule will lead to clear-cut *non-p* and *non-q* patterns in a WST. Hence, the results of Experiment 1 differ from the evolutionary social contract theory of Cosmides and colleagues, which has explicitly broken with any logical foundation and which has not differentiated between different logical deontic connectors.

But the deontic logic theory advocated, as well as the results of this experiment, are at odds with pragmatic reasoning schema theory (Cheng & Holyoak, 1985, 1989), since that theory was presented as a set of pragmatic reasoning schemas (obligation and permission schema), and not – as it is proposed here – as a system of pragmatic reasoning schemas based on deontic logic. Holyoak & Cheng (1995b) indeed made a step in the direction of reformulating their theory in more general terms of complementary rights and duties. FDL theory of the WST indeed may perhaps be consistent with an approach based on rights and duties. However, Holyoak and Cheng (1995b) did not themselves elaborate and test a deontic logic. Consequently, pragmatic reasoning schema theory did not come up with the prediction of a schema for a prohibition and a conditional permission to refrain. Thus, deontic logic theory and the confirming results of Experiment 1 could be interpreted either as inconsistent with pragmatic reasoning schema theory, or as extensions of such a theory, conciliating it with the much older deontic logic (cf. general discussion).

Furthermore, the results of this experiment are inconsistent with the traditional domain-general theories of the WST, since these theories have not applied deontic logic to the WST. Nonetheless, it may be possible that these theories will be accommodated to account for such findings. Indeed some first steps towards a deontic logic approach for the WST have been made from this side (cf. e.g., Manktelow & Over, 1991, 1995; Johnson-Laird & Byrne, 1992, and lately in the different context of syllogistic reasoning: Bucciarelli & Johnson-Laird, 2005; cf. pp. 275 f.). In the General Discussion, we will consider these approaches in detail. In any case, in the context of the WST it was previously argued, tested and confirmed that in checking prescriptive conditionals all four cells of an ought table can systematically become a focus of cheater detection.

In the next Chapter, the FDL theory claim will be tested that even this deontic logic of cheater detection is not general enough. It shall be shown that human rationality is even more flexible and that cheater detection is only one kind of goal-dependent focusing on particular cells of an ought table.

12 Third Aspect of FDL Theory: Focus Effects Based on Deontic Logic

This chapter is aimed at providing some evidence for the third central prediction of flexible deontic logic theory that in testing prescriptive rules people focus flexibly on particular cells of induced ought tables. It is expected that in a context of punishment of rule violators, participants usually focus on forbidden cases and engage in cheater detection. In contrast, a context of reward is expected to elicit a search for rule followers and a corresponding cooperator focus. This focus mechanism is here combined with deontic logic. Here two experiments on this matter are presented.

In Experiment 11 the two factors of a cooperator versus a cheater focus and a conditional obligation rule versus a conditional prohibition rule will be varied, aiming to show an interaction of both factors. Here the goals were varied only by changing the role description of the tester. This Experiment was part of a *Diplomarbeit* of

Nicole Metzner, which has been supervised by the author (Metzner, 2005). The experiment has been initiated by the author and it was planned together. It has first been published in a joint publication (cf., v. Sydow, Hagmayer, Metzner, Waldmann; 2005). Because this experiment was designed as a central tests of FDL theory by the author, its main results (using an own loglinear analysis) are reported here (for more details cf.: Metzner, 2005; cf. also Anke Gummelt, 2005).

Experiment 12 provides a test of the interaction of focus and rule with a slightly changed procedure, and additionally introduces double focus conditions in order to elicit a simultaneous goal of cheater *and* cooperator detection. If this would result in the predicted novel double focus effects this would be a strong additional argument in favour of FDL theory and against alternative interpretations of the interaction effects.

Before I am going to present the experiments, I want to outline the similarities and differences of our experiments to former WSTs in the extensive literature on this topic.

12.1 Similarities and Differences to Former WSTs

In the WST tradition much research has been done on subtle manipulations of the task instructions. Some of these manipulations on the first view resemble the manipulation of the following Experiment 11.

However, most of that research was concerned with instructions which emphasize falsifying cases, counter examples or cases of a logical violation of a rule. Mental model theory assumes that such modifications lead to a more complete representation of the situation (e.g., Johnson-Laird & Wason, 1970b; van Duyne, 1974; Jackson & Griggs, 1990, cf. also the reply by Kroger, Cheng & Holyoak, 1993; Johnson-Laird & Byrne, 1991, 80-82; Platt & Griggs, 1993; Green, 1995; Manktelow, Sutherland & Over, 1995, Exp. 2, 3; Fiedler & Hertel, 1994; Love & Kessler, 1995; Handley, Feeney, & Harper 2002). From the perspective of FDL theory some of these experiments would presumably need to be reinterpreted, since it is not clear what has been tested, the representation of the situation or the test focus. Also experiments on obligation rules in which more *p* and *non-q* patterns were elicited either by a rationale for the test (Cheng & Holyoak, 1985; Manktelow, Sutherland & Over, 1995, Exp. 2, 3) or a cheater detection motivation (Cosmides, 1989, Exp. 5-7; Gigerenzer & Hug, 1992, Exp. 2; Kirby, 1994, Exp. 4; Love & Kessler, Task 4 versus Task 5) may be reinterpreted by FDL theory as experiments where a cheater focus was elicited in one condition and the focus was absent (without a clear alternative focus) in another condition.

Other experiments, which we have discussed in Chapter 10, involuntarily or voluntarily investigated the difference between a cheater detection instruction for a prescriptive task and the true-false instruction of descriptive tasks (e.g., Cosmides, 1989, Exp. 1, 2; Gigerenzer & Hug, 1992, Exp. 1; Platt & Griggs, 1993, Exp. 3).

Moreover, from the viewpoint of FDL theory research on perspective effects may perhaps also be interpreted as focus effects. But they are only concerned with effects *within* different cheater detection cells, not with a cooperation focus. To decide whether perspective effects can be interpreted as a particular case of focus effects, further research would be needed. Former research has not distinguished between different cheater foci within a complete representation or different constructions of the rule as a permission or an obligation rule in the first place (cf. e.g., Johnson-Laird & Byrne, 1991, 78-79, 1992, 1995; Manktelow & Over, 1991; Gigerenzer & Hug, 1992; Politzer, & Nguyen-Xuan, 1992; Holyoak & Cheng, 1995a, 1995b; Fairly, Manktelow & Over, 1999; Almor & Sloman, 2000; Staller, Sloman, & Ben-Zeev, 2000). However, we will here be concerned with an explicit cheater versus a cooperator focus.

Only a few experiments have made use of something like alternative cheater and cooperator focus instructions (Manktelow & Over 1990, Love & Kessler, 1995; Sperber & Girotto 2002, 2003). The resulting focus effects were either not intended by the authors, interpreted in a different way, or the effects were not very pronounced (cf. the discussion of this experiment). The week effects might perhaps be due to the need of a fit between goal of the task and focused cells. The goal of cooperator detection needs to be equally plausible in that situation as the goal of cheater detection. Therefore, in Experiment 11 a context was used in which a social system of sanctions is plausible, both in regard of punishment and in regard of gratification. Moreover, in Experiment 11 the goal of the imagined role was manipulated and not only the formal instruction at the end of the task.

Furthermore, until now no one has yet investigated a systematic effect of a cooperator versus cheating focus instruction combined with testing deontic conditionals with different ought tables. To my knowledge, Experiment 11 is the first experiment that combines deontic logic with focus effects.

Experiment 12 can even more easily be distinguished from former experiments in the WST tradition. Here double focus conditions provide a novel test (or control) condition for FDL theory, simultaneously eliciting the two goals of cooperator and cheater detection. The double focus effects, which should be elicited here, have to be distinguished from any setting intended to investigate perspective effects, since our double focus effects are concerned with the simultaneous focus on a cheater *and* a cooperator cell, and not with a specific cheater perspective. Hence, above the novel aspects already mentioned for Experiment 11, the double focus conditions of Experiment 12 are not found in any former experiment on WSTs.

12.2 Experiment 11 – Cooperator Detection and Cheater Detection for Different Deontic Conditionals (with Nicole Metzner[59])

Method of Experiment 11

Design and Participants. The experiment had a 2 (obligation vs. prohibition rule) × 2 (cheater vs. cooperator focus) between-subjects design. Eighty students from the University of Göttingen (48 female, 32 male; mean age: 24 years) voluntarily participated in the experiment. The largest groups of participants studied law (26 %), economics (16 %) or social sciences (11 %). The participants got a little present, and were randomly assigned to the four conditions.

Materials and Procedure. In all four conditions deontic WSTs were used. Like in Experiment 11 participants were asked to imagine they were members of a council of elders. Now the goal of the council was varied. The council had to check whether members of a tribe have either violated or followed the laws of the tribe (cheater versus cooperator conditions).

In two obligation rule conditions, the council of elders had to check the rule: “If someone is a bachelor, then he must abduct a virgin from a hostile and dangerous tribe”. In the two prohibition rule conditions, the following rule was to be tested: “If someone is a bachelor, then he is forbidden from fleeing from a battle, which is about to be lost.” Both rules were novel and unfamiliar to ensure that no direct prior experience with the rule was available. Moreover, both rules were culturally alien and

[59] See the introduction of Chapter 12 (cf. Metzner, 2005; v. Sydow, Hagmayer, Metzner, & Waldmann, 2005).

it can be assumed that most participants did not approve to such rules personally. Nevertheless, the schema-based FDL theory predicts that participants are able to put themselves into the position of someone who has to check such deontic rules.

The goals which, according to FDL theory, should influence which cells are focused were manipulated by assigning different responsibilities to the council of elders. In cheater detection conditions of the two rules, participants were told, “The council of elders is responsible for law enforcement. The task of the council is to punish those who had violated the laws of the tribe.” In the two cooperator detection conditions, participants were instructed, “The council of elders each year decorates members of the tribe with honour feathers. The task of the council is to honour those who followed the laws of the tribe.”

In all conditions, four male members of the tribe were presented to the participants for possible checks. These four tribesmen were represented by four cards. Like in Experiment 10 participants were instructed that on one side of each card information was given about whether the man was a bachelor or not, and on the other side whether he has abducted a woman (or has fled from a battle) or not. The cards read: “Bachelor” (*p*), “Husband” ($\neg p$), “Virgin abducted” (or: “Has fled from a lost battle”; *q*), and “No virgin abducted” (or: “Has persisted in a lost battle”; *non-q*). Participants had to decide which card(s) they really needed to turn over to test whether the rule had been followed or violated. They were requested to indicate all cards necessary to complete the given task.

Results of Experiment 11

Selections of Card Combinations

For each condition Table 60 shows the percentage of participants who selected particular card patterns. The predicted answers have been darkened. As predicted, the goal of the council of elders in interaction with the type of rule strongly influenced the preferred pattern of selected cards.

In order to test the interaction effects we carried out a hierarchical loglinear analysis with three factors for each card pattern: card pattern (2) × rule type (2) × goal (2). For the analysis we used SPSS.

Let us have a look at the *p & non-q pattern* first. As predicted, backward elimination ($p < .05$) retained the generative model class of a second order interaction term ‘card pattern × rule × goal’ (saturated model). For the saturated model the analysis of *single levels* consistently showed the corresponding second order interaction level ($k = 3$) to be highly significant (Pearson $\chi^2(1) = 13.08$, $p < .001$). Thus also confirming the hypotheses of FDL theory, there was no significant first order interaction level ($k = 2$, $\chi^2(3) = 1.32$, $p = .72$). A significant effect on the general level of main effects ($k = 1$, $\chi^2(3) = 16.4$, $p < .001$) indicated over all conditions a predicted predominance of other card patterns than *p & non-q*. This interpretation is warranted by the next step of analysis. The even more specific tests of the *single parameters* of the saturated model showed that only two parameters significantly differed from zero: the predicted second order interaction term ($\lambda = -.57$, $z = -3.18$, $p < .01$) and the particular main effect term of the pattern *p & non-q* ($\lambda = .72$, $z = 4.05$, $p < .001$).

Table 60
Selections of Card Combinations for Different Rules and Different Foci in Experiment 11 (N = 80)

Card pattern	Obligation rule				Prohibition rule			
	Cheater detection		Co-operator detection		Cheater detection		Co-operator detection	
P, ¬*Q*	50 %	10	5 %	1	10 %	2	35 %	7
P, *Q*	5 %	1	55 %	11	70 %	14	35 %	7
P	10 %	2	20 %	4	5 %	1	0 %	0
¬*P*	5 %	1	0 %	0	0 %	0	0 %	0
Q	0 %	0	0 %	0	0 %	0	5 %	1
¬*Q*	5 %	1	0 %	0	0 %	0	15 %	3
P, ¬*P*	0 %	0	5 %	1	5 %	1	5 %	1
¬*P*, *Q*	5 %	1	0 %	0	0 %	0	0 %	0
¬*P*, ¬*Q*	5 %	1	0 %	0	0 %	0	0 %	0
Q, ¬*Q*	0 %	0	0 %	0	0 %	0	0 %	0
P, ¬*P*, *Q*	5 %	1	5 %	1	0 %	0	0 %	0
P, ¬*P*, ¬*Q*	5 %	1	0 %	0	0 %	0	0 %	0
P, *Q*, ¬*Q*	0 %	0	10 %	2	5 %	1	5 %	1
¬*P*,*Q*, ¬*Q*	5 %	1	0 %	0	0 %	0	0 %	0
P, ¬*P*, *Q*, ¬*Q*	0 %	0	0 %	0	5 %	1	0 %	0
-	0 %	0	0 %	0	0 %	0	0 %	0
n		20		20		20		20

Note: Percentage and number of participants selecting each card pattern. Predicted answers in darkened cells.

Likewise, the analysis of *p & q* patterns showed that the predicted second order interaction term was not eliminated in backward elimination. The analysis of the different levels showed that only the predicted level of the second order interaction ‘card pattern × rule × goal’ ($k = 3$, $\chi^2(1) = 16.47$, $p < .0001$) was significant. As expected there were no first order interactions ($k = 2$, $\chi^2(3) = 3.08$, $p = .37$) or main effects ($k = 1$, $\chi^2(3) = 1.85$, $p = .60$). The testing of each particular parameter of the saturated model showed that only the predicted second order interaction ($\lambda = .58$,

$z = 3.52$, $p < .001$) and a first order interaction between *p & q* pattern and rule type ($\lambda = .37$, $z = 2.27$, $p < .05$) reached significance. The latter parameter is unpredicted, but its interpretation is problematic, since the whole level ($k = 2$) had not been significantly different from zero. However, this deviation from the predictions seems to refer to more *p & q* selections in the prohibition rule conditions than in the obligation rule conditions. This may be due to the phrasing of the prohibition. Terms such as ‘forbids’ might be less open for a cooperator focus than terms such as ‘must’, which was used to phrase the obligation rule.

For each condition the remaining 14 alternative unpredicted selection patterns (*p; ¬p; q; ¬q; p & ¬p; p & ¬p & q*; etc.) were collapsed into a single category ‘alternative patterns’. For this resulting variable backward elimination reduced the model to a term referring only to the predicted main effect ‘alternative patterns’. This only shows that even if taken together these patterns were significantly more often absent than present. The further calculations led to the similar results. Only the main effect level ‘alternative patterns’ ($k = 1$, $\chi^2(3) = 8.10$, $p = .04$) and the corresponding main effect parameter ‘alternative patterns’ were significant ($\lambda = .35$, $z = 2.90$, $p < .01$). This again refers to the fact that the collapsed patterns are less often selected than not selected. Hence, as expected for this rest category, there was no effect of conditions ($k = 2$, $\chi^2(3) = 2.72$, $p = .44$) and their interaction ($k = 3$, $\chi^2(1) = .58$, $p = .45$).

In conclusion, the predicted interactions ‘card pattern × rule × goal’ were significant only for the predicted *p & q* pattern, and vice versa for the *p & non-q* pattern, but not for the rest category. For the two card patterns, the obligation and the prohibition rule led to vice versa effects in the cooperator detection conditions and in the cheater detection conditions.

Selections of Specific Cards

Table presents the number and percentage of participants who selected *specific* cards for each condition.

Table 61

Selections of Single Cards for Different Rules and Different Foci in Experiment 11 (N = 80)

Card selected	Obligation rule		Prohibition rule	
	Cheater detection	Co-operator detection	Cheater detection	Co-operator detection
P	75 % 15	100 % 20	100 % 20	80 % 16
¬P	30 % 6	10 % 2	10% 2	5% 1
Q	20 % 4	70 % 14	80 % 16	45 % 9
¬Q	70 % 14	15 % 3	20 % 4	55 % 11
n	20	20	20	20

Note. Percentage and number of participants selecting each card. Predicted answers in darkened cells.

On this level of analysis, the hierarchical log-linear calculations showed that the selection of the theoretically decisive cards (*q* and *non-q*) depended on the predicted interaction of rule type and goal (co-operation or cheater detection). For both cards, backward elimination retained the interaction term ‘card × rule × goal’ as generative model class.

Then the remaining saturated model was analyzed. For the *q* card only the expected second order interaction level was found ($k = 3$, $\chi^2(1) = 14.96$, $p < .001$). The analysis of single parameters showed that only the predicted interaction term ‘*q* card × rule × goal’ significantly differed from zero ($\lambda = .48$, $z = 3.70$, $p < .001$).

Similarly, the loglinear model for the *non-q* card only showed the predicted second order interaction level ($k = 3$, $\chi^2(3) = 16.98$, $p < .0001$) and, again, only the interaction term ‘*non-q* card × rule × goal’ significantly differed from zero ($\lambda = -.52$, $z = -3.90$, $p < .001$).

The analysis of the *p* card selections corroborated the expected main effect of the predominance of *p* card selections in all conditions (main effect level: $k = 1$, $\chi^2(3) = 41.80$, $p < .0001$; card parameter: $\lambda = -1.22$, $z = -4.54$, $p < .001$). Over all conditions, *p*-cards were more often selected than not selected. But the analysis also revealed an unpredicted, weaker but also highly significant, second order interaction level (*non-q* card × rule × goal; $k = 3$, $\chi^2(1) = 10.15$, $p < .01$; $\lambda = -.63$, $z = 2.37$, $p < .05$). However, the number of *p*-selections was high in all conditions.

For the *non-p* card, backward elimination in seven iterative steps led to a resulting model only consisting of the predicted main effect of a low selection frequency of this card (main effect level: $k = 1$, $\chi^2(3) = 38.78$, $p < .0001$; main effect parameter *non-p* card: $\lambda = -.92$, $z = 5.42$, $p < .001$).

The most interesting result, showing that the selection of *q* or *non-q* cards was strongly dependent on the rule and the goal of cheater or cooperator detection, is depicted in Figure 17.

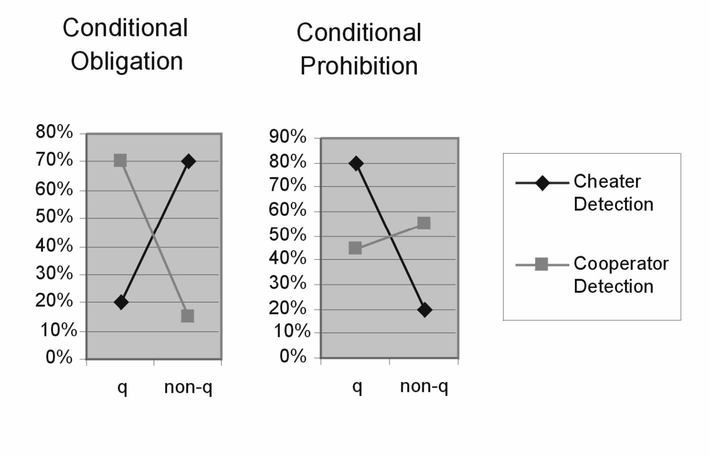


Figure 17. Graph of the percentage of participants selecting *q*- or *non-q*-cards in the cheater or cooperator conditions of 11.

Discussion of Experiment 11

The results of this experiment provide evidence for the predicted interaction of focus effects and types of conditionals. The predictions of FDL theory were largely confirmed, even though the used manipulation was less pronounced than the manipulations used by Cosmides for cheater detection only (1989; Cosmides & Tooby, 1992).[60] More importantly, it was shown that the goal of cheater detection and the goal of cooperator detection lead to reversed selection patterns for the rules of a conditional obligation and a conditional prohibition. Hence, the predictions of FDL theory were largely confirmed.

In the prohibition rule, the results for the cooperation condition were not as pronounced as for the obligation rule. In my opinion, this finding may well be due to the phrasing of the prohibition. Terms like ‘forbids’ might be less open for a cooperator focus than terms like ‘must’, which has here been used to phrase the two rules. It is consistent with FDL theory that even the formulation of the rules may have a connotation rendering a particular focus more or less plausible. However, one may also argue that this is evidence only for a residual cheater focus bias for the

[60] In Cosmides’ experiments, the task of looking for cheaters was not only elicited by a context, but also by the *final* WST instruction to look for cheaters. In Experiment 12 the instructions formally left it to the subjects to check whether someone violated or complied with a rule. Only the variation of the goal of the person induced different selection patterns.

prohibition rule. In any case, even a residual cheater bias would be a much weaker claim than that of the isolated cheater detection module postulated by social contract theory.

The results are problematic for other approaches of the WST. At this point, we will only treat social contract theory (SC theory). Other theories will be considered in the Discussion of Experiment 12 and in the General Discussion.

The Results of Experiment 12 and Social Contract Theory

With regard to the 'cheater detection theory' (e.g., Cosmides, 1989; Cosmides & Tooby, 1992; Gigerenzer and Hug, 1992) the results of Experiment 12 show that not only the goal of cheater detection, but also the goal of cooperator detection can lead to clear-cut and rational selection patterns. In contrast, perspective effects were concerned with *cheater* cases only (cf. Gigerenzer & Hug, 1992).

However, from the viewpoint of social contract theory, the deviations found in the prohibition rule may be seen as evidence in favour of SC theory. To argue in this way, one would need to extend the 'cheater detection theory' to include deontic logic. Although this, in my view, would be incoherent with the illogical spirit of social contract theory, in principle this is possible. SC theory has made no explicit predictions for testing prohibitions, but a proponent of an extended SC theory based on deontic logic would have to postulate a *p & q* pattern for both focus conditions of the prohibition rule. This pattern would be predicted for the cheater condition, because it refers to the cheater case (cf. FDL theory). Likewise, it would be predicted for the cooperator condition, because from the viewpoint of SC theory this pattern (mixed with some other selections) has been assumed to be a general default selection pattern if no cheater detection algorithm is elicited. Hence, the high number of unpredicted *p & q* patterns in the cooperator condition actually found may – alternatively to my above explanation – count in favour of SC theory. However, even in the prohibition condition the cooperator detection focus elicited 55 % *non-q* choices, significantly more than in the cheater condition. This is a high result for WSTs, particularly if we take into account that the manipulation of cheater vs. cooperator instructions were weaker than in the original cheater instructions used by Cosmides (1989, cf. Cosmides & Tooby, 1992). Hence, also the results of the prohibition rule conditions confirm the focus predictions of FDL theory.

Moreover, the obligation conditions clearly favour FDL theory over the cheater detection approach. Here a clear-cut focus on the *p & q* cell as well as on the *p & non-q* cell of an ought-table has been elicited. To interpret the *p* & *q* selections as standard for testing deontic obligations in the WSTs would be misleading. Although Gigerenzer and Hug (1992, cf. their fig. 4) indeed predicted a reduction of *p & non-q* patterns if a cheater focus is removed, they did not predict and did not find the clear predominance of a *p* & *q* patterns, as we have found here. Instead, their results show that the (reduced) *p & non-q* selection pattern remained by far the most frequent pattern in their relevant obligation-testing tasks.[61] Also the research of other authors showed that without an explicit cheater detection instruction the testing of prescriptive obligations (if they are not reinterpreted as descriptive obligations) rather leads to cheater detection selections than to cooperation detection selections (see particularly Sperber & Girotto, 2003, with 'mixed Wason/FCT rules', cf. Manktelow & Over, 1990). Thus, although formally using an unspecified selection instruction, the introduction of a cooperator detection goal, led to more cooperation detection selections than found in any other comparable deontic WST. Nonetheless, an additional control condition will be introduced in Experiment 12.

Finally, the results provide evidence for the predicted *interaction* of deontic logic and focus effects. This is the straightest argument against SC theory and this favours the claim that in testing prescriptive rules, the 'algorithm' of cheater detection is only one way of focusing on different cells of different ought tables.

The findings may seem to contradict previous results of Cosmides (1989; Cosmides & Tooby, 1992) on altruist detection. In her studies, it was found that altruist detection instructions did not lead to any clear-cut selection patterns. A closer analysis of the chosen tasks reveals that this finding may be due to the phrasing of the instructions used in the scenarios. Altruistic behaviour in these scenarios did not clearly correspond to a specific selection of cards; altruist selections were consistent with both *p & q* or *non-p & q* selections.

Since the current results are at odds with central claims of Cosmides and colleagues, it seems that cheater detection looses its unique position. Correspondingly, the findings cast doubt on the modular understanding of cheater detection.

[61] Other results of Cosmides, 1989, Exp. 1, Exp. 2, and Gigerenzer and Hug, 1992, Exp. 1, cannot be considered for this question, since the number of *p* & *q* selections have not been reported.

Are these Results Consistent with Previous Studies?

Some of the earlier mentioned studies on so-called facilitation effects may perhaps turn out to be actually due to focus effects. For example, the manipulation of van Duyne (1974) was clearly confounded with a focus instruction. However, there have been only very few studies which voluntarily or accidentally investigated focus effects in a symmetrical way with different foci (cf. Manktelow & Over, 1990; Love & Kessler, 1995; and particularly Sperber & Girotto, 2002, 2003). These studies differed in their theoretical background as well as in the experimental method from the present study. We will briefly discuss to what extent these studies differ and to what extent they came to similar conclusions.

Manktelow and Over (1990, cf. 1992) have taken a decision theoretic perspective and have tried to vary the subjective pay-off of different card selections. Based on varying pay-offs, not goals, they did not find what we treated here as focus effects for confirming cases. Partly based on these results Manktelow and Over (1992, 185) distanced themselves from normative decision theory. In contrast, here participants were found to check also for cooperation detection cases of an obligation rule in the context of social sanctioning. These different results may perhaps favour a focus account over a more general utilitarian account. In any case, the focus account provides a different manipulation and more specific prediction than a general ‘decision theoretic account’ (cf. General Discussion).

Love and Kessler (1995) conducted some WSTs, which may also be reinterpreted as focusing experiments in our current sense (although they used this term ‘focusing’ in a different meaning). The probabilistic attention effects, which they have postulated in their first experiment, are confounded with focus effects in the current sense. Moreover, they did not interpret their results based on something like a flexible deontic logic, but as facilitation effects of logically correct *p* & *non-q* responses.

Theoretically, my proposal combines the idea of a focus mechanism, as it has also been mentioned in the work of Oaksford and Chater (1994) and then elaborated in the work of Sperber and Girotto (2002/2003), with deontic logic. The results of Sperber and Girotto (2002) in testing conditionals (Exp. 2) have already supported focus effects. Sperber and Girotto (2003) found some clear focus effects, but found in so-called mixed Wason/FCT-rules (in which participants are instructed to check for cheaters *or* followers of the rule) that participants clearly tended to check for cheaters.

In the experiment discussed here, the participants were also formally requested to check “whether the rule has been violated or followed”. Nonetheless, here a variation of their role descriptions led to both, either a cheater focus or a cooperator focus. Hence, Experiment 11 supports the idea of focus effects using the goals of cheater and cooperator detection. Moreover, Experiment 12 combined focus effects with deontic logic.

The present experiment was the first WST, which combined the testing of different types of deontic conditionals with the variation of a cheater or a cooperator detection focus. We found the interaction of types of conditionals and induced goals, which was predicted by FDL theory.

12.3 Experiment 12 – Deontic Logic, Cooperator Detection and Double Focus Effects

The goal of Experiment 12 was to test the prediction of a *double focus effect* for different kinds of conditionals, which would provide specific support for FDL theory. Additionally, the interaction of single focus effects with different deontic rules of Experiment 11 should be replicated, now using the double focus conditions also as control.

The Concept of Double Focus Effects

A double focus is here understood as the conjunction of both goals, cooperator detection *and* cheater detection. A double focus may lead to a double focus effect, that is, a selection of the cheater detection cards and the cooperator detection cards at the same time. Since such patterns have never been predicted by other theories and have never been found to dominate a deontic WST before, this would strongly support FDL theory.

Let us consider a fictitious tribal law in the form of a conditional obligation “If one is a bachelor, one is obliged to bring a tiger’s coat to the medicine man” and its corresponding ought table (cf. Table 62). In a context of punishment of rule violators (cheater detection) it would be rational to inspect “bachelors” (*p*) and those who “did not bring a tiger’s coat” (*non-q*), since these are the cases were a cheater can be found. If one’s aim is to reward rule followers (cooperator detection) “bachelors” (*p*) and persons who “did bring a tiger’s coat” (*q*) would have to be inspected. In the context of a double focus, when one has to punish violators *and* has to reward rule followers, one should to check all three cases “bachelors” (*p*), those who “did bring the tiger’s coat” (*q*) and those who “did not bring the tiger’s coat” (*non-q*) (cf. Table 62).

Table 62
Conditional Obligation And a Double Focus: Combined Cheater (Full Circle) and Cooperator Focus (Dotted Circle)

Obligation – Double Focus		
	Brings a tiger’s coat (*q*)	Does not bring a tiger’s coat (*non-q*)
Bachelor (*p*)	Allowed	Forbidden
Husband (*non-p*)	Allowed	Allowed

Table 63
Conditional Prohibition And a Double Focus Combining Cheater (Full Circle) and Cooperator Focus (Dotted Circle)

Prohibition – Double Focus		
	Has drunk from the glogg (*q*)	Has not drunk from the glogg (*non-q*)
Bachelor (*p*)	Forbidden	Allowed
Husband (*non-p*)	Allowed	Allowed

If the auxiliary hypothesis that subjects are able to check two foci at the same time is added to FDL theory, a *p & q & non-q* card pattern is predicted for the WST. This pattern is our main prediction for double focus conditions – although such patterns have neither been predicted nor observed to be dominant in deontic and descriptive WSTs. For the prohibition rule “if one is a bachelor one is forbidden from drinking the popular glogg” analogous predictions can formulate for the prohibition rule (cf. Table 63).

In contrast, social contract theory, pragmatic reasoning theory, and mental model theory would all have difficulties in accounting for this double focus effect (cf. Discussion).

A weaker auxiliary hypothesis, which in principle would also be consistent with FDL theory, is the assumption that participants only check for one focus at a time (a cheater *or* a cooperator focus). Such a ‘one cell focus’ assumption has some prior plausibility, because it is empirically known that checking for cheaters of a ‘biconditional obligation’ (with two forbidden cells) normally leads to the selection of only two cards connected with one of the two forbidden cells. This is the case although the task would logically require participants to select all four cards (cf. e.g.: Johnson-Laird & Byrne, 1991, pp. 78-79; Manktelow & Over, 1991; Johnson-Laird & Byrne, 1992, 1995; Almor & Sloman, 2000; Beller & Spada, 2003; but see Politzer & Nguyen-Xuan, 1992). Hence, a general tendency to check only for one focus appears plausible – also in regard of double focus conditions with a cheater *and* a cooperator focus in the instruction. In this case, double focus conditions could still serve as interesting control conditions, and also in this role they could provide indirect evidence for FDL theory (the double focus conditions could test which foci are checked by default).

However, I think that two foci of a cooperator-cheater double focus can be elicited at the same time, if both goals appear plausible and applicable. Hence, here the foci were emphasized in the instructions.

In Experiment 12 these novel double focus conditions are combined with a replication of the found interaction between a cooperator versus a cheater focus and an obligation versus a prohibition rule, found before in Experiment 11. The instruction was changed slightly using the examples from above (Table 62, Table 63). Unlike varying the goals only indirectly at the beginning of the instruction, here the goals were explicitly mentioned in the final instruction of the WSTs. In this respect the tasks more closely resemble the original deontic cheater detection WSTs of Cosmides (1998, cf. Cosmides & Tooby, 1992). Additionally, the context of the rules was changed and the conditional obligation rule was formulated without the modal ‘must’.

Method of Experiment 12

Design and Participants

Experiment 12 had a three foci (cheater focus vs. cooperator focus vs. double focus) × two rules (conditional obligation vs. conditional prohibition) between-subjects design.

120 members of the University of Göttingen voluntarily took part in the experiment (48 % female, 52 % male; mean age: 25 years). Apart from a few staff members, the majority of participants were students. They were members of different departments (the largest group of 33 % studied law). The participants were randomly assigned to the six conditions. Three participants were excluded and replaced because they did not follow the instructions. After the experiment, all participants received a little present and were able to win additional prizes.

Materials and Procedure

Participants were instructed to read the text carefully and informed they should ask the experimenter if they had any questions concerning comprehension of the task. The WSTs were formulated in German.

In the WSTs the participants, as in Experiment 11, had to imagine they were members of a tribal council of elders. The council is described as being generally responsible for checking whether the rules of that tribe have been followed or violated.

Subsequently a particular rule was presented. The rules and the goals were highlighted using a larger font size. In the three conditional obligation conditions the rule read: “If one is a bachelor, then one is this year obliged to bring a tiger skin to the medicine man”. In the prohibition rule conditions it read: “If one is a bachelor, then one is this year forbidden from drinking a popular fruit glogg”.

Then the goals were introduced. For the three conditions, the instruction read as follows:

- Cheater conditions: “In checking this rule you have *only one goal*: You should *punish those, who have violated the above rule*.”
- Cooperator conditions: “In checking this rule you have *only one goal*. You should *reward those, who have followed the above rule*.”
- Double focus conditions: “In checking this rule you have *two goals* at the same time: [First bullet.] You should *reward those, who followed the above rule*. [Second bullet.] You should *punish those, who violated the above rule*.”

Four male members of the tribe were put forward for possible checks. Each of the four members of the tribe was represented by a card. As in the other experiments participants were instructed, that one side of each card provided information about whether the clansman is a bachelor or not, and the other side provided information whether he has brought a tiger skin (or has drunk from the fruit glogg) or not. It was clarified that the four cards shown (below the text) each showed only one side of a card.

Unlike Experiment 11 the varied goals were here explicitly repeated in the task description. The instruction read as follows: “Which card(s) do you have to turn over, to check exactly for all possible cases of violating [alternatively: "following"; "violating and following"] the above rule. Please mark all the cards necessary."

The cards read: "Bachelor" (*p*), "Husband" (*non-p*), "has brought tiger skin" (or: "has drunk from the glogg", *q*), and "has not brought tiger skin" (or: "has not drunk from the glogg", *non-q*).

After finishing, the participants were asked to give comments on the task and their understanding of it, and to provide some demographic data.

Results of Experiment 12

We first analyse the results on the level of card combinations and then on the level of single card selections.

Selections of Card Combinations

The resulting patterns of card selections for different rules and foci are shown in Table 64.

The predicted answers are darkened. Descriptively, all differences between conditions were in the predicted direction. This is the case as well for the *p & q* pattern, the *p & non-q* pattern and the *p & q & non-q* pattern. However, the results in the cooperator and double focus condition of the prohibition rule appear to be less pronounced than one may have hoped.

Table 64
Selections of Card Combinations for Different Rules and Different Foci, Experiment 12 (N = 120)

	Obligation rule			Prohibition rule		
Card pattern	Cheater detection	Double focus	Co-operator detection	Cheater detection	Double focus	Co-operator detection
P, *Q*	10 % 2	5 % 1	50 % 10	85 % 17	35 % 7	20 % 4
P, *¬Q*	70 % 14	25 % 5	10 % 2	0 % 0	5 % 1	35 % 7
P, *Q*, *¬Q*	0 % 0	60 % 12	20 % 4	0 % 0	35 % 7	25 % 5
P	5 % 1	0 % 0	5 % 1	5 % 1	0 % 0	0 % 0
Q	0 % 0	0 % 0	5 % 1	0 % 0	5 % 1	0 % 0
¬Q	0 % 0	0 % 0	5 % 1	0 % 0	0 % 0	5 % 1
P, *¬P*	0 % 0	0 % 0	0 % 0	0 % 0	5 % 1	10 % 2
¬P, *Q*	5 % 1	0 % 0	0 % 0	5 % 1	15 % 3	0 % 0
¬P, *¬Q*	5 % 1	0 % 0	0 % 0	0 % 0	0 % 0	0 % 0
P, *¬P*, *Q*	0 % 0	0 % 0	0 % 0	5 % 1	0 % 0	0 % 0
P, *¬P*, *Q*, *¬Q*	5 % 1	10 % 2	5 % 1	0 % 0	0 % 0	5 % 1
Other combinations	0 % 0	0 % 0	0 % 0	0 % 0	0 % 0	0 % 0
n	20	20	20	20	20	20

Note. Percentage and number of participants selecting each card pattern. Predicted answers in darkened cells.

Since it is suitable to present the statistical results separately for the obligation and the prohibition rule, I refrained from calculating a loglinear analysis of the data (also because the parameters of the three foci conditions are more difficult to interpret). Instead, all relevant single comparisons are presented successively. We first analyze the results obtained for the obligation rule and only then we turn to the prohibition rule.

Conditional obligation rule. All tests of the comparisons of the different focus conditions of the obligation rule confirmed the predictions of FDL theory (Table 65).

Firstly, the comparison between the cheater and the cooperator condition of that rule confirmed the predictions. There were significantly more *p* & *non-q* selections in the cheater than in the cooperator condition (Pearson test: $\chi^2_{(1,\, n\, =\, 40)} = 15.00, p < .001$), more *p* & *q* selections in the cooperator than in the cheater condition (Pearson test: $\chi^2_{(1)} = 7.62, p < .01$), and – also as expected – no significant difference between these conditions for the *p* & *q* & *non-q* patterns (exact Fisher test, *df* = 1, *p* = .11) and for the remaining conflated thirteen patterns (exact Fisher test, *p* = 1.00).

Table 65

For the Obligation Rule: Hypotheses and Tests of the Comparisons of Card Combinations between the Cheater, the Cooperator and the Double Focus Condition

Hypotheses of FDL theory	Result		Positive?
$f(p \,\&\, q)_{\text{Cheater}} < f(p \,\&\, q)_{\text{Cooperator}}$	$\chi^2_{(1,\, n\, =\, 40)} = 7.62,$	$p < .01$	yes
$f(p \,\&\, \neg q)_{\text{Cheater}} > f(p \,\&\, \neg q)_{\text{Cooperator}}$	$\chi^2_{(1,\, n\, =\, 40)} = 15.00,$	$p < .001$	yes
$f(p \,\&\, q \,\&\, \neg q)_{\text{Cheater}} = f(p \,\&\, q \,\&\, \neg q)_{\text{Coop.}}$	exact Fisher test,	$p = .11$	yes
$f(\text{Rest})_{\text{Cheater}} = f(\text{Rest})_{\text{Cooperator}}$	exact Fisher test,	$p = 1.00$	yes
$f(p \,\&\, q)_{\text{Cheater}} = f(p \,\&\, q)_{\text{Double}}$	exact Fisher test,	$p = 1.00$	yes
$f(p \,\&\, \neg q)_{\text{Cheater}} > f(p \,\&\, \neg q)_{\text{Double}}$	$\chi^2_{(1,\, n\, =\, 40)} = 8.12,$	$p < .01$	yes
$f(p \,\&\, q \,\&\, \neg q)_{\text{Cheater}} < f(p \,\&\, q \,\&\, \neg q)_{\text{Double}}$	$\chi^2_{(1,\, n\, =\, 40)} = 17.14,$	$p < .001$	yes
$f(\text{Rest})_{\text{Cheater}} = f(\text{Rest})_{\text{Double}}$	exact Fisher test,	$p = 0.66$	yes
$f(p \,\&\, q)_{\text{Double}} < f(p \,\&\, q)_{\text{Cooperator}}$	$\chi^2_{(1,\, n\, =\, 40)} = 10.16,$	$p < .001$	yes
$f(p \,\&\, \neg q)_{\text{Double}} = f(p \,\&\, \neg q)_{\text{Cooperator}}$	exact Fisher test,	$p = .41$	yes
$f(p \,\&\, q \,\&\, \neg q)_{\text{Double}} > f(p \,\&\, q \,\&\, \neg q)_{\text{Coop.}}$	$\chi^2_{(1,\, n\, =\, 40)} = 6.67,$	$p < .01$	yes
$f(\text{Rest})_{\text{Double}} = f(\text{Rest})_{\text{Cooperator}}$	exact Fisher test,	$p = .66$	yes

Note. If not indicated differently, the empirical *p* values are given for two-sided tests. For the predictions of alternative theories, cf. Table 70.

Also for the comparison between the cheater detection condition and the double focus condition, only the differences predicted became significant. There was a higher number of *p* & *non-q* patterns in the cheater condition than in the double focus condition (Pearson test: $\chi^2_{(1)} = 8.12, p < .01$) and more selections of *p* & *q* & *non-q* in the double focus condition than in the cheater condition (Pearson test: $\chi^2_{(1)} = 17.14$, *p* < .001). As expected, there was neither a difference between the number of selected *p* & *q* patterns (exact Fisher test: *p* = 1.00) nor between the remaining conflated patterns (exact Fisher test: *p* = 0.66).

Finally, the comparisons between the cooperator focus and the double focus condition also corroborated FDL theory. The *p* & *q* & *non-q* pattern was selected significantly more frequently in the double focus than in the cooperator condition (Pearson test: $\chi^2_{(1)} = 6.67, p < .01$). While, opposed to this, the *p* & *q* pattern was selected more frequently in the cooperator condition than in the double focus condition (Pearson test: $\chi^2_{(1)} = 10.16, p < .001$). Furthermore, there was no significant difference for the *p & non-q* card selections (exact Fisher test: $p = .41$) and the remaining conflated patterns (exact Fisher test: $p = .66$).

Conditional prohibition rule. In Table 66 an overview is given of the tests of all comparisons concerning the prohibition rule.

The comparison between cheater condition and cooperator condition yielded the reversed differences predicted for the *p & q* pattern and *p & non-q* pattern. For the prohibition rule there were significantly more *p & q* selections in the cheater condition than in the cooperator condition (Pearson test: $\chi^2_{(1)} = 16.94$, $p < .001$) and more *p & non-q* patterns the other way round (exact Fisher test: $p < .001$). Deviating from the predictions, the number of selected *p & q & non-q* double focus patterns significantly increased in the cooperator condition relative to the cheater detection condition (exact Fisher test: $p = .047$). It appears that some participants have adopted a double focus strategy also in this condition. As predicted there was no difference for the other card patterns (exact Fisher test: $p = 1.00$).

Table 66
For the Prohibition Rule: Hypotheses and Tests of the Comparisons of Card Combinations between the Cheater, the Cooperator and the Double Focus Condition

Hypotheses of FDL theory	Result		Positive?
$f(p \,\&\, q)_{\text{Cheater}} > f(p \,\&\, q)_{\text{Cooperator}}$	$\chi^2_{(1)} = 16.94$,	$p < .001$	yes
$f(p \,\&\, \neg q)_{\text{Cheater}} < f(p \,\&\, \neg q)_{\text{Cooperator}}$	exact Fisher test,	$p < .001$	yes
$f(p \,\&\, q \,\&\, \neg q)_{\text{Cheater}} = f(p \,\&\, q \,\&\, \neg q)_{\text{Coop.}}$	exact Fisher test,	$p = .047$	no
$f(\text{Rest})_{\text{Cheater}} = f(\text{Rest})_{\text{Cooperator}}$	exact Fisher test,	$p = 1.00$	yes
$f(p \,\&\, q)_{\text{Cheater}} > f(p \,\&\, q)_{\text{Double}}$	$\chi^2_{(1)} = 10.42$,	$p < .001$	yes
$f(p \,\&\, \neg q)_{\text{Cheater}} = f(p \,\&\, \neg q)_{\text{Double}}$	exact Fisher test,	$p = 1.00$	yes
$f(p \,\&\, q \,\&\, \neg q)_{\text{Cheater}} < f(p \,\&\, q \,\&\, \neg q)_{\text{Double}}$	$\chi^2_{(1)} = 10.42$,	$p < .001$	yes
$f(\text{Rest})_{\text{Cheater}} = f(\text{Rest})_{\text{Double}}$	exact Fisher test,	$p = .45$	yes
$f(p \,\&\, q)_{\text{Double}} = f(p \,\&\, q)_{\text{Cooperator}}$	$\chi^2_{(1)} = 1.13$,	$p = .29$	yes
$f(p \,\&\, \neg q)_{\text{Double}} < f(p \,\&\, \neg q)_{\text{Cooperator}}$	exact Fisher test,	$p < .05$	yes
$f(p \,\&\, q \,\&\, \neg q)_{\text{Double}} > f(p \,\&\, q \,\&\, \neg q)_{\text{Coop.}}$	$\chi^2_{(1)} = .73$,	$p = .72$	no
$f(\text{Rest})_{\text{Double}} = f(\text{Rest})_{\text{Cooperator}}$	$\chi^2_{(1)} = .53$,	$p = .47$	yes

Note. If not indicated differently, the empirical *p* values are given for two-sided tests. For the predictions of alternative theories, cf. Table 71.

If the cheater detection condition is compared with the double focus condition of the prohibition rule, the found predicted differences are reversed to the obligation rule. There were more *p & q* selections in the cheater condition than in the double focus condition (Pearson test: $\chi^2_{(1)} = 10.42$, $p < .001$). In regard of *p & non-q* card selections there were, as expected, equally few selections in both conditions (exact Fisher test: $p = 1.00$). With respect to the *p & q & non-q* selections, the difference could have been higher, but, as hypothesized, there were significantly more such selections in the double focus condition than in the cheater condition (Pearson test: $\chi^2_{(1)} = 10.42$, $p < .001$). Again the remaining conflated selection patterns did not differ between conditions (exact Fisher test: $p = .45$).

Comparing the cooperator and the double focus conditions, it was corroborated that there was no difference in the number of *p & q* card selections (Pearson test: $\chi^2_{(1)} = 1.13$, $p = .29$), and more *p* & *non-q* selections in the cooperator condition than in the double focus condition (exact Fisher test: $p < .05$). However, in contrast to the predictions of FDL theory (strong auxiliary hypothesis, see above) the difference in *p & q & non-q* selections between the double focus condition and the cooperator condition was not significant (Pearson test: $\chi^2_{(1)} = .73$, $p = .72$). As expected there was no difference in the remaining card selections (Pearson test: $\chi^2_{(1)} = .53$, $p = .47$).

In summary, all tests in the obligation rule condition led to the results predicted by FDL theory. For the prohibition rule, the effects were less pronounced than one may have hoped, but most comparisons confirmed the general predictions; only two tests concerning the *p & q & non-q* patterns diverged from the expected pattern (cf. discussion).

Selections of Single Cards

Table 67
Selections of Single Cards for Different Rules and Different Foci in Experiment 12 (N = 120)

	Obligation rule			Prohibition rule		
Card selected	Cheater detection	Double focus	Co-operator detection	Cheater detection	Double focus	Co-operator detection
P	90 % 18	100 % 20	90 % 18	95 % 19	80 % 16	95 % 19
¬P	15 % 3	10 % 2	5 % 1	10 % 2	20 % 4	15 % 3
Q	20 % 4	75 % 15	80 % 16	95 % 19	90 % 18	50 % 10
¬Q	80 % 16	95 % 19	40 % 8	0 % 0	40 % 8	70 % 14
n	20	20	20	20	20	20

Note. Percentage and number of participants selecting each card. Predicted answers in darkened cells.

The results for the selections of single cards are presented in Table 67. All predicted differences between conditions were found and they descriptively went in the predicted direction.

Conditional obligation rule conditions. An overview of the main comparisons of on the single card level is given in Table 68. For the obligation rule, the contrast between cheater detection condition and cooperator condition did replicate the hypothesized differences also found in Experiment 2. There was no difference for the *p* selections (exact Fisher test: $p = 1.00$) and for the *non-p* selections (exact Fisher test: $p = .61$), but more *q* selections in the cooperator condition (Pearson test: $\chi^2_{(1)} = 14.40$, $p < .001$) and more *non-q* selections in the cheater condition (Pearson test: $\chi^2_{(1)} = 6.67$, $p < .01$).

Table 68

For the Obligation Rule: Comparisons of the Single Card Selections Q and Non-Q between Cheater, Cooperator and Double Focus Condition

Hypotheses of FDL theory	Result		Positive?
$f(q)_{\text{Cheater}} < f(q)_{\text{Cooperator}}$	$\chi^2_{(1)} = 14.40$,	$p < .001$	yes
$f(\neg q)_{\text{Cheater}} > f(\neg q)_{\text{Cooperator}}$	$\chi^2_{(1)} = 6.67$,	$p < .01$	yes
$f(q)_{\text{Cheater}} < f(q)_{\text{Double}}$	$\chi^2_{(1)} = 12.13$,	$p < .001$	yes
$f(\neg q)_{\text{Cheater}} = f(\neg q)_{\text{Double}}$	exact Fisher test,	$p = .34$	yes
$f(q)_{\text{Double}} = f(q)_{\text{Cooperator}}$	exact Fisher test,	$p = 1.00$	yes
$f(\neg q)_{\text{Double}} > f(\neg q)_{\text{Cooperator}}$	$\chi^2_{(1)} = 13.79$,	$p < .001$	yes

Note. If not indicated differently, the empirical *p* values are given for two-sided tests. For the predictions of alternative theories, cf. Table 72.

When additionally comparing the cheater detection condition with the double focus condition there was a higher number of *q* ‘cooperator’ card selections in the double focus condition (Pearson test: $\chi^2_{(1)} = 12.13$, $p < .001$), and there was no significant difference in the number of *non-q* card ‘cheater’ selections (exact Fisher test: $p = .34$). This is in accordance with FDL theory, which predicts as many *non-q* card selections for the double focus condition as for the cheater condition, but a relative increase of *q* cards in the double focus condition. Moreover, as expected there were no reliable differences between the conditions for the *p* (exact Fisher test: $p = .49$) and *non-p* card selections (exact Fisher test: $p = 1.00$).

The comparison between cooperator focus and double focus condition again led to no difference for the *p* selections and for the *non-p* selections (exact Fisher tests: $p = .49$; $p = 1.00$). More interestingly, it was also confirmed that there was no difference in the (high) number of *q* card selections (exact Fisher test: $p = 1.00$), but a higher number of *non-q* selections in the double focus condition (Pearson test: $\chi^2_{(1)} = 13.79$, $p < .001$).

Conditional prohibition rule conditions. An overview over the tests of the comparisons in the prohibition rule is given in Table 69.

Comparing the cheater condition and the cooperator condition, the *q* selections and the *non-q* selections varied in the reverse direction than it did in the obligation rule conditions. There were significantly more *q* selections in the cheater than in the cooperator condition (Pearson test: $\chi^2_{(1)} = 10.16$, $p < .01$), and more *non-q* selections in the cooperator condition than in the cheater condition (Pearson test: $\chi^2_{(1)} = 21.54$, $p < .001$). The *p* selections and the *non-p* selections were not affected (exact Fisher test: $p = 1.00$; exact Fisher test: $p = 1.00$).

Table 69

For the Prohibition Rule: Comparisons of Single Card Selections Q and Non-Q Between Cheater, Cooperator and Double Focus Condition

Hypotheses of FDL theory	Result		Positive?
$f(q)_{\text{Cheater}} > f(q)_{\text{Cooperator}}$	$\chi^2_{(1)} = 10.16$,	$p < .01$	yes
$f(\neg q)_{\text{Cheater}} < f(\neg q)_{\text{Cooperator}}$	$\chi^2_{(1)} = 21.54$,	$p < .001$	yes
$f(q)_{\text{Cheater}} = f(q)_{\text{Double}}$	exact Fisher test,	$p = 1.00$	yes
$f(\neg q)_{\text{Cheater}} < f(\neg q)_{\text{Double}}$	exact Fisher test,	$p < .01$	yes
$f(q)_{\text{Double}} > f(q)_{\text{Cooperator}}$	$\chi^2_{(1)} = 7.62$,	$p < .01$	yes
$f(\neg q)_{\text{Double}} = f(\neg q)_{\text{Cooperator}}$	$\chi^2_{(1)} = 3.64$,	$p = .057$	(yes)

Note. If not indicated differently, the empirical *p* values are given for two-sided tests. For the predictions of alternative theories, cf. Table 73.

If the cheater detection and double focus conditions are compared, an increase of *non-q* (cooperator) selections was found in the double focus condition (exact Fisher test: $p < .01$). As predicted there was no difference between the (high) number of *q* (cheater) selections in the two conditions (exact Fisher test: $p = 1.00$). Again, the conditions had no impact on the *p* selections and the *non-p* selections (exact Fisher test: $p = .34$; exact Fisher test: $p = .66$).

Finally, the differences between the cooperator detection condition and the double focus condition were tested. As hypothesized, there were significantly more *q* selections (in this condition: cheater detection cards) in the double focus condition than in the cooperator condition of the prohibition rule (Pearson test: $\chi^2_{(1)} = 7.62$, $p < .01$). But here one deviation from the predictions is also found at this level of analysis. There seem to have been more *non-q* (cooperator) card selections in the cooperator condition than in the double focus condition. Although this difference formally did not reach significance, it was very close to become significant (Pearson test: $\chi^2_{(1)} = 3.64$, $p = .057$). Also for this comparison, there were no differences in the number of *p* card selections (exact Fisher test: $p = .34$) and in the number of *non-p* card selections (exact Fisher test: $p = 1.00$).

To sum up, on the level of the single card selections all but perhaps one comparison confirmed the predictions of FDL theory.

Discussion of Experiment 12

Experiment 12 will first be discussed with respect to FDL theory. It will be shown that FDL theory has predicted most of the results and can even account for the few deviations found. Then the three main alternative theories of deontic WSTs, pragmatic reasoning schema theory, social contract theory and mental model theory will each be discussed separately. It will be shown that neither of them can account for the results of Experiment 12. Finally, there will be a brief simultaneous overview of the results and of the visualised predictions of all discussed theories.

Confirmation of FDL Theory

Experiment 12 strongly supports the predictions of FDL theory both on the level of card combinations and on the level of single card selections.

Firstly, Experiment 12 replicated the interaction of cheater detection versus cooperator detection condition with the tested prohibition versus obligation rules (cf. Experiment 11). The two rules led in the cooperator and cheater conditions to reversed *q* and *non-q* selection patterns (cf. Table 65, Table 66, Table 68, Table 69).

Secondly, a predicted clear-cut double focus effect with a *p & q & non-q* selection pattern was found at least for the obligation rule (Table 64, Table 65, Table 67, cf. Figure 18a, b). Until now such a pattern has neither been explicitly predicted nor found to dominate any deontic WST.

Only in the prohibition rule conditions did a few comparisons diverge from the predictions (Table 64, Table 66, Table 68). These departures do need an explanation, particularly since the manipulation was stronger than in Experiment 11. Nonetheless, despite some weak effects, most of the expected effects were significant also for the prohibition rule. The analysis of card combinations only showed deviations for the *p & q & non-q* pattern. As expected, there were significantly more such selections in the double focus conditions than in the cheater detection conditions. However, it was not expected that there were more such selections in the cooperator than in the cheater condition and not more in the double focus condition than in the cooperator condition. But by a plausible assumption this can easily be made explainable in the framework of FDL theory; one may even argue that also this result uniquely supports FDL theory. As asserted by FDL theory the subjective goal of participants and the resulting focus, is not only based on the *explicit* focus given in the instruction, but also on the implicit connotations of the formulation of the tested rule itself. As in Experiment 11 the formulation of a conditional prohibition, using a term like ‘forbidden’, seems to have a tendency to elicit a subjective cheater focus. The formulation of the prohibition rule seems to be less neutral than the formulation of conditional obligation rules using terms like ‘must’ or ‘is obliged to’. If the formulation of the conditional prohibition increased the probability of a subjective cheater focus, the results become perfectly intelligible. Consistent with this explanation, some participants in the cooperator focus condition seem to have combined the formally given cooperator focus with this implicit cheater focus and hence have selected the double focus pattern *p* & *q* & *non-q*. In the cooperator condition there were clearly more participants who selected a cooperator or a double focus pattern (taken together) than a cheater pattern, whereas this was completely reversed in the cheater condition (Table 64). Moreover, it also becomes understandable, why in the double focus conditions of the prohibition rule some participants only checked for the cheater focus, which was explicitly and implicitly plausible, and not for the cooperator focus, which was explicitly plausible but implicitly implausible. Hence, these patterns and the double focus effect in the cooperator condition of the prohibition rule seem to confirm my interpretation of Experiment 11. I see no other way to explain these double focus effects, also found in the prohibition rule (Table 66, Table 69). Hence, also these effects appear to confirm FDL theory. In any case, even in the prohibition rule, all comparisons for the *p* & *q* patterns and the *p* & *non-q* patterns were confirmed. Moreover, on the level of single cards there where also significantly more *q* selections (here: cheater detection selections) in the double focus and cheater conditions than in the cooperator condition and more *non-q* selections (here: cooperator detection) in the double focus and the cooperator condition than in the cheater condition.

In summary, the results of the obligation rule conditions in particular, but also the results of the prohibition rule conditions support the predictions of FDL theory. For both rules, double focus effects were shown and the interaction of rules and foci was replicated.

Other Theories of the WST and the Results – the Specific Theories

The main three alternative theories of the WST and their most important predictions will now be discussed separately. Many arguments are applicable both to Experiment 12 and Experiment 11 (For a visualisation of the predictions of these theories cf. Figure 18, Figure 19, and Table 70 to Table 73.)

(a) Social contract theory (SC theory, see pp. 14 f., 175 f., 264 f.) has advocated an adapted and specialized cheater detection module and has broken with a normative logical standard of checking deontic rules (Cosmides, 1989; Cosmides & Tooby, 1992; cf. Gigerenzer & Hug, 1992; Fiddick, Cosmides, & Tooby, 2000; Fiddick, 2003, 2004). Proponents of SC theory have claimed that the activation of a cheater detection 'module' leads to clear-cut patterns in social contracts (Cosmides, 1989; Cosmides & Tooby, 1992; Gigerenzer & Hug, 1992). Likewise, perspective effects, which in hindsight might be reinterpreted as particular kinds of focus effects, were exclusively concerned with cheater detection cases.

SC theory may account for some results of Experiment 12. For instance, a clear-cut *p* & *non-q* pattern in the cheating condition of an obligation rule is predicted by SC theory and FDL theory alike. But without giving up the essential claim of SC theory of an illogical and modular cheater detection module, SC theory is incoherent with the found interaction of cheater and cooperator conditions in the obligation or prohibition rule conditions. However, this only replicates Experiment 11 (cf. Discussion of Experiment 11).

Additionally, the double focus conditions in Experiment 12 yielded a *p* & *q* & *non-q* pattern, which is clearly inconsistent with SC theory. For the double focus condition of the obligation rule, SC theory predicts a *p* & *non-q* cheater detection pattern. The cheater detection algorithm should have become activated, since cheating was mentioned as frequently as in the cheater condition. Instead, the double focus condition led to a significant increase of *p* & *q* & *non-q* patterns. The same argument can be made on the level of single card selections. Because the number of *non-q* selections was equally high in the double focus condition as in the cheater condition of the obligation rule, a proponent of SC theory would have to assume that both conditions activated the 'cheater detection module'. But in this case, it becomes inexplicable – without a second cooperator focus – why the number of *q* (cooperator detection) selections increased significantly in the double focus condition relative to the cheater condition. Hence, these findings are inexplicable on the basis of an isolated cheater detection module.

In addition, the increase of double focus selections in the prohibition rule cannot be explained by SC theory. The additionally found *residual cheater bias* has been discussed before. It has been argued that a cheater bias is a much weaker claim than the original claim of SC theory that a specialized cheater detection algorithm explains the selections in deontic WSTs. However, it has been outlined before that also for the prohibition rule cooperator and double focus effects were observed, which cannot be explained by SC theory, but by FDL theory.

In conclusion, Experiment 12, with single and double focus effects, provides strong support for the idea that the cheater detection module is to be explained by a more general and more rational focus on different cells of deontic ought tables.

(b) *Pragmatic reasoning schema theory* (PRS theory, see pp. 13 f., 176 f., 270 f.) has postulated an obligation schema and a permission schema (Cheng & Holyoak, 1985; Cheng, Holyoak, Nisbett, & Oliver, 1986; Politzer & Nguyen-Xuan, 1992; Holyoak & Cheng, 1995b). According to PRS theory, schemas are activated if a rule accords with certain production rules. Even if PRS theory were extended to include prohibition rules, it would not be able to account for any of the obtained focus effects. PRS theory would actually have to postulate cheater detection answers invariantly for all three different focus conditions. This is inconsistent with the empirical evidence.

(c) *Mental model theory* (MM theory, see pp. 11 f., 173 f., 272 f.) reconstructs selection patterns based on specific incomplete or completed logical representations, which are called mental models (Johnson-Laird and Byrne, 1991, 1992, 1995, 2002). MM theory predicts *p* patterns for incomplete represented conditionals and falsificationist *p* & *non-q* card selection patterns for a fleshed out model with an explicit representation of the false consequent (Johnson-Laird & Byrne, 1991, 79-80). This prediction for a complete and incomplete representation can also be applied to conditional obligations (Johnson-Laird & Byrne, 1992, 1995, 2002). Connected to the debate about perspective effects (Manktelow & Over, 1991; Cheng & Holyoak, 1985; Cheng et al., 1986; Cosmides, 1989; Cosmides & Tooby, 1992; Gigerenzer & Hug, 1992) MM theory has explicitly been extended to include conditional permissions (Johnson-Laird & Byrne, 1992, 1995, cf. 2002). Only Bucciarelli and Johnson-Laird

(2005) provided a MM account of a conditional prohibition (see below, and cf. pp. 275 f. for details). However, Bucciarelli and Johnson-Laird were concerned with syllogistic reasoning, not with the WST. Thus, it may be not completely clear what selections are predicted in the case of a prohibition rule.

In any case, the results for the obligation alone cannot be explained by the postulated complete or incomplete mental models of a conditional obligation, without introducing the focus concept advocated here. MM theory may account for the *p* & *non-q* selections in the cheater detection condition of the obligation rule, if a fleshed out model is assumed (cf. Johnson-Laird, 1991). Moreover, it seems that MM theory may also postdict the results of the cooperator detection condition by an assumed incomplete representation. One only needs the additional assumption that not deontic rules in general (cf. Johnson-Laird & Byrne, 1991), but only cheater contexts elicit a fully fleshed-out model (cf. Cosmides, 1989; Gigerenzer & Hug, 1992; Johnson-Laird & Byrne, 1992, 1995). But the assumption of an incomplete model of a conditional in the cooperator condition can only explain the absence of the *non-q* selections, not the significant increase of *q* selections. For *p* & *q* selections, MM theory of the WST (cf. p. 11) has postulated an incomplete *biconditional* representation of the tested rule. However, there is no reason why a cooperator focus should lead to a biconditional understanding and the cheater focus to a conditional one (*p* and *non-q* selections refer to a conditional interpretation), since all other things were kept equal. This argument is equally applicable in case of Experiment 11 and in case of Experiment 12.

Moreover, MM theory cannot explain the double (conjunctive) focus effects tested in Experiment 12, even if we evaluate the results for the obligation rule alone. MM theory has neither explicitly predicted nor found a significant increase of *p* & *q* & *non-q* selection patterns for deontic WSTs. It is essential to the MM theory of the WST that counterexamples are selected if the mental model is fleshed out so that the counterexample becomes subjectively represented (e.g., Johnson-Laird & Byrne, 1991, p. 80). On this basis, MM theory may explain why there are more falsifying (*non-q*) selections in the double focus condition than in the cooperator condition, since mentioning cheating may have fleshed out the conditional obligation. But if this would have been the case, there should be no additional *q* selections relative to the cheater condition. Contrary to this implication of MM theory and consistent with FDL theory there were more *q* selections in the double focus condition than in the cheater condition. For the obligation rule, Experiment 12 shows an increase of *p* & *q* & *non-q* selection in the double focus condition relative to both the cheater and the cooperator condition (Table 65). Also if a biconditional interpretation is assumed, based on the *q* selections, this leads to inconsistencies, since *q* selections together with the *non-q* selections would imply a *fleshed-out* biconditional model. However, this would also imply a significant increase of *non-p* selections, which has not been found. Hence, despite some degrees of freedom in the theory, MM theory cannot account for the double focus effects in the obligation rule. The found selection patterns can only be explained by a double focus on a given representation and not by the selective representations postulated by current MM theory.

Finally, also my reversed results in the prohibition rule (cf. Table 71, Table 73) and the found interaction of rules and foci in Experiment 11 and 12 are problematic for MM theory.

However, MM theory is only applicable to prohibition rules, if we make use of the recent proposal of Bucciarelli and Johnson-Laird's (2005, cf. General Discussion of Part III for details; pp. 275 f.). Although this proposal was made in the context of deontic syllogistic reasoning, prediction can be derived using the traditional domain-general MM theory of the WST (Johnson-Laird, & Byrne, 1991, 1992, 1995). Bucciarelli and Johnson-Laird did not change the domain-general MM account of the WST. Hence, MM theory has not advocated different foci within a given representation and it has not broken with its universal adherence to normative falsificationism. But Bucciarelli and Johnson-Laird (2005) did formulate the incomplete mental model of a prohibition (cf. Table 75, pp. 275 f.). This model is made out of a forbidden case *p & q*, an allowed case *p & non-q*, and an ellipsis. This proposal, if applied to the WST in the standard way, may account for the relatively high rate of 'falsificationist' *q* (cheater) selections also in the cooperator focus condition of the prohibition rule. However, this proposal cannot account for any of the focus effects in the prohibition rule without adopting a focus mechanism itself. Based on the mental model of the prohibition, selections of the falsifying *p* & *q* cards would have to be predicted equally often for all conditions. Therefore, MM theory cannot account for the found interaction of a cheater and a cooperator focus with an obligation or prohibition rule found both in Experiment 11 and Experiment 12.

MM theory has not adopted any focus mechanism independent of representations, neither in a domain-specific nor in a domain-general way. Moreover, to adopt such a mechanism would mean to give up the essential claim of the MM theory of the WST that selections are – in a domain-general way – exclusively based on complete and incomplete representations (cf. Johnson-Laird & Byrne, 1995). Nonetheless, such an extension of MM theory may well be reasonable, but it is questionable what would remain of the MM theory of deontic WSTs, if this would be done.

In conclusion, the reversed flexible focus effects in the obligation and prohibition rules and the found double focus effects in Experiment 12 are neither explicable by social contract theory, by pragmatic reasoning theory, nor by mental model theory.

Other Theories of the WST and the Results – an Overview

After having discussed the predictions of FDL theory and subsequently the predictions of the main alternative theories, it becomes possible to visualise the predictions of these theories to allow a simultaneous comparison to the results of Experiment 12. The reasons for the predictions were discussed before. The visualised predictions and results partly have to be presented in a simplified manner to allow also for a quick and schematic comparison.

On the level of *card combinations* Figure 18 shows an overview of the results and the predictions of FDL theory, PRS theory, SC theory and MM theory. All theories only make qualitative predictions whether a selection pattern is expected to be high or low. (For the visualisation, a chance selection rate was generally assumed to be 10 % for each of the three main card combinations and 20 % cumulative for the remaining selections.) Descriptively, the schematic comparison of Figure 18 shows that FDL theory provides a better fit to the data then all the other theories, particularly with regard to the obligation rule conditions, but also for the prohibition rule conditions.

For the level of card combinations, Table 70 (obligation) and Table 71 (prohibition) summarise the earlier presented results for the tested comparisons of specific card selection patterns and combine this with an overview whether the discussed theories would have predicted this result. The tables present in a condensed form that the findings are problematic for all other discussed theories and that only FDL theory can account for most of the data.

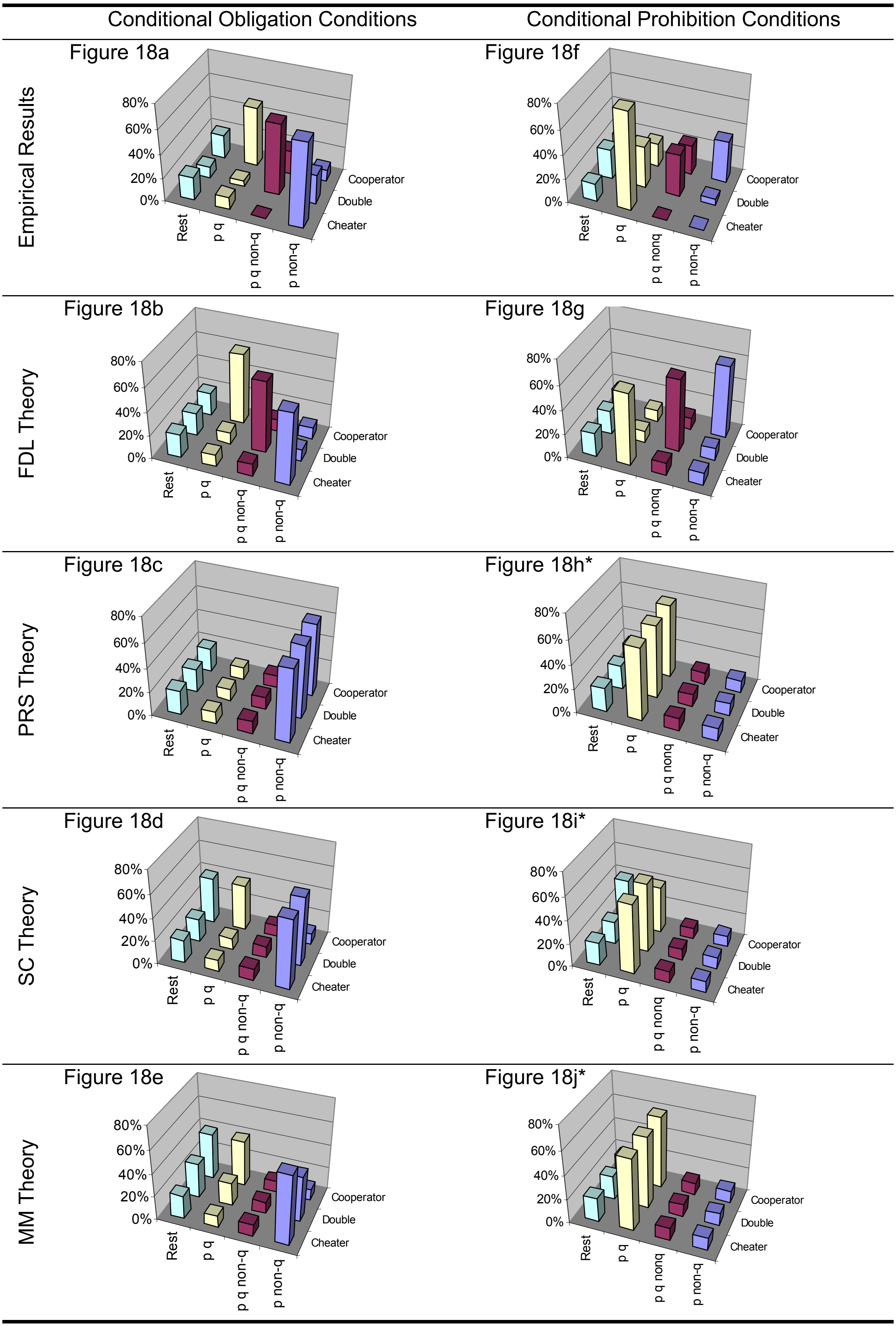


Figure 18. Bar graphs of *p & q, p & q & non-q,* and *p & non-q* selection patterns in the cheater, cooperator and double focus conditions of the obligation and prohibition rule together with predictions for FDL, PRS, SC and MM theory. (*These theories have not made explicit predictions for a prohibition rule. Cf. text for details.)

Table 70

Overview of the Results for the Comparisons of Card Patterns in the Obligation Rule and the Predictions of FDL, PRS, SC, and MM Theory

The results of the comparisons	FDL	PRS	SC	MM
$f(p \,\&\, \neg q)_{\text{Cheater}} > f(p \,\&\, \neg q)_{\text{Cooperator}}$	✓	-	✓	✓
$f(p \,\&\, q \,\&\, \neg q)_{\text{Cheater}} = f(p \,\&\, q \,\&\, \neg q)_{\text{Cooperator}}$	✓	✓	✓	✓
$f(p \,\&\, q)_{\text{Cheater}} < f(p \,\&\, q)_{\text{Cooperator}}$	✓	-	✓	✓
$f(\text{rest})_{\text{Cheater}} = f(\text{rest})_{\text{Cooperator}}$	✓	✓	-	(-)
$f(p \,\&\, \neg q)_{\text{Cheater}} > f(p \,\&\, \neg q)_{\text{Double}}$	✓	-	-	(✓)
$f(p \,\&\, q \,\&\, \neg q)_{\text{Cheater}} < f(p \,\&\, q \,\&\, \neg q)_{\text{Double}}$	✓	-	-	-
$f(p \,\&\, q)_{\text{Cheater}} = f(p \,\&\, q)_{\text{Double}}$	✓	✓	✓	(-)
$f(\text{rest})_{\text{Cheater}} = f(\text{rest})_{\text{Double}}$	✓	✓	✓	(-)
$f(p \,\&\, \neg q)_{\text{Double}} = f(p \,\&\, \neg q)_{\text{Cooperator}}$	✓	✓	-	-
$f(p \,\&\, q \,\&\, \neg q)_{\text{Double}} > f(p \,\&\, q \,\&\, \neg q)_{\text{Cooperator}}$	✓	-	-	-
$f(p \,\&\, q)_{\text{Double}} > f(p \,\&\, q)_{\text{Cooperator}}$	✓	-	✓	✓
$f(\text{rest})_{\text{Double}} = f(\text{rest})_{\text{Cooperator}}$	✓	✓	-	(-)

Note: The first column describes the achieved results for the comparisons of the cheater detection condition, the cooperator detection condition and the double focus condition. Column two to five list whether the result has been predicted by FDL theory, PRS theory, SC theory, or MM theory. Unclear predictions are presented in brackets.

Table 71

Overview of the Results for the Comparisons of Card Patterns in the Prohibition Rule Together with the Predictions of FDL, PRS, SC, and MM Theory

The results of the comparisons	FDL	PRS[b]	SC[b]	MM[b]
$f(p \,\&\, \neg q)_{\text{Cheater}} < f(p \,\&\, \neg q)_{\text{Cooperator}}$	✓	-	-	-
$f(p \,\&\, q \,\&\, \neg q)_{\text{Cheater}} < f(p \,\&\, q \,\&\, \neg q)_{\text{Cooperator}}$	-[a]	-	-	-
$f(p \,\&\, q)_{\text{Cheater}} > f(p \,\&\, q)_{\text{Cooperator}}$	✓	-	✓	-
$f(\text{rest})_{\text{Cheater}} = f(\text{rest})_{\text{Cooperator}}$	✓	✓	-	✓
$f(p \,\&\, \neg q)_{\text{Cheater}} = f(p \,\&\, \neg q)_{\text{Double}}$	✓	✓	✓	✓
$f(p \,\&\, q \,\&\, \neg q)_{\text{Cheater}} < f(p \,\&\, q \,\&\, \neg q)_{\text{Double}}$	✓	-	-	-
$f(p \,\&\, q)_{\text{Cheater}} > f(p \,\&\, q)_{\text{Double}}$	✓	-	-	-
$f(\text{rest})_{\text{Cheater}} = f(\text{rest})_{\text{Double}}$	✓	✓	✓	✓
$f(p \,\&\, \neg q)_{\text{Double}} < f(p \,\&\, \neg q)_{\text{Cooperator}}$	✓	-	-	-
$f(p \,\&\, q \,\&\, \neg q)_{\text{Double}} = f(p \,\&\, q \,\&\, \neg q)_{\text{Cooperator}}$	-[a]	✓	✓	✓
$f(p \,\&\, q)_{\text{Double}} = f(p \,\&\, q)_{\text{Cooperator}}$	✓	✓	-	✓
$f(\text{rest})_{\text{Double}} = f(\text{rest})_{\text{Cooperator}}$	✓	✓	-	✓

Note: The first column describes the results for the comparison of card combinations between the cheater detection, the cooperator detection, and the double focus condition. Column two to five list whether these results have been predicted by FDL theory, PRS theory, SC theory or MM theory.

[a] It has been shown that these deviations can be explained in a way which is consistent with FDL theory.

[b] PRS theory, SC theory and traditional MM theory have all made no explicit predictions for prohibitions. The predictions made for MM theory are derived from Johnson-Laird and Bucciarelli (2005). See text for details.

Level of single cards. Figure 19 shows bar graphs for the single *q* and *non-q* selections together with the predictions of FDL theory and of the discussed alternative theories. For each theory it is shown in Figure 19 whether a high or a low level of card selections is expected (assuming a general error level of 20 %).

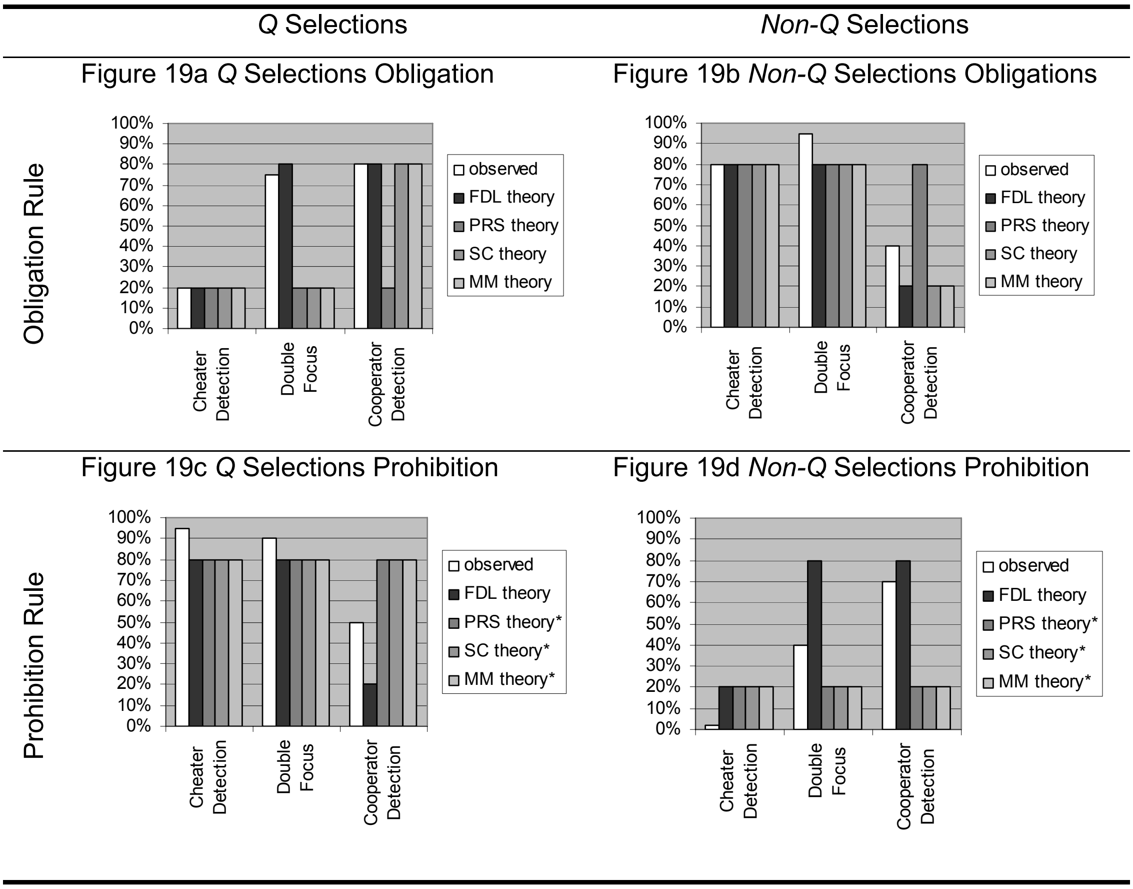


Figure 19. Bar graphs of percentage of *q* and *non-q* selections in the obligation and the prohibition rule with the predictions of FDL theory, PRS theory, SC theory and MM theory. See text for details. (*These theories have made no explicit predictions for a prohibition rule. See the earlier discussion of these theories.)

Particularly, in the double focus condition of the obligation rule all other mentioned theories would have needed to predict a lower number of *q* (cooperator) selections than in the cheater condition: According to PRS theory, an obligation rule schema should generally activate *p* and *non-q* selections only. According to SC theory, a cheater detection algorithm should have been triggered in both cases, leading to *p* & *non-q* selections only. Also according to MM theory only these selections and no *q* selections are predicted, since the deontic context and the cheater context should have fleshed out the model. The additional appearance of q selections is only intelligible based on a flexible focus mechanism. Another unique prediction of FDL theory is the increase of *non-q* selections in the double focus and in the cooperator conditions of the prohibition rule.

Table 72
Overview of the Results for the Comparisons of the Single Card Selections in the Obligation Rule Together with the Predictions of FDL, PRS, SC, and MM Theory

The results of the comparisons	**FDL**	**PRS**	**SC**	**MM**
$f(non\text{-}q)_{\text{Cheater}} > f(non\text{-}q)_{\text{Cooperator}}$	✓	-	✓	✓
$f(q)_{\text{Cheater}} < f(q)_{\text{Cooperator}}$	✓	-	✓	✓
$f(non\text{-}q)_{\text{Cheater}} = f(non\text{-}q)_{\text{Double}}$	✓	✓	✓	✓
$f(q)_{\text{Cheater}} < f(q)_{\text{Double}}$	✓	-	-	-
$f(non\text{-}q)_{\text{Double}} > f(non\text{-}q)_{\text{Cooperator}}$	✓	-	✓	✓
$f(q)_{\text{Double}} = f(q)_{\text{Cooperator}}$	✓	✓	-	-

Note: The first column describes the results for the comparisons of card selections between the cheater detection, the cooperator detection and the double focus condition. Column two to five list whether these results have been predicted by FDL theory, PRS theory, SC theory or MM theory.

The results of the tested comparisons on the single card level are summarised in Table 72 (obligation rule) and Table 73 (prohibition rule) together with an overview whether these results had been predicted by FDL theory or other theories. The tables show that FDL theory can account for more results than any other discussed theory.

Table 73
Overview of the Results Together with the Predictions of FDL, PRS, SC, and MM Theory for the Comparisons of Single Card Selections in the Prohibition Rule

The results of the comparisons	**FDL**	**PRS**[b]	**SC**[b]	**MM**[b]
$f(non\text{-}q)_{\text{Cheater}} < f(non\text{-}q)_{\text{Cooperator}}$	✓	-	(-)	-
$f(q)_{\text{Cheater}} > f(q)_{\text{Cooperator}}$	✓	-	(-)	-
$f(non\text{-}q)_{\text{Cheater}} < f(non\text{-}q)_{\text{Double}}$	✓	-	-	-
$f(q)_{\text{Cheater}} = f(q)_{\text{Double}}$	✓	✓	✓	✓
$f(non\text{-}q)_{\text{Double}} = f(non\text{-}q)_{\text{Cooperator}}$	✓[a]	✓	(-)	-
$f(q)_{\text{Double}} > f(q)_{\text{Cooperator}}$	✓	-	(-)	-

Note: The first column describes the comparisons of card selections between different focus conditions. Column two to five list whether these results have been predicted by FDL theory, PRS theory, SC theory or MM theory. Unclear predictions are given in brackets.
[a] The marginal significance of this comparison may be explained in a way which is consistent with FDL theory.
[b] PRS theory, SC theory and traditional MM theory have all made no explicit predictions for prohibitions. The predictions made for MM theory are derived from Johnson-Laird and Bucciarelli (2005). See text for details.

The schematic comparison of the predictions of the different theories have expressed the results of the earlier discussion in a pronounced way: despite a few deviations from the predictions of FDL theory, the findings clearly favour FDL over all discussed alternative theories of deontic WSTs.

13 General Discussion of Part III: The Flexible Deontic Logic of the WST

13.1 The Flexible Deontic Logic of the WST

The five experiments of this part have supported some main claims of the postulated flexible deontic logic theory of testing prescriptive rules.

In Experiment 8 and in Experiment 9 the postulated differences for the standard test strategies in prescriptive and descriptive tasks have been investigated. In Experiment 8, as predicted, full frequency effects were found for the descriptive version of a social rule. Moreover, the frequency effects partially disappeared in the prescriptive task conditions. It has been argued that this specific pattern of frequency effects can only be due to some kind of rational frequency effects, which are fully consistent with the dichotomy postulated by FDL theory. In Experiment 9 it was tested whether an exclusion of these additional frequency effects leads to the predicted contrast of two test strategies. This was the case. Experiment 8 and Experiment 9 taken together confirm the postulated existence of at least two different test strategies for the standard testing of prescriptive and descriptive WSTs and suggest a more elaborated typology of frequency effects.

All further experiments in the part were exclusively concerned with tasks that were clearly formulated as prescriptive tasks.

In Experiment 10 the testing of prescriptive conditionals was shown to be determined by a deontic logic of checking different ‘forbidden’ cells of an ought table. Although participants invariably had to check *conditional* rules, the forbidden cell was varied according to deontic logic. All four possible types of conditionals with one forbidden cell were tested: a conditional prohibition rule, a conditional obligation rule, a conditional permission rule and a conditional permission to refrain rule. For all four forbidden cells of an ought table the predicted card combinations were obtained and became dominant: *p* and *q* (conditional prohibition), *p* and *non-q* (conditional obligation), *non-p* and *q* (conditional permission), *non-p* and *non-q* (conditional permission to refrain). The results confirm the systematics as predicted on the basis of the flexible deontic logic theory of the WST.

In Experiment 11, the prediction of FDL theory was tested that there is an systematic interaction of the goal of cheater detection versus the goal of cooperator detection with different ought tables of deontic conditionals. Although many social sanction systems are concerned with cheater detection and punishment, there are also others that are concerned with cooperator detection and gratification. There are human reward systems concerned for example with providing laud, awards or medals. Experiment 11 demonstrated that the goal of cooperator detection can equally elicit clear-cut selection patterns. Contexts and rule formulations were used, for which the idea of punishment and gratification seemed equally possible. Despite some departures from the predictions, the results generally favour FDL theory. The predicted interaction of rule type and goal has been confirmed.

In Experiment 12 double focus conditions were introduced. Additionally, the aim was to replicate the rule-focus interaction effects of Experiment 11 with slightly altered instructions. In the obligation rule conditions, the predictions of FDL theory were fully confirmed. In the double focus condition the predominant selection pattern was *p & q & non-q*, a pattern neither ever predicted nor found to dominate the selections in deontic WSTs. Moreover, the cooperator and cheater conditions varied from this control in the way predicted. In the prohibition condition there seems to have been a generally lower level of cooperator selections and a higher level of cheater selections, leading also to double focus effects in the cooperator condition. Although most predictions were confirmed, a residual cheater detection bias in the prohibition rule condition was found. This can be explained in a way which is coherent with FDL theory. The formulation of a prohibition, using the word ‘forbidden’, seems not only to convey a deontic rule, but also some tendency to adopt a cheater focus. Here the subjective focus seems not only to be based on the focus instruction but to some extent also on the formulation or connotation of the rule. Nonetheless, also in the prohibition rule condition reverse differences between the cheater and the cooperator conditions, and also double focus effects, have been observed. Overall, Experiment 12 supports the predictions of FDL theory.

In summary, the results of the three experiments confirm central predictions of FDL theory. Despite some deviations, a flexible focus combined with the different ought tables of deontic conditionals seem to account for essential aspects of testing deontic conditionals. The results favour FDL theory particularly over the cheater detection approach (Cosmides, 1989; Cosmides & Tooby, 1992; Gigerenzer & Hug, 1992). Although there may indeed be a residual cheater detection bias in some tasks, the irrational concept of specific schemas and of an isolated cheater detection module appears to be explicable by the more general and systematic approach of a deontic logic combined with focus effects. It appears that the *zoon politicon* is much more rational in testing prescriptive rules than most psychological theories of the WST have assumed.

Although there are many (mathematical and psychological) open questions in regard to how deontic logic is to be formulated in detail (e.g., Hilpinen, 1970, 1981; Nortmann, 1989; von Wright, 1994; cf. Bucciarelli & Johnson-Laird, 2005), the results of the experiments show that deontic logic does provide a possible basis for explaining deontic WSTs. Deontic logic contains propositional logic but also assumptions about the relations between prohibitions, obligations and permissions (cf. e.g., Hilpinen, 1970, 1981). The very fact that deontic logic is often seen as a more powerful logic than propositional logic raises interesting questions about rationality. Whereas propositional logic would treat different deontic conditionals, like a conditional permission and a conditional prohibition, in the same way, deontic logic is sensitive about differentiating between these conditionals. A difference in content becomes a difference in form. Moreover, FDL theory distinguishes different kinds of foci. The extensions blur the distinction between content and form, and raise interesting questions concerning the logicality of the world and the possibilities of enriching logical systems.

The flexible deontic logic theory of testing prescriptive rules is to be understood as a synthesis of domain-general and domain-specific ideas.

On the one hand, a deontic logic and a flexible focus mechanism on particular cells of an ought table are clearly more systematic and more general processes than the assumption of an irrational cheater detection algorithm or the isolated schemas of pragmatic reasoning schema theory. Thus, despite all associations with these theories, the FDL account is also opposed to current domain-specific approaches. It is based on the normative deontic logic and a flexible focus mechanism and aims to provide a more general account than current domain-specific theories.

On the other hand, although the FDL theory postulates partly normative and more general mechanisms within the realm of deontic reasoning than current domain-specific theories, it is a more domain-specific theory than current domain-general theories. Firstly, although deontic logic is a kind of logic it is by definition a logic for the prescriptive realm. Secondly, the focus mechanism investigated here is understood as the specific standard mechanism for testing prescriptive rules, which may be distinguished from normal (Bayesian) hypothesis testing (of descriptive rules). The experiments on frequency effects confirmed at least two (if not more) test strategies, which we need to distinguish. Finally, despite its emphasis on relatively general concepts like ought tables and a flexible focus, FDL theory acknowledges the importance of domain-specific knowledge in the process of *constructing* ought tables or in constructing a focus in the first place. For example, the seriousness of rule violations or the praiseworthiness of following a rule might differ depending on a variety of factors, including biological preparedness, social role, individual values, interests, justifications, custom, utility, upbringing, culture, ethics, and religious belief systems. More research is needed to understand this integration of specific knowledge with the here proposed and supported flexible deontic logic of the testing of prescriptive rules.

13.2 Discussion of Other Theories of the WST

Although it has been emphasized that flexible deontic logic theory is not a *creatio ex nihilo* but a synthesis of ideas, the predictions of FDL theory differ from all current formulations of other accounts of the WST. The combined results of the five experiments of this part are inconsistent with social contract theory, pragmatic reasoning theory and with current mental model theory. They also challenge relevance theory or decision theoretic accounts. The results either disconfirm these theories or they show a necessity to extend these theories. Future research will have to show how FDL theory may be combined with aspects of these other theories or whether other theories can perhaps be revised to become compatible with the presented findings.

Evolutionary Social Contract Theory (SC Theory)

SC theory (Cosmides, 1989; Cosmides & Tooby, 1992; cf. pp. 14 f., 175 f.) differs in crucial aspects from flexible deontic logic theory and our results favour FDL theory over SC theory.

The Main Claims of SC Theory

Apart from ‘ostensible links to evolutionary biology’ (Lloyd, 1999, 213) the SC theory of the WST can in my opinion be characterized by three main predictions (for critical evaluations of the empirical content of SC theory cf. Cheng & Holyoak, 1989; Pollard, 1990; Lloyd, 1999; Almor, 2003):

Firstly, the proponents of SC theory have originally argued that clear-cut selection patterns are only found in social exchange situations: „No thematic rule that was not a social contract had ever produced a content effect that was both robust and replicable.“ (Cosmides & Tooby, 1992, p. 183; cf. Cosmides, 1989, pp. 200)

This has been criticized early on (e.g., by Cheng & Holyoak, 1989). In addition, precaution rules, like "If you clean up split blood, then you must wear rubber gloves" (Manktelow & Over, 1990; Cheng & Holyoak, 1989), can elicit clear-cut selection patterns and it would be arbitrary to call them social contracts. However, precaution rules may be understood as prescriptive rules (here a conditional obligation) for ones own security or well being. This can be seen as problematic for SC theory (for a different position, cf. Fiddick, 2004). However, the name-giving concept of a social contract anyway lost its essential position for SC theory. Gigerenzer & Hug (1992, cf. Cosmides & Tooby, 1992) dissociated social contracts and cheater detection and showed that an activation of the assumed cheater detection module is the crucial theoretical factor. „The key concepts of a social contract rule and a cheater-detection algorithm can be experimentally separated. The decisive concept is the cheater detection algorithm“ (Gigerenzer & Hug, 1992, p. 145, cf. p. 165).

Secondly, SC theory claims that the empirical selection patterns in WSTs are not coherent with any rational calculus, like formal propositional logics. Particularly their own results showed "that social contract performance is not caused by the activation of a logical faculty" (Cosmides & Tooby 1992, 189).

Thirdly, and most important, it is claimed that selection patterns in the WST are determined by an innate specific adaptation (Darwinian algorithm) of cheater detection. This postulated cheater detection module is essential to SC theory (cf. Gigerenzer & Hug, 1992; Cosmides & Tooby, 1992; Almor, 2003). It was even argued that a cheater detection or violation detection strategy has become uncontroversial in the literature on deontic WSTs (Cummins, 2000).

In the following, we first discuss some main predictions of SC theory empirically, in the light of the obtained experimental results. Subsequently, some (further) aspects of SC theory are briefly discussed theoretically.

SC Theory and the Empirical Results of Part III

The results of the experiments in Part III contradict central predictions of social contract theory.

(a) Experiment 8 and Experiment 9 have shown that there are different test strategies for testing prescriptive rules with an *a priori* focus and normal descriptive rules. SC theory may account for the constantly high cheater selection rate in the prescriptive conditions of Experiment 9, since here the postulated cheater detection algorithm may have been elicited. This may also explain the difference to one of the two descriptive conditions in the experiments. However, two aspects of Experiment 8 and Experiment 9 are inexplicable based on SC theory. Firstly, SC theory cannot explain the frequency effects found for descriptive tasks, here even found for descriptive social contracts. Although SC theory is mainly concerned with social contracts, it has not been completely silent on descriptive tasks (they served as control conditions). In contrast to our predictions, SC theory has generally predicted chaotic default selections in descriptive tasks, particularly *p* selections or *p* & *q* selections (cf. General Discussion of Part II, pp. 175 f.). Secondly, SC theory cannot explain the different kind of frequency effects for deontic conditionals found in Experiment 8. In my discussion, I have shown that this effect is fully consistent with the advocated dichotomy. Based on the postulated modular cheater detection algorithm, which is either elicited or not elicited, SC theory cannot account for such effects.

(b) Experiment 10 shows that domain effects and cheater detection appear to be embedded into a more general deontic logic (cf. Beller, 2001, 2003; Bucciarelli & Johnson-Laird, 2005). Deontic logic, an essential part of the logic of practical reasoning, provides a much more systematic and rational account of deontic WSTs than it has ever been asserted by the evolutionary account of SC theory (e.g., Cosmides, 1989, 229; Cosmides & Tooby, 1992, 189; Fiddick, Cosmides, & Tooby, 2000; Fiddick, 2004). SC theory would not have predicted the novel conditional schemas derived from ought tables. Here for the first time the four deontic types of conditionals with one forbidden cell were tested on equal footing in a WST. Cosmides has even explicitly limited her SC approach to permissions and obligations: “all social contracts are either permission or obligation rules” (Cosmides, 1989, 235). Hence, the advocated and confirmed systematics of four predicted types of clear-cut selection patterns goes beyond the explicit formulation and the spirit of evolutionary social contract theory.

(c) Experiment 11 has to be interpreted as a further strike against the account of Cosmides and Tooby. Although the results were not perfect, the focus instructions (cheater versus cooperator focus) had a strong effect in interaction with the type of deontic conditional. Since the focus had different effects on different types of deontic conditionals, a flexible focus appears to explain cheater detection effects (cf. discussion of Experiment 11).

(d) Experiment 12 replicated the findings of Experiment 11 and additionally provided evidence for double focus effects, which even more clearly cannot be explained by SC theory. In checking an obligation rule a double focus, that is, the simultaneous goal of cheater and cooperator detection, elicited a *p* & *q* & *non-q* pattern – not predicted by SC theory. For this rule, there were more *non-q* cheater selections in both the cheater and the double focus condition than in the cooperator condition. SC theory in both cases would need to assume that the cheater detection algorithm had been elicited. But in this case SC theory cannot explain why there were differences between this cheater and the double focus condition. Without postulating another focus it is inexplicable, why there were substantially more *q* selections in the double focus condition than in the cheater condition. SC theory cannot explain the found double focus effects as it cannot explain the replicated interaction pattern of rules and foci (cf. discussion of Experiment 12).

Although there was evidence for the interaction of foci and rules in both Experiment 11 and Experiment 12 there was also a residual cheater detection bias in the prohibition rule of both experiments. The term ‘forbidden’ presumably led to a moderate tendency to prefer or to add a cheater focus. Such a bias combined with the found focus and even double focus effects has not much in common with the original prediction concerning the cheater-detection algorithm of SC theory. Nonetheless, the explanation of this phenomenon by FDL theory relies on domain-specific considerations. Only if a context allows for the construction of a goal a corresponding focus is elicited. Even if there is a system of deontic logic and of different foci, it is still an important and one may say domain-specific matter to determine the context conditions under which goals are elicited and applied in the first place (similar: Gigerenzer, personal communication, 9th September 2005). Despite all differences, in this respect FDL theory is fully in line with a domain-specific approach.

Moreover, FDL theory is not opposing the idea that there are quite different social regulation systems of distributing the burdens and the benefits of cooperation and that these systems are, for example, linked to different kinds of emotions as proximate mechanisms (Fehr & Gächter, 2002; Fiddick, 2003, 2004; Beller, Bender & Kuhnmünch, 2005). However, differently to the latest defence of a highly modular approach (Fiddick, 2003, 2004) I think this needs not to be an argument against the existence of more general and integrated mechanisms, like a deontic logic and a flexible focus. Moreover, whether cheater detection and cooperator detection in their evolution may have had different evolutionary *roots* remains an open question; but the evidence provided here in regard to their present function, speaks against a modular view and favours a more general focus mechanism (cf. Sperber & Girotto 2002, 2003), which here has been elaborated and combined with deontic logic.

Overall, the experiments favour much more systematic and rational mechanisms than advocated by SC theory. An integration of deontic logic and a general flexible focus mechanism differs essentially from the assumption of a specific cheater detection module.

Theoretical Problems of SC Theory

Apart from the question of its empirical truth it is an important theoretical question, whether the claims of Cosmides and colleagues can actually be derived from evolutionary biology. This question is of particular importance since critics have claimed that for SC theory „ostensible links to evolutionary biology – rather than the experimental evidence – are doing much of the work of eliminating rival psychological hypotheses“ (Lloyd, 1999, 213; cf. also Cheng & Holyoak, 1989; Almor, 2003). Although I do regard it as a merit of Cosmides and colleagues that they have connected the rationality debate around the WST with debates in evolutionary biology, I support the view that some central evolutionary assumptions of SC theory are not warranted by evolutionary theory (cf. Cheng & Holyoak, 1989; Pollard, 1990; Chater & Oaksford, 1996; Lloyd, 1999; Almor, 2003). There are assumptions of Cosmides & Tooby (1992) that are uncontroversial, for instance that Pleistocene hunter-gatherer communities faced the ‘adaptive problem’ to secure cooperation in small groups of non-relatives. Nonetheless, in my view, the following three aspects of SC theory cannot be directly derived from evolutionary biology:

(a) Based on evolutionary considerations, and particularly on the theory of direct reciprocity (Trivers, 1971; Axelrod & Hamilton, 1981), Cosmides and Tooby (1992) have *excluded the possibility of true altruism* (cf. Dawkins, 1982). However, the altruism debate and the partly linked unit of selection debate are undecided and highly controversial (e.g., Hull & Ruse, 1998; Sober & Wilson, 1998; von Sydow, 2001; Fehr & Fischbacher, 2003). For instance, E. Sober and D. S. Wilson (1998) have prominently defended evolutionary models of sacrificing biological altruism on the population level – taking into account the problem of invasions of egoistic strategies (cf. v. Sydow, 2001). Fehr and Gächter (2002) have shown that even cheater detection in humans – paradoxically the central mechanism of SC theory – is often to be interpreted as altruistic behaviour (cf. also Fehr & Fischbacher, 2004). Moreover, also in a biological context there are findings that do support the existence also of positive truly altruistic behaviour (e.g., Sober & Wilson, 1998, Warneken & Tomasello, 2006; see Fehr & Fishbacher, 2003, for an overview). This does of course not imply a general predominance of altruistic behaviour. Nonetheless, it is unwarranted, to predict the absence of altruism *generally* on evolutionary grounds. Psychology of higher cognition should not ignore evolutionary theory, but needs to become aware of its controversies.

(b) Cosmides and colleagues justified their prediction of a *specific module* of cheater detection on evolutionary grounds. FDL theory accepts the importance of cheater detection to stabilize cooperation in groups of non-relatives, but – in appropriate contexts – also assumes the importance of cooperator detection and even altruist detection. Although cheater detection and cooperator detection may have domain-specific evolutionary roots, these mechanisms appear to have become integrated during evolution into a quite unified and rational ability to focus on different cells of an ought table. Almor (2003) argued that the flexibility of evolutionary contexts would rather favour "a complex and flexible mechanism" than a specific module – FDL theory advocates such a mechanism.

(c) Finally, it has to be questioned whether the cheater detection 'algorithm' or other more complex deontic reasoning faculties need to be inherited or whether they may also partly be learned (cf. Cheng & Holyoak, 1989; Chater & Oaksford, 1996; see also Cosmides & Tooby, 1992). Sanction systems obviously have evolutionary, historical, cultural, linguistic, biographical and even moral aspects. The question of heritability should not be ignored, but needs to be addressed empirically – the standards for a proof of inheritance in humans should not fall short of the standards for drosophila or grey geese.

Hence, although it is honourable that Cosmides established the need to link evolution and higher cognition, the discussed empirical and theoretical problems of Cosmides' approach are reminiscent of a historic paper of the biologists Gould and Lewontin (1979), in which they have warned that many adaptationist predictions boil down to little more than 'just-so stories' – referring to Rudyard Kipling's fables that offered fanciful explanations for certain animals' distinctive qualities. (On the problem of tautological formulations of adaptation, see: Popper, 1974; Gould & Lewontin, 1979; Dawkins, 1982; von Sydow, 2001.)

In conclusion, our empirical results disconfirm SC theory. Additionally, it has been outlined that the postulated theoretical justification of SC theory on the grounds of evolutionary biology can be called into question (cf. Cheng & Holyoak, 1989; Pollard, 1990; Lloyd, 1999; Almor, 2003).

Pragmatic Reasoning Schema Theory (PRS Theory)

Cheng and Holyoak (1985; Cheng, Holyoak, Nisbett, & Oliver, 1986; Politzer & Nguyen-Xuan, 1992; cf. Holyoak & Cheng, 1995a, 1995b) proposed the existence of specific reasoning schemas for testing rules in the WST and turned against the rational interpretation of conditionals based on propositional logic. For deontic WSTs they proposed the existence of two reasoning schemas, a permission schema and an obligation schema (for more details see Section 2.3, p. 13 f., cf. p. 176 f.).

FDL theory follows PRS theory in the important respect that it also postulates reasoning schemas that do not match the prescriptions of propositional logic. Moreover, FDL theory incorporates the proposed permission and obligation schemas.

Nevertheless, FDL theory also departs from PRS theory in four respects. In all four respects, the experiments support FDL theory. The second point is only of minor importance.

(a) PRS theory cannot account for any frequency effects in testing descriptive rules (Experiment 8, Experiment 9). Although PRS theory was mainly concerned with what I call prescriptive tasks, I have shown previously that also claims concerning descriptive tasks had been made in the proposal of PRS theory (cf. pp. 176 f.). Additionally, PRS theory cannot account for any frequency effects for prescriptive rules which have been found in Experiment 8 and which have been suppressed, as predicted, in Experiment 9. Although PRS theory would have predicted the result for the prescriptive task conditions in Experiment 9, it cannot account for the differences between Experiment 8 and Experiment 9.

(b) FDL theory has reformulated the obligation and permission schemas, first postulated by PRS theory, in terms of ought tables and a cheater focus. Hence, with respect to these schemas and this focus FDL theory and PRS theory come to largely the same predictions. Nonetheless, the reformulation by FDL theory solves a theoretical problem of PRS theory. PRS theory gave a circular account of the introduced if-then schemas by defining them by four other if-then production rules. Moreover, the reformulation makes clear that prescriptive conditionals need not to be formulated using the modals ‘may’ and ‘must’. In accordance with this view, the permission rule in Experiment 10 and the obligation rule in Experiment 12 were phrased without modals but still elicited clear-cut selection patterns.

(c) Instead of proposing two specific schemas, which explicitly should not refer to any formal theory of reasoning (cf. p. 13 f.), here a generalization and systematization is proposed, based on ought tables and deontic logic. PRS theory (Cheng & Holyoak, 1985; Holyoak & Cheng, 1995b; Cheng, Holyoak, Nisbett, & Oliver, 1986; Politzer & Nguyen-Xuan, 1992) has only predicted patterns for the obligation and the permission rule. Perhaps the modification of Holyoak and Cheng (1995) on rights and duties may be understood as a step towards a more systematic treatment. But a deontic logic of the WST was neither theoretically explicitly developed nor empirically tested. Here, explicitly based on deontic logic additional schemas of conditional prohibitions and conditional permission to refrain were derived. Experiment 1 confirmed all four predicted selection patterns, not predicted by PRS theory.

(d) The proposed *flexible* deontic logic theory combines deontic logic with the concept of checking prescriptive rules by focusing on persons who violated or persons who followed a rule. In contrast, PRS theory formulated production rules independent of the goal of cheater and cooperator detection. Thus, PRS theory cannot account for any differences found between a cheater-detection, a cooperator-detection or a double-focus condition. PRS theory cannot account for any focus effect. Because the results of Experiment 11 and Experiment 12 clearly support the predicted role of focus effects, they are not explicable by PRS theory.

In conclusion, the four aspects and the five described experiments favour flexible deontic logic theory over the more specific approach of PRS theory. The experiments show either that PRS theory is false or that it needs to be extended in the sense of a flexible deontic logic theory of deontic WSTs.

Mental Model Theory (MM Theory)

It will be shown that MM theory (cf. 11 f., 173 f.) – without substantial modifications – cannot account for the data of the experiments in this paper; particularly the findings of Experiment 12 on double focus effects are incoherent with MM theory (cf. Discussion of Experiment 12). Before turning to the experiments, MM theory will be discussed more generally in order to understand its relationship to FDL theory and to derive the MM predictions for the current experiments.

The Essential Characteristics of MM Theory of the WST

The MM theory of the WST is generally characterized by the assertion that reasoning and reasoning errors are based on rudimentary representations of the logical possibilities or impossibilities conveyed by a sentence, for instance by a conditional. These representations are called mental models. In contrast to MM theory, FDL theory assumes that selection patterns for simple deontic WSTs are not (only) based on the selectivity of representations, but on full ought tables and a flexible focus *within* a given representation. Particularly the flexible focus aspect of FDL theory (as has similarly been formulated by Sperber & Girotto, 2002, 2003) is in contradiction with current MM theory of the WST. MM theory has no concept of a focus within a given representation. FDL theory combines this focus concept with deontic logic. Hence, experiments supporting the interaction of focus effects and deontic logic provide evidence for the falsity or incompleteness of MM theory.

Nonetheless, the focus mechanism may be compatible with a MM theory of more complex situations, like syllogistic reasoning (Bucciarelli & Johnson-Laird, 2005). Although there are also other promising theories of syllogistic reasoning (cf. e.g., Chater & Oaksford, 1999b), it indeed seems plausible that processing limitations and resulting selective representations have to play a role in complex situations. Also, for instance, for judgments concerning the equivalence of two sentences, some kind of selective representations may play a role. To stress the selectiveness of representations is no doubt an important contribution of MM theory. However, in this work we are concerned with the MM theory of normal simple WSTs. MM theory is discussed in the version as formulated by its main proponents Johnson-Laird and

Byrne (1991, 1992, 1995, 2002, cf. Bucciarelli & Johnson-Laird, 2005). In Part III we concentrate on deontic WSTs. FDL theory claims that for simple deontic WSTs, the role of selective representations is to be replaced (or at least to be supplemented) by the mechanism of a focus based on deontic logic.

The traditional MM theory of the WST (Johnson-Laird & Byrne, 1991, 1992, 1995, 2002), is not only characterized by the assertion that mental representations *somehow* play a role in testing hypotheses – which indeed would be a quite general assumption – but that these representations determine card selections in the WST in a *particular* way:

Firstly, MM theory makes predictions about what is represented. In principle, the mental models of a logical connection of two objects can represent all four logical possibilities of the (semantically, pragmatically and syntactically evoked) dyadic logical connector. This corresponds to all logical cases of a truth table: *p & q, p & non-q, non-p & q, non-p & non-q* (cf. Table 1). But representations based on mental models differ from a representation of *all* logical cases in the two following respects: (a) According to a 'principle of truth' only true cases are normally represented. As mentioned earlier, the false (impossible) cases, namely the complement of the set of true cases, are inferred only with some effort. (b) Moreover, in the case of a conditional the true cases are normally also represented incompletely. It is claimed that conditionals are often initially represented only by an incomplete model of *p & q* (with a 'mental footnote' that only *p* is represented completely). Knowledge of the situation, experience with rule violations, and a deontic context may cause that these 'implicit models' are 'fleshed out' to become 'explicit models' representing all true cases.

Secondly, MM theory makes predictions about the process of selecting cards in the WST based on these given representations. Although stressing the *representation* of true cases, Johnson-Laird and Byrne (1991, 1992, 1995) have maintained a *falsificationist* norm of hypothesis testing (for a Bayesian alternative see Oaksford & Chater, 1994, 1998, 2003; Evans & Over, 1996, von Sydow, 2004). MM theory has not distinguished the testing of descriptive and prescriptive conditionals. For obligations, like for normal descriptive conditionals, only *p & non-q* selections are interpreted as correct selections. The frequently found *q* selections (found particularly for descriptive conditionals) have generally been assumed to be due to incomplete representations. Falsifying *non-q* cards would generally be selected if the corresponding case is represented: “In short, the model theory predicts that people will select the card falsifying the consequent whenever the models are fleshed out with explicit representations of that card” (Johnson-Laird & Byrne, 1991, 80). The ‘facilitation effects’ in deontic contexts are explained not by a flexible focus or by domain-specific modules but by the (domain-general) assumption that these contexts will often lead to a complete representation of the falsifying cases.

Before we come in a second step to access whether MM theory is incoherent with our data, we first need to clarify how one has to apply MM theory to the deontic conditionals tested here.

MM Theory and the Four Types of Deontic Conditionals

It is problematic to access MM theory not only because it is not precisely specified which situations should lead to fully represented deontic conditionals, but also because it needs to be determined how the (complete or incomplete) models of a conditional prohibition and a conditional permission to refrain look like in the first place. I discuss traditional MM theory explicitly formulated for the WST and subsequently an interesting recent proposal of Bucciarelli and Johnson-Laird (2005) which has been formulated in a different context. Although the latter proposal has not been proposed for or applied to the WST, we will consider it here.

(a) Traditional mental model theory of the WST (as represented by the joint work of Johnson-Laird & Byrne, 1991, 1992, 1995, 2002) to my knowledge has not explicitly treated conditional prohibitions. MM theory has integrated the obligation and permission schemas (Johnson-Laird & Byrne, 1992, 1995), also discussed in the context of pragmatic reasoning schema theory (Cheng & Holyoak, 1985, Holyoak & Cheng, 1995b) and in the context of perspective effects (cf. Manktelow & Over, 1991; Gigerenzer & Hug, 1992). The integration of these schemas by MM theory proceeded along the lines of the older idea that a conditional may be interpreted as a biconditional (cf. Johnson-Laird, 1970). This has been combined with the post hoc explanation that a biconditional, depending on one’s perspective, can be represented either as obligation or as permission, each referring only to one of the forbidden cells of the biconditional. In contrast, conditional prohibitions have not even been discussed in the 2002 paper of Johnson-Laird and Byrne, in which they provide an interesting MM account of very different usages of conditionals.

One may try to reconstruct the missing models by analogy using the domain-general MM account, which is based on (alethic or deontic) possibilities or impossibilities.

But firstly, in my point of view, the MM assumption that general (alethic) modal logic is fully equivalent to deontic logic is problematic. Although the similarity of alethic and deontic modalities has been a fertile analogy (cf. Hilpinen, 1971), recent accounts of modal logic have also elaborated some disanalogies (Hilpinen, 1981; cf. even von Wright, 1981). For example, whereas alethic modal logic normally accepts the formula $p \rightarrow Possible(p)$, deontic logic must reject the analogue formula $p \rightarrow Allowed(p)$. Moreover, without different representations for allowance and forbiddance on the one hand and for possibility and impossibility on the other hand, the differences of the *meaning* of descriptive and prescriptive rules cannot be represented.

Secondly, even if deontic logic would be directly derivable from (alethic) modal logic, it remains unclear how to construct a rudimentary model for a prohibition: If the representation of “if *p* then it is forbidden to *q*” would be constructed in analogy to a normal conditional, the mentioned *p* & *q* case would be represented. However, for a prohibition, this case is the impermissible case and its representation in an incomplete model would contradict the principle of truth. According to this principle, incomplete representations consist of possible but not of impossible (impermissible) models. Alternatively, if this impermissible case were not represented, the representation of a prohibition would be empty, which would be absurd: a prohibition differs from saying nothing. Hence, in my view traditional MM theory has no precise prediction concerning prohibitions and consequently cannot account for most of our data.

(b) Recently, Bucciarelli and Johnson-Laird (2005, cf. Beller, 2003) proposed a deontic mental model theory of syllogistic reasoning, which is also based on deontic logic, and at the same time continues to explain reasoning errors by incomplete representations. Since this account goes beyond traditional MM theory, it is not clear how to apply it to the WST. Nonetheless, when using the existing MM theory of the WST, it will be shown below that even the new kinds of representations – despite its similarities – cannot account for our findings.

The completed models of deontic conditionals are equivalent to the ought tables based on deontic logic, as postulated here. Bucciarelli and Johnson-Laird (2005) for the first time presented incomplete models of all four types of deontic conditionals, which are also treated here. Table 74 shows these complete mental models of a conditional prohibition and a conditional obligation. However, Bucciarelli and Johnson-Laird (2005) were concerned with deontic syllogistic reasoning and have not applied their account to the WST.

Table 74
Complete Mental Models of a Conditional Prohibition and a Conditional Obligation

Conditional prohibition “if *p* then one is forbidden to *q*”	Conditional obligation “if *p* then one must *q*”
[a]	*p* *q*
p *non-q*	
non-p *q*	*non-p* *q*
non-p *non-q*	*non-p* *non-q*

Note. According to Bucciarelli and Johnson-Laird (2005, Table 1).
[a] The forbidden case of the prohibition is also generally represented as an impermissible case (see text below).

The MM theory of the WST crucially depends on the formulation of the incomplete models, since any difference in card selections in the WST, also with regard to deontic WSTs, has to be based on differences in representations (Johnson-Laird & Byrne, 1991, 1992, 1995, cf. 2002). These representations (and the search for counterexamples using these representation), not any additional inference rules, have been postulated to be the very means (the algorithmic level account) of deductive reasoning (Johnson-Laird & Byrne, 1991, 35 f.).

Now the two corresponding rudimentary models of a conditional obligation and that of a conditional prohibition are presented (Table 75). The incomplete prohibition has not been formulated by Johnson-Laird and colleagues before. The incomplete obligation rule, ‘if *p* then one must *q*’, remains similar to the normal incomplete representation of a descriptive conditional. It contains a permissible model of ‘*p q*’ and an ellipsis ‘…’ (including the implicit permissible *non-p* models). In contrast, for the prohibition rule, ‘if *p* then one is forbidden to *q*’, they claim an impermissible model ‘*p q*’ and a permissible model ‘*p non-q*’ with the standard ellipsis. Please note, that the impermissible model ‘*p q*’ is assumed to be represented, although normally only positive cases are represented until the model is fleshed out (Johnson-Laird & Byrne, 1991, cf. 2002). The impermissible model is assumed salient and endowed with a mental footnote, marking it to be a negative case.

Table 75

Incomplete Mental Models of a Conditional Prohibition and a Conditional Obligation

Conditional prohibition “if *p* then one is forbidden to *q*”	Conditional obligation “if *p* then one must *q*”
p *q* (mental footnote: forbidden)	*p q*
p non-q	…
…	

Note. According to Bucciarelli and Johnson-Laird (2005, Table 2). See text for details.

Although the fleshed-out models (see Table 74) were in their presentation (also originally) depicted without forbidden cases, the forbidden case of the conditional prohibition is generally assumed to be represented.

It should be noted that this deontic theory of syllogistic reasoning departs from traditional mental model theory and that this theory has not been proposed as a theory of the WST:

Firstly, this formulation of deontic MM theory departs from the truth principle of traditional MM theory, which claimed that incomplete representations only refer to possible – true or allowed – cases (cf. Johnson-Laird & Byrne, 1992, 1995, 2002). This is problematic for MM theory, since this principle can be seen as a main unifying principle of the construction of selective models (e.g., Johnson-Laird & Byrne, 2002). Moreover, this principle has not been replaced by another one. Without such a principle, MM theory is in danger to describe data only in a post hoc way.

In my opinion, one may perhaps resolve this theoretical problem by proposing a new principle for a deontic MM theory that is specific for the realm of prescriptive rules. For the prescriptive realm the allowance to act freely may be seen as a baseline, so that it would be parsimonious not to represent all such ‘allowances’ explicitly. In analogy to the ‘principle of truth’, one may call this the ‘principle of freedom’. In contrast, it appears plausible that on the prescriptive level forbidden and clearly honourable cases are represented explicitly. Presumably, this is to be combined with a second principle of representing the mentioned cases first (‘matching principle’). Perhaps, such a MM theory might account for our data. However, current MM theory would be at odds with such a proposal.

Secondly, deontic MM theory of syllogistic reasoning is underspecified to count as a full theory of the WST. Bucciarelli and Johnson-Laird have formulated selective representations for the four types of conditionals, but they have discussed this neither theoretically nor empirically for the WST. Hence, we have to rely on former versions of MM theory (Johnson-Laird & Byrne, 1991, 1992, 1995, 2002) to extend this modified MM theory to the WST. Despite some difficulties to derive exact predictions for our WSTs (MM theory has some free parameters) it will become clear that it is possible to criticize MM theory based on the data provided here.

MM Theory and the Experiments

Subsequently, it will be shown that current MM theory of the WST cannot account for the combined obtained results of the experiments. Even if the proposal of Bucciarelli and Johnson-Laird (2005) would be applied to the WST, MM theory is not able to explain the findings. Instead, the data show that deontic MM theory is either incomplete or false. The frequency effects in Experiment 8 and Experiment 9 cannot be explained by MM theory. Experiment 10 might be seen as neutral. The interaction of rule and focus shown in Experiment 11 and particularly the double focus effects of Experiment 12 are problematic for MM theory.

(a) MM theory has not predicted frequency effects and any difference between descriptive and prescriptive task. It cannot account for the findings of Experiment 8 and Experiment 9. MM theory may account for the logical results in the prescriptive task of Experiment 9, assuming that the cheater focus led to a fleshed out model. But MM theory cannot explain the *p* versus *non-p* and *q* versus *non-q* frequency effects found in descriptive tasks, elicited by pure manipulations of frequencies (cf. General Discussion of Part II, pp. 173, for details).

(b) The results of Experiment 10 on deontic logic, always using a cheater focus, go beyond traditional MM theory (Johnson-Laird & Byrne, 1991, 1992, 1995, 2002), since it was inapplicable to prohibitions (see discussion above). Nonetheless, if the extended deontic MM theory of Bucciarelli and Johnson-Laird (2005), formulated for syllogistic reasoning, would be applied to the WST (which has not been done before), the results would be explicable by this extended MM theory. Either the general deontic context (Johnson-Laird & Byrne, 1991) or the cheater context (Johnson-Laird & Byrne, 1995) may be claimed to lead to fleshed-out models. Nonetheless, it should be noted that these assumptions and deontic logic would be the explaining components, not the selective representations postulated by MM theory. If selective representations played any role here, the results would have been different. Hence, also this MM theory can account for the results of Experiment 10 only by assuming that the specific explanatory mechanisms of MM theory, selective representations, played no role here.

(c) The results of Experiment 11 are problematic for MM theory because they show the predicted reversed difference between a cheater and a cooperator focus condition for the obligation and the prohibition rules.

Alternatively, if one assumes fleshed out models for the cheater detection conditions and incomplete models for cooperator detection conditions (cf. Table 75), MM theory is indeed able to account for two effects:

Firstly, MM theory would (like FDL theory) predict for the *obligation rule* a decrease of *non-q* selections in the cooperator condition relative to the cheater condition, since the incomplete representation in the cooperator conditions does not contain a (falsifying) model with the *non-q* case.

Secondly, for the prohibition rule MM theory may account for the unpredicted high rate of *q* card selections in both corresponding conditions, since MM theory assumes the forbidden *p q* case to be represented both in the complete as well as in the incomplete model.

Nonetheless, in regard of both points there are reasons why Experiment 11 favours FDL theory over MM theory.

Firstly, the above explanation in favour of MM theory assumes for the cooperator condition of the obligation rule a general predominance of *p* and *q* answers for deontic obligations without explicit cheater focus. However, former research on deontic WSTs makes this assumption implausible. Although a *p & q* selection pattern has been predominant for descriptive conditionals, it has not at all been the base line with regard to testing the prescriptive (concrete) obligations, even if an explicit cheater instruction is removed. This was even the case in the findings of Gigerenzer and Hug (1992, cf. their Table 15). Although the removal of the cheater focus has reduced *p & non-q* patterns in that experiment, the *p & non-q* selections remained dominant in all their tasks (all without an alternative focus) and they did not find any resulting clear-cut *p & q* pattern for a non-cheater condition of the obligation rule. Likewise, results of other authors show rather a remaining predominance of *p* & *non-q* selections for an obligation rule, if a cheater instruction was removed without introducing an alternative focus (Manktelow & Over, 1990, 1991, 1992; Johnson-Laird & Byrne 1992, 1995; Sperber, Cara, & Girotto, 1995, 85 f., and, particularly, Sperber & Girotto, 2003). Hence, the current finding of a clear-cut *p & q* selection pattern for an obligation rule in a situation with a formally unspecified focus is unique and hence is most plausibly attributed to the introduced goal of cooperator detection (cf. the discussion of SC theory in Experiment 11).

Secondly, although MM theory of the WST (Johnson-Laird & Byrne, 1991) for the conditional obligation indeed predicts a decrease of *non-q* selections in the cooperator condition relative to the cheater condition (assuming an incomplete model, because no cheater detection instruction elicited a fleshing-out of the model), it simply does not predict an increase of *q* selection. The *q* selection would have to be explained by an (incomplete) biconditional interpretation. But there is no reason why there should be a conditional interpretation for the cheater detection task and a biconditional interpretation for the cooperator task (cf. Discussion of Chapter 12, pp. 253).

Thirdly, also concerning the prohibition rule FDL theory provides a better fit to the data than MM theory. FDL theory – like MM theory – can provide a post-hoc explanation for the partial deviation from the initial FDL prediction: It seems plausible that the term "forbidden" in the prohibition rule contributes not only to the construction of the rule, but also to the construction of a focus, that is, this formulation evokes a tendency to elicit a cheater detection focus. But FDL theory additionally explains the found differences between the cheater and the cooperator conditions of the prohibition rule. MM theory has no explanation for these effects. MM theory predicts no difference between both conditions, since the complete representation (cheater condition) and the incomplete representation (cooperator condition) both contain the impermissible *p & q* model. Hence, MM theory cannot account for this difference without adopting a focus mechanism itself. For these three reasons, Experiment 11 favours FDL theory also over MM theory.

(d) Experiment 12 provides additional evidence against the sufficiency of current MM theory. It replicates the interaction of foci and rules found in Experiment 11, which, as we have seen, is problematic for MM theory. Furthermore, Experiment 12 also provides evidence in favour of novel double focus effects, when both goals (cheater and cooperator detection) were evoked simultaneously. The found *p & q & non-q* patterns have neither been predicted nor found empirically by MM theory.

For the obligation rule, the increased number of *non-q* selections in the double focus condition relative to the cooperator condition can only be explained by MM theory if it postulated a fleshed out model for the double focus condition, as postulated for the cheater condition. However, if this is done, it becomes inexplicable why the number of *q* 'cooperator' selections in the double focus condition is significantly higher than in the cheater condition. They are even about as high as in the cooperator detection condition, which is assumed to elicit an incomplete model. MM theory only postulates two kinds of representations of an obligation (an incomplete and a complete model), but here three kinds of patterns were predicted and found. Likewise MM theory cannot explain the results and the double focus effects in the prohibition condition (for details, cf. Discussion of Experiment 12).

(e) In conclusion, MM theory cannot account for the data of the presented experiments. Flexible deontic logic has predicted most of the results. No reference to selective representation is needed to account for these results. Moreover, several results were inconsistent with MM theory.

MM theory of the WST in its current formulation (Johnson-Laird & Byrne, 1992, 1995, 2002), which is presumably the most popular account of the WST, cannot explain most of the findings. The recent proposal of Bucciarelli and Johnson-Laird (2005) shares with FDL theory the reference to deontic logic. Nonetheless, if applied to the WST, also this deontic MM theory cannot account for the pattern of results without substantial modifications.

MM theory can only account for our data by advocating a focus mechanism itself. In principle, this is possible and I would favour this solution; but this would change the essence of the MM theory of the WST. Moreover, the essential explanatory mechanism, selective representations, would play no role at least in explaining the present data. Whether a revised MM theory could be formulated to delineate the selection phenomena, which are caused by focus mechanisms and those, which are caused by selective representations, is a question, which may be addressed by future research. But even if this would be successful, MM theory could not account for the probability effects found here and also in Part II.

Relevance Theory and Decision Theoretic Approaches

Relevance theory and decision theoretic approaches in principle may both be compatible with our findings, but in their current versions, they too cannot account for our findings without substantial extensions.

Relevance Theory of the WST

Relevance theory, as applied to the WST (Sperber, Cara & Girotto, 1995; Sperber & Girotto, 2002, 2003; cf. also Sperber & Wilson, 1986; cf. Evans, 1994), claims that those cards that are selected seem to be relevant in the light of what is already known. In the WST those cards are understood as being relevant which are connected with a high 'cognitive effect' and a low 'cognitive effort'. This general cognitive cost-benefit assertion brings relevance theory close to decision theoretic approaches (cf. below). But here also more specific features of relevance theory come into play.

Although relevance theory has not advocated a deontic logic, the concept of a flexible focus has been proposed and tested for the first time in this theoretical framework (Sperber & Girotto 2002, 2003). The focus aspect of FDL theory can be regarded as a direct descendent from relevance theory. Moreover, FDL theory is, in my view, fully consistent with the findings of Sperber and Girotto (2002, 2003), although they advocated focus effects without the distinction of descriptive and prescriptive rules. In contrast, FDL theory was formulated to be compatible with the results of the Bayesian approaches. I have asserted that the truth or falsity of descriptive rules is normally tested not by focusing, but in a Bayesian way. It has been proposed that this is the standard way of testing descriptive rules in the WST (cf. e.g., Oaksford & Chater, 1994, 1995, 2003; v. Sydow, 2004). However, in my view the seemingly opposed claims of relevance theory and of Bayesian theories are compatible: If one formulates WSTs as categorization tasks (Sperber & Girotto, 2002, 2003) one can indeed also find focus effects for descriptive WSTs. Only the testing of the *truth or falsity* of a descriptive rule results in a Bayesian account (cf. e.g., Oaksford & Chater, 1994, 2003; v. Sydow, 2004) and the normal deontic testing of rule compliance or rule breaking involves focusing (cf. e.g., Cosmides, 1989; Cosmides & Tooby, 1992; Cheng & Holyoak, 1985; Holyoak & Cheng, 1995b). The decisive point is the goal of the task and the used instruction (cf. Sperber & Girotto, 2003). The instruction of Experiment 2 differed only in minor aspects from the one used by Sperber and Girotto (2002, 2003). Although Experiment 2 used a balanced question to check whether the rule had been followed *or* violated, strong focus effects were achieved by varying the goal of the tester in a context of social sanctioning. For this balanced formulation the found effects were stronger than those found by Sperber and Girotto (2003, mixed WST/FCT rules).

Despite these similarities, there are also clear differences between the relevance theory of the WST and FDL theory.

Firstly, relevance theory can at best only provide a post hoc explanation of frequency effects in descriptive tasks (Sperber, Cara & Girotto, 1995; Oaksford & Chater, 1995). In Experiment 8 and 9, the observed frequency effects in descriptive tasks and the observed partial or complete reduction of frequency effects in prescriptive tasks both cannot be explained by relevance theory, since this theory does not provide predictions for mere frequency manipulations (cf. General Discussion Part II, see pp. 177 f.).

Secondly, relevance theory has not advocated a deontic logic to explain the findings in the WST and has not combined focus effects with deontic logic. Additionally, no double focus effects have been investigated, which are particularly decisive for a comparison of FDL theory with competing theories. In this respect, relevance theory can neither explain the results of Experiment 10, Experiment 11 or Experiment 12. Relevance theory has not postulated a logical systematics of different conditionals and has not combined focus effects with this systematics.

However, relevance theory not only has its own virtues, but, in my view, it is formulated openly enough that it in principle may become extended to incorporate deontic logic. Nonetheless, this would be an essential extension of relevance theory, particularly since relevance theory has explicitly tried to disconnect the WST from any normative theory of reasoning (Sperber, Cara, & Girotto, 1995, 36-39). Moreover, the link to deontic logic (cf. Bucciarelli & Johnson-Laird, 2005) would again connect relevance theory to the reasoning standard of deontic logic, which is traditionally seen as a normative account. Furthermore, our understanding of what is ‘relevant’ in the WST would substantially be altered by employing deontic logic. Hence, it is questionable, whether we would still be concerned with the same theory of relevance.

This leads us to a more fundamental problem. An extension of relevance theory to include deontic logic makes use of an almost tautological meaning of relevance. The concept of relevance would become unfalsifiable if one claimed that relevant cards are selected, while defining those cards that had become selected as relevant. Similarly, if other theories, like information gain theory or MM theory would be integrated *a posteriori* into relevance theory, one would come close to such a tautological understanding of relevance (cf. the dispute Sperber, Cara, & Girotto, 1995; Oaksford and Chater, 1995). In my opinion, the very question at issue in the WST debate is not whether subjectively relevant cards are selected, but *what* makes cards relevant. Hence, an extension of relevance theory to be combined with deontic logic may be justifiable, but would alter the core of relevance theory.

In conclusion, FDL theory does owe much to relevance theory, but our results show that current relevance theory is at least incomplete, in not combining the flexible focus mechanism with a systematics of deontic logic.

Decision Theoretic Approaches

In addition, current decision theoretic approaches to the WST (Manktelow & Over, 1990, 1991, 1992, 1995; Evans, Over, & Manktelow, 1993; Over & Manktelow, 1993; Evans & Over, 1996; cf. also Cosmides, 1989; Kirby, 1994; Oaksford & Chater, 1994; Perham & Oaksford, 2005; but also Johnson-Laird & Byrne, 1992; Fairley, Manktelow & Over, 1999; Manktelow & Over, 2000) have not predicted the found pattern of results. However, theses approaches – like relevance theory – can perhaps be modified to account for the results.

Utilities are connected with the testing of descriptive and deontic rules. Particularly for deontic rules, I do share the assumption that there is a complex relationship between utilities and deontic concepts (e.g., Over & Manktelow, 1993). This has even been conceded by authors not advocating a decision theoretic approach (Holyoak & Cheng, 1995, but cf. Johnson-Laird & Byrne, 1992). To me it is for instance plausible that the four types of deontic conditionals, discussed and tested here, do convey some default information about utilities (cf. Manktelow & Over, 1995). Moreover, in my opinion prescriptive rules are often constructed in the first place to prevent prisoner dilemma situations, and to allow participants to reach a better result than achieved by uncoordinated action. Although this has not been emphasised by former decision theoretic accounts of the WST, I would indeed subscribe to such an account.

However, the current decision theoretic accounts of the WST by Oaksford and Chater (1994, cf. Perham & Oaksford, 2005) and by Manktelow and Over (1990, 1991, 1992, and 1995) cannot (fully) account for our data for the following reasons:

Firstly, although Manktelow & Over (1991, 1992, 1995) contributed to a differentiation of deontic and descriptive rules, they did neither postulate nor investigate two different test strategies, of an a priori and an *a posteriori* (frequency dependent) focus. The main predecessor of the present distinction is Oaksford and Chater (1994), who in principle distinguished an information gain account from a utility account for specific cells. FDL theory is in my view fully compatible with their account. However, Oaksford and Chater (1994) did not elaborate and test this distinction and were later mainly concerned with descriptive tasks. (Oaksford and Chater have postulated neither a deontic logic nor explicitly a focus mechanism of cheater and cooperator detection, see below.)

Secondly, a cost-benefit analysis is incomplete without a deontic logic. Although Manktelow and Over (1995) combined their decision theoretic approach with the general idea of a deontic logic, they have not elaborated a full deontic logic of the WST. Oaksford and Chater (1994; cf. Perham & Oaksford, 2005) have made no reference to deontic logic at all.

Thirdly, these approaches have not explicitly predicted and tested the focus and double focus effects based on the goals of cooperator detection or cheater detection or both. Only in a post hoc way, one might reinterpret the found focus and double focus effects as changes in the cards' pay-offs.[62]

Fourthly, although the reasons behind the selections may well be reinterpreted as social conventions about utility distributions, the current experiments are under-specified to substantiate such a claim. The tasks did not refer to subjective utility. A proponent of a decision theoretic approach might argue that FDL theory is essentially incomplete, since it does not explain under which conditions the goals of punishment or of gratification become plausible. This is correct. Additionally, a proponent of a utility approach may argue that these goals are based on subjective utility. But does this lead us any further? If subjective utility would be defined by the individual pay-off then this indeed becomes an empirically testable claim; but this means we cannot assume its truth a priori:

My view on this matter is that the individual pay-off is only one factor determining our selection goals. The construction of our goals in the context of social sanctions is also influenced by other factors, like emotions, custom or our concepts of justice.

For instance, differently from Manktelow and Over (1990, 1992), I would predict that apart from subjective benefits and costs, considerations of justice (cf. Rawls,

[62] Perham and Oaksford (2005) even treated cooperator cases as costs only.

1999/1971; MacIntyre, 1985) also play a role in regulating the sanctions of punishment and gratification. Such an account would be in line with recent research, for instance, on the ultimatum game or the dictator game, which has shown that subjects are not only influenced by their individual pay-offs, but also by the concept of justice (e.g., Güth, Schmittberger, & Schwarze, 1982). This would also be coherent with the finding of Fehr and Gächter (2002) that humans tend to punish norm deviating behaviour, even if this is not individually advantageous ('altruistic punishment'). Thus, this question needs to be addressed empirically. Of course, if justice, altruism, deontic logic, emotions, customs, and all other aspects influencing card selection are all interpreted as contributing to subjective utility, I would advocate that the WST is 'only' based on subjective utility – but I doubt that this would still be an empirical (falsifiable) theory of the WST.

Fifthly, it has become a crucial assumption of decision theoretic approaches of the WST that people are much more sensitive to costs than to benefits. In the understanding of FDL theory this means that participants in deontic WSTs focus more on cases of rule violations than on cases of rule following (e.g., Manktelow & Over, 1990, 1992; Oaksford & Chater, 1994; Sperber, Cara, & Girotto, 1995, 85 f.). For example, Manktelow and Over (1990) tested the checking of a bingo rule "If you have a winning line, then you must shout 'Bingo' to win a prize!" This condition was intended as a strong manipulation with a high benefit for testing conforming cases. Manktelow and Over predicted that this rule should elicit high numbers of *p* and *q* answers. However, their results did not confirm this prediction. Partly based on these findings, Manktelow and Over (1992, 185) abandoned *normative* decision theory, and instead claimed "that people are more sensitive to costs than to benefits in deontic selection tasks, particularly the serious costs which can result from being cheated […]." Contrary to this revised utility approach (and in difference to the findings of Cosmides, 1989, and Sperber, Cara & Girotto, 1995, 85 f.) in Experiment 11 and Experiment 12 clear-cut focus effects have been found in obligation rules using the context of social sanctions and the goal of cooperator detection.

Finally, it remains problematic for any utility account that a high pay-off and low risks condition did not lead to *p* & *q* answers (Manktelow & Over, 1990, 1992) whereas a deontic context combined with a plausible goal of cooperator detection did.

In conclusion, although the findings appear to be inconsistent with the revised predictions of Manktelow and Over (1992), they need not be in contradiction with some other decision theoretic approach. The findings of this paper, including the findings on deontic logic, may be made compatible with a decision theoretic approach, by assuming that the four deontic connectives and the different foci change the pay-off matrices in a suitable way. I fully subscribe to the claim that utilities need to be integrated into a full account of the WST. Although a utility-based interpretation of our results may be possible and reasonable, the present results of the experiments (like those of Manktelow & Over, 1990) do not provide any independent evidence for such an interpretation. Even if the decision theoretic approach of the WST were to be modified based on the current data, the explanatory work would be done by a deontic logic and by a focus mechanism, and not by decision theory itself. Hence, any such interpretation – without further evidence – would only be a "just so" extension of a decision theoretic account of the WST.

13.3 Summary of the Discussion of Part III

The experiments provide support for central predictions of FDL theory. Experiment 8 and Experiment 9 showed differential effects of frequency information for WSTs with descriptive and prescriptive rules. In Experiment 10 we obtained support for the four postulated types of deontic conditionals based on deontic logic. In Experiment 11 the predicted interaction of focus effects (cheater detection versus cooperator detection) with different deontic rules was corroborated. In Experiment 12 this was replicated and the novel prediction of double focus effects with resulting *p* & *q* & *non-q* selection patterns has been confirmed. FDL theory was once more described as combining aspects of the domain-specific and the domain-general approaches of the WST.

The results have been discussed separately for the variety of current theories of the WST. Some theories have been shown to be inconsistent with the experimental evidence provided here. All theories have not predicted the obtained results which were mostly consistent with FDL theory. Hence, no theory apart from FDL theory can account for our findings. However, a few theories are formulated openly enough that they may perhaps be extended to account for the results. Future research has to investigate whether and in how far these theories can integrate the results of FDL theory or in how far FDL theory may integrate additional evidence from these theories.

In any case, currently only the proposal of FDL theory provides a coherent and rational explanation for the provided experimental results.

Part IV Towards a Domain-Specific but Rational Theory of Testing Rules

In this work, a domain-specific but rational account of testing descriptive and prescriptive rules in the WST has been theoretically elaborated and empirically corroborated.

The WST has been the Pandora’s box of the psychology of reasoning. Not only has the WST “become the single most investigated problem in the psychology of reasoning“ (Evans & Over, 1996, 356), it “has raised more doubts over human rationality than any other psychological task“ (Oaksford & Chater, 1998, 174). Research on the WST revealed two main anomalies for the falsificationist logicist norm of hypothesis testing: firstly, the ‘false’ confirmatory *p* & *q* selections in most descriptive tasks, and, secondly, the content effects mainly found for deontic rules. These problems of the WST have become notorious. They contributed to the development of several major theoretical proposals (e.g., mental model theory, social contract theory). Much of the research on the WST was concerned with working out peoples’ limitations in solving the WST logically. Here it is advocated that humans – despite their obvious computational limitations – are much more rational, systematic and flexible in testing rules than previously assumed in the WST debate. The main limitation may not have been with the task-solvers, but with us, the experimenters, in, for instance, not differentiating between descriptive and prescriptive tasks or in not understanding the use of additional quantitative and qualitative prior knowledge for optimal Bayesian information selection.

Before making some final remarks on rationality, knowledge, and domain-specificity, I would like to give a summary of the main parts of this work.

14 Summary

In Part I the WST and the main domain-general and domain-specific theories concerning the WST were introduced. Parts II and Part III made up the main parts of the thesis, the former being concerned with the testing of descriptive rules, the latter with the checking of prescriptive rules. The theoretical and empirical results and their relations to other theories of the WST have been summarised and discussed in depth in the corresponding general discussions (pp. 167 f., 261 f.). To avoid any battology here only a general summary of the results is given.

14.1 Part II – The Flexible Bayesian Logic of Testing Descriptive Rules

In Part II, a flexible Bayesian logic for testing descriptive rules has been elaborated. In Chapter 3 the advocated knowledge-based Bayesian account was developed and justified philosophically. First of all, it was argued that Hume's fundamental problem of induction cannot be solved by falsificationism, but only, perhaps, by a knowledge-based or synductive approach. Secondly, a knowledge-based Bayesian approach was advocated as a solution to Hempel's more specific problem of induction. These two arguments provide the philosophical basis for advocating a knowledge-based Bayesian account, which has here also been advocated in the context of the WST debate.

In the WST context the advocated Bayesian account is based on the refined modelling approach of Oaksford and Chater (1994, 1998a) (Chapter 4). Generally, a Bayesian account makes use of additional quantitative knowledge going beyond a mere distinction of logical categories. Falsificationism makes no use of this additional quantitative information and it would demand that only potentially falsifying cases should be investigated. In contrast, the Bayesian approach would consider additional frequency information.

For instance, testing the hypothesis "if I eat haddock then I drink gin", we may have reason to assume that P(eat haddock) = .10 and P(drink gin) = .15.

Falsificationism would demand that only 'haddock' cases (p) and 'not gin' cases (*non-q*) should always be selected to test the hypothesis, independent of any frequency information.

In contrast, from a Bayesian perspective, the selection of 'haddock' cases (p) and 'gin' cases (q) are rational (under the above condition). The mathematics of the Bayesian approach is complicated, but the basic idea is simple. If $P(p)$ and $P(q)$ are low, it is not only informative to check p but also the 'gin' case, q: If the hypothesis were false (and 'haddock', p, and 'gin', q, occurred only independently together), it would be improbable to find a 'haddock' case (q), if 'gin' (p) has been drunken. In contrast, if the hypothesis were true, 'gin' cases (q) would more often occur together with 'haddock' cases (p): $P((p \mid q) \mid M_D) > P((p \mid q) \mid M_I$. Hence, this low probability condition also the q case is – probabilistically – informative as well.

In the case of a conditional with high probabilities $P(p)$ and $P(q)$, like in the example 'if one eats rice then one drinks tea in Asia', the tea case, q, is less informative. Independently, of whether the hypothesis is true or not true, a tea case should mostly occur together with a rice case anyway (P$((p \mid q) \mid M_D) \approx P((p \mid q) \mid M_I)$.

Whereas a Bayesian approach makes use of this information, falsificationism leads to suboptimal results, since it throws away this additional source of quantitative evidence.

Additionally, it was argued here that it is suboptimal to neglect structural information. Unlike the universal model approach of Oaksford and Chater, here a knowledge-based Bayesian account was advocated (Chapter 5; cf. Sydow, 2002). This approach assumes that there is no *universal* Bayesian model, whose specific constancy assumptions are given *a priori* for testing any conditional, but that these constancy assumptions depend on our knowledge or the situation we are in. This approach differs from models postulating a universal approach for testing of conditionals or for their meaning (Oaksford & Chater, 1994, 1998, 2002, 2004; Oaksford & Wakefield, 2003; Hattori, 2003; cf. Evans, Handley & Over, 2003; Over & Evans, 2003; Over, 2004; Evans, Over, Handley 2005; and, ex negativo, Oberauer et al., 1999, 2004). Whereas previous direct tests of the predictions of the Bayesian approach were shown to have led either to negative or inconclusive results (e.g. Oaksford et al., 1999, Oberauer et al. 2004), I here achieved clearly positive results in the Experiments 1a, b, and 2a, b, by explicitly introducing all preconditions of the tested model. Unlike claims made by Oaksford and Wakefield (2003), this was achieved in non-successive tasks. Experiment 1 used ravens material (linking to the Raven's Paradox debate) and Experiment 2 used the original letter-number material (Wason, 1966). The confirmation of the (refined knowledge-based) Bayesian approach suggests the solution of the first anomaly of the WST debate, the dominant *p* and *q* selections were only dominant under (standard) conditions in which the mentioned atomic propositions can be assumed to have a low probability.

In Chapter 6, the knowledge-based approach was additionally tested more directly by inducing different structural Bayesian models, all of which tested the truth or falsity of a conditional. This was tested for the first time *within* an experiment (Experiment 3). The results were not completely positive but strongly confirmed different selections for different models. Moreover, it was shown that the participants were sensitive to additional implications of these different models.

In Chapter 7, the Bayesian account was generalised for the first time to other logical connectors. A 'Bayesian logic of testing hypotheses in the WST' was proposed theoretically and tested with regard to four 'Bayesian connectors'. Although the Experiments 4 to 7 also corroborated a logical explanatory factor, they provided a first evidence in support of the postulated Bayesian logic.

In Chapter 8, the General Discussion of Part II, it was reasoned that the results favour the postulated flexible Bayesian logic over all other theories of the WST. It was worked out in detail that neither mental model theory, nor mental logic theory, nor the domain-specific theories, nor relevance theory, nor matching bias theory can account for the pattern of findings obtained in Part II.

All in all, Part II developed a knowledge-based Bayesian logic of testing descriptive rules in the WST and provided substantial support for this novel position.

14.2 Part III – The Flexible Deontic Logic of Testing Prescriptive Rules

Part III proposed and elaborated a flexible deontic logic theory (FDL theory) (cf. also von Sydow, Hagmayer, Metzner, Waldmann; 2005; von Sydow, submitted).

In Chapter 9, FDL theory was presented as a synthesis of converging lines of research, which may explain the second anomaly of the WST. The clear-cut logical or illogical selection patterns found particularly in deontic WSTs were claimed to be explicable in a systematic way by FDL theory.

This theory on the one hand combines a deontic logic of different ought tables for four different deontic kinds of conditional (cf. Beller, 2003; Johnson-Laird & Byrne, 2005) and on the other hand a goal-based focus on particular cases (cf. Sperber & Girotto, 2002, 2003) of the postulated ought tables. This approach claims to explain cheater detection (Cosmides, 1989; Cosmides & Tooby, 1992; Gigerenzer & Hug, 1992; cf. Fiddick, 2004) and pragmatic reasoning schemas (Cheng & Holyoak, 1985; Holyoak & Cheng, 1995) on the rational basis of deontic logic, but in a way, which differs from traditional mental model theory (Johnson-Laird & Laird, 1991, 1992, 1995, 2002; cf. Bucciarelli & Johnson-Laird, 2005). (The relation to other theories was considered both in Chapter 9 and in Chapter 13.)

Moreover, it is postulated that the standard way of checking prescriptive rules needs to be differentiated from the standard way of assessing the truth or falsity of a hypothesis. In checking prescriptive rules, a police context, for instance, should lead to an *a priori* cheater detection focus; one tries to find the rule violation cases in the corresponding deontic ought table. This differs from our Bayesian account of testing the truth or falsity of descriptive rules, according to which the selected cells should result *a posteriori*, depending on which cell becomes most informative given the known probabilities. Hence, one should find standard frequency effects in the testing of descriptive rules but not in prescriptive rules.

In Chapter 10, this first aspect of FDL theory was tested. The postulated dichotomy for frequency effects was tested for the first time in two experiments. Although the results of Experiment 8a/b were a bit more intricate than a simple dichotomy suggests, the results were in line with distinguishing (at least) the two postulated kinds of test strategies. In Experiment 9a/b the intended dichotomy was tested under improved ideal conditions. This evidence supported the proposed difference of testing descriptive and prescriptive rules.

In Chapter 11 all four kinds of deontic conditionals (all with only one forbidden cell), derived from deontic logic, were tested using a cheater detection goal. The four predicted kinds of different selection patterns were corroborated in Experiment 10.

In Chapter 12, the postulated *interaction* of deontic goals and deontic logic was investigated. The essential role of goals, postulated by FDL theory, was examined in two experiments. Experiment 11 was concerned with the interaction of obligation and prohibition rules and the goals of cheater or cooperator detection. Experiment 12 additionally contrasted this interaction with double focus effects, simultaneously checking for cooperators and cheaters. The results favour FDL theory over the alternative theories of the WST. For instance, social contract theory (Cosmides, 1989; Cosmides & Tooby, 1992; cf. Fiddick, 2004), the flagship of evolutionary psychology, has exclusively predicted a specific and unsystematic cheater detection module. The results of this work show that cheater detection appears to be part of a much more systematic capacity of deontic logic and goal-dependent focusing on particular cases of ought tables. Also mental model theory cannot account for the findings using the models it has postulated, without itself introducing a focus mechanism. However, if a focus mechanism were introduced, it would become highly questionable which phenomena would remain explicable by the original mechanisms postulated by mental model theory itself.

In Chapter 13, it is discussed in detail that – despite debts to other theories – the complete pattern of findings cannot be accounted for by any other current theory of the WST; only the synthesis of FDL theory predicts most of the findings.

Overall, the obtained results of Part II and Part III corroborate the two proposals tested in this work, the flexible Bayesian logic of testing descriptive rules and the flexible deontic logic of testing prescriptive rules. The accounts provide reasonable resolutions for the two anomalies of a falsificationist understanding of the WST, the confirmatory *p* and *q* answers in most descriptive WSTs and the clear-cut ‘logical’ *p* and *non-q* patterns or illogical *non-p* and *q* patterns in deontic WSTs. Additionally, both proposals also lead to the confirmation of novel predictions.

It was discussed in detail in Part II as well as in Part III that all other main theories of the WST, that is mental model theory, mental logic theory, pragmatic reasoning schema theory, social contract theory, relevance theory, and matching bias theory cannot account for the findings in the corresponding parts. Hence, there is no need to show that each of these theories even more clearly cannot account for the *combined* results of the two main parts of this work.

15 Concluding Remarks – A Domain-Specific but Normative Approach

The advocated approach for the testing of descriptive and prescriptive rules is neither a rational domain-general theory (like mental logic theory) nor an 'irrational' domain-specific theory (like social contract theory). Instead, the postulated flexible Bayesian logic and the postulated flexible deontic logic may both be seen as *rational domain-specific theories*. On the one hand, they both refer to established rational systems of reasoning or hypothesis testing; on the other hand, both are only applicable to the domain of testing either descriptive or prescriptive rules.

15.1 The Rationality of the Use of Additional Knowledge

Since the knowledge-based Bayesian approach has here been proposed as a normative approach, one may ask what it means to be normative. How could the knowledge-based Bayesian approach be said to be more rational than a universal Bayesian approach? Why should we prefer a normative Bayesian approach to the also 'normative' falsificationist approach?

In Chapter 3, it was argued that only a Bayesian approach that is knowledge-based appears to solve Hume's and Hempel's philosophical problems of induction. However, even without referring to these philosophical argumentations, it is obvious, why a reference to additional quantitative and qualitative knowledge is rational. The more complete usage of additional information, understood as additional premises, provides additional constraints, which allow us to derive more adequate conclusions than when premises are neglected.

Let us consider the analogous case of a mathematical proof. The Pythagorean Theorem cannot be proved if one leaves out the premises that the triangle in question is right-angled or that a Euclidean metric is to be used. The relation of the three sides of the triangle described by the Pythagorean Theorem is false (or inapplicable) if for instance a non-Euclidean Riemannian geometry is used. Likewise, if we were actually concerned with a Euclideian metric and a right-angled triangle, but did not use this information for our calculations, we could not derive the Pythagorean Theorem and this would clearly be suboptimal.

Analogously, the knowledge-based Bayesian account goes beyond falsificationist logicism, since it makes additional use of knowledge about quantitative or qualitative information. However, the Bayesian approach advocated here is not opposed to logic and the postulated Bayesian logic of hypothesis testing even refers to the connectors distinguished by propositional logic. However, the knowledge-based Bayesian approach uses more information than propositional logic and, interestingly, its implications partly lead to opposed normative predictions for the WST. Moreover, the advocated knowledge-based account in a similar way goes beyond universal Bayesian account, in considering additional knowledge about structural constraints.

Philosophically such a knowledge-based approach to rationality on the one hand still aims at obtaining a normative basis of rationality, in this sense it still adheres to the research programme of a *mathesis universalis* (Leibniz), but on the other hand this approach seems to be opposed to domain-general accounts that regard it as rational to ignore domain-specific knowledge in the process of hypothesis testing. Aspects of the latter view have been highly influential in different currents of modern philosophy, not only in falsificationism, but also for instance in naïve inductionism (cf. Chapter 3). Moreover, within psychology the advocated knowledge-based account of rationality may not only be relevant to the WST debate but also for the more general rationality debate (e.g., Kahneman & Tversky, 1996; Gigerenzer, 1996; Gigerenzer & Goldstein, 1996; Gigerenzer, Todd, & The ABC Group, 1999; Gigerenzer, 2000; Hertwig & Hoffrage, 2001; Sloman, Over, Slovak, & Stibel, 2003; Chater, Oaksford, Nakisa, & Redington, 2003; Over, 2004). There is no antagonism between rationality and domain-specificity in hypothesis testing, only knowledge-based models of induction are rational.

15.2 Between Domain-Generality and Domain-Specificity

The knowledge-based understanding of rational hypothesis testing in the WST also makes apparent that the question of whether human reasoning is (and ought to be) either domain-general or domain-specific is ill posed.

There may be not two but several levels of generality and specificity. The use of additional knowledge provides additional constraining premises, and the derived conclusions are only valid if these premises are valid – in this sense, they can be called domain-specific on quite different levels. Even *within* the necessarily incomplete approach advocated here, one may distinguish different levels of generality and specificity.

On a first level, the postulated flexible Bayesian logic of testing descriptive rules and the flexible deontic logic of prescriptive rules have a logical core, since both refer to aspects of logic. Both distinguish logical classes, with regard to the tested or checked hypothesis. Logical classes may be understood to be domain-general.

On a second level, the two sub-theories, either concerned with the testing of descriptive or prescriptive rules, are domain-specific or 'realm-specific' theories. By definition, deontic logic is only concerned with prescripts. The advocated goal-dependent cheater or cooperator focus, at least normally, refers to the checking of prescripts. Likewise, the basic models of information gain only refer to the testing of (descriptive) hypotheses, since they distinguish a dependence model for the truth of a hypothesis and at least one alternative model for its falsity. The distinction of two hypotheses normally does not make sense for prescripts; the validity of them is neither falsifiable nor in question. The two strategies of either testing the truth or checking a particular cell of an ought table seem to be goal-dependent. Additionally, there may be schemas that trigger the one or the other test strategy. Despite this goal dependence, there are rational and irrational strategies (given a particular goal). Moreover, although aspects of the two strategies are 'domain-specific', they are much more general and systematic than anything postulated by the previous domain-specific theories of the WST (PRS theory and SC theory). Hence, the testing descriptive hypotheses or checking prescripts may be more adequately called 'realm-specific'.

There is a further level of specificity. In testing descriptive rules, different Bayesian models have to be distinguished, based on specific constraints given in situations or classes of situations. In addition, in a particular deontic situation the plausibility of *specific* objectives may be regarded to be a feature of that situation.

Hence, instead of a dichotomy of domain-specific or domain-general mechanisms the advocated perspective seems to suggest a – perhaps schema-based – hierarchy of different levels of domain-specificity.

However, such a hierarchy of generality and specificity raises the fundamental question of whether the postulated rational systems of testing rules may be enriched by additional further kinds of domain-specific knowledge. Only further research will show. But the advocated synductive account of knowledge acquisition (Chapter 3) suggests that other kinds of structural knowledge may plausibly have a rational influence (in the above sense) on WSTs as well, e.g., the temporal or causal order in descriptive WSTs, or the behavioural options and constraints in prescriptive WSTs.

In any case, the flexible Bayesian logic theory of testing hypotheses and the flexible deontic logic theory of testing prescripts seem to explain the two main anomalies of the falsificationist research programme on the WST, by introducing rational accounts of testing rules, which refer to established mathematical theories. Since the WST has been obsessed with finding irrational selection patterns, it was of primary interest here to work out that the seemingly chaotic patterns are at least partly based on an underlying systematic and rational basis. Even if future research were to adopt and corroborate the perspective advanced here, new fields of future research may arise. Despite many positive results, one must again also explain the few deviations found. For instance, subjects did not reason according to the 'complicated' Laming model. The current research only briefly pursued the question why these deviations arise, since the main goal has been to establish the fundamental systematics postulated. However, perhaps there may be partial representations of Bayesian or causal models, which might lead to something like a 'mental Bayesian model theory' or 'mental causal model theory' (cf. the General Discussions of Chapter 6 and 7).

The presented research will not close the Pandora's box of the WST – more research is needed and novel problems may well occur –, but the theoretical proposals (flexible Bayesian logic and flexible deontic logic) and the confirmatory evidence of the current work seem to contribute to a more differentiated and systematic understanding of a domain-specific but comparatively rational testing of descriptive and prescriptive rules.

16 Acknowledgements

Although the author is wholly responsible for the entire work – including the errors – I would like to thank several persons without whom the work would not have been finished in its present form.

Above all, I would like to express my gratefulness to a collective entity, the research group of Professor Dr. Michael Waldmann at the University of Göttingen. This group provided an inspiring background for my research. The excellent group climate enabled me to pose questions and develop my own ideas.

Individually, I first would like to thank my supervisor Michael Waldmann for support, and for encouraging me to develop my own position in this field. Nonetheless, my results partly gravitated back to his causal model theory. Additionally, I particularly want to thank Dr. York Hagmayer for help, valuable discussions, and a friendly atmosphere. I am also grateful to my colleague and friend Björn Meder for his cooperative attitude, which made the work on the DFG project and on the thesis enjoyable.

For corrections and helpful comments on the dissertation itself, I would like to thank Michael Waldmann, Sandra Többe, Paul Embleton, and Friedrich von Sydow. I also thank Professor Dr. Margarete Boos for being second examiner on this work and her interest in it.

More generally, I am grateful to my father Friedrich von Sydow for sparking my interest in questions of logic and rationality, and to my mother Anne von Sydow for inspiring a critical stance towards a merely logical point of view. Particularly, I have to thank Sandra Többe who had to bear a boy friend who too often spent his weekends in front of a monitor. Moreover, I want to thank her for her love and emotional support.

I also would like to thank Professor Dr. Edgar Erdfelder, who supervised my MA thesis. He is responsible for infecting me with the ‘Wason-selection-task’ virus. Furthermore, I want to thank Sieghard Beller, Nick Chater, Mike Oaksford, Magda Osman, and Gerd Gigerenzer for interesting discussions of some of my proposals concerning the WST. Last but not least, I thank Nicole Metzner, Nora Thiemann, Maximiliane Janski and Alexander Wiegmann for collecting some of the data.

17 List of Abbreviations

C_i	cell *i* of a contingency matrix
df	degrees of freedom
EIG	expected information gain
FBL	flexible Bayes logic
FDL	flexible deontic logic
IG	information gain
$\mathrm{M_D}$	dependence model
$\mathrm{M_I}$	independence model
ML	mental logic
MM	mental model
MST	many cards selections task
N	number of participants in an experiment
n	number of participants in particular conditions of an experiment
PRS	pragmatic reasoning schema
r_φ	Phi correlation (normally φ or Φ).
SC	social contract
SEIG	scaled expected information gain
WST	Wason selection task

18 References

Ahn, W., & Graham, L. M. (1999). The impact of necessity and sufficiency in the Wason four-card selection task. *Psychological Science, 10*, May, 237-241.

Alexander, H. G. (1958). The paradoxes of confirmation. *British Journal of Philosophy of Science, 9*, 227-233.

Almor, A. (2003). Specialized behaviour without specialized modules. In D. E. Over (Ed.), *Evolution and the psychology of thinking: The debate* (pp. 101-119). Hove, UK: Psychology Press.

Almor, A., & Sloman, S. A. (1996). Is deontic reasoning special? *Psychological Review*, *103*, 374-380.

Almor, A., & Sloman, S. A. (2000). Reasoning versus text processing in the Wason selection task: A nondeontic perspective on perspective effects. *Memory & Cognition, 28*, 1060-1070.

Anderson, J. R., & Sheu, C. F. (1995). Causal inferences as perceptual judgments. *Memory and Cognition, 23*, 510-524.

Aristoteles (1990/about 350 BC). *Lehre vom Beweis oder Zweite Analytik* (*Organon IV;* transl. by E. Rolfes, introduction by O. Höffe).Hamburg: Meiner.

Aristoteles (1992/about 345 BC). *Topik* (*Organon V;* transl. by E. Rolfes, introduction by H. G. Zekl). Hamburg: Meiner.

Axelrod, R., & Hamilton, W. D. (1981). The evolution of cooperation. *Science, 211*, 1390-1391.

Barres, P. E., & Johnson-Laird, P. N. (2003). On imagining what is true (and what is false). *Thinking and Reasoning, 9*, 1-42.

Beller, S. (1997). *Inhaltseffekte beim logischen Denken - Der Fall der Wason'schen Wahlaufgabe. Eine wissensbasierte Lösung für ein altes Problem*. Psychologia Universalis. Lengerich, Germany: Pabst.

Beller, S. (2001). A model theory of deontic reasoning about social norms. In J. D. Moore & K. Stenning (Eds.), *Proceedings of the Twenty-Third Annual Conference of the Cognitive Science Society* (pp. 63-68). Mahwah, NJ: Erlbaum.

Beller, S. (2003). The flexible use of deontic mental models. In R. Alterman & D. Kirsh (Eds.), *Proceedings of the Twenty-Fifth Annual Conference of the Cognitive Science Society* (pp. 127-132). Mahwah, NJ: Erlbaum.

Beller, S., & Bender, A. (2004). Cultural differences in the cognition and emotion of conditional promises and threats – comparing Germany and Tonga. In K. Forbus, D. Gentner, & T. Regier (Eds.), *Proceedings of the Twenty-Sixth Annual Conference of the Cognitive Science Society* (pp. 1411-1416). Mahwah, NJ: Erlbaum.

Beller, S., & Spada, H. (2003). The logic of content effects in propositional reasoning: The case of conditional reasoning with a point of view. *Thinking and Reasoning, 9,* 335-378.

Bender, A., & Beller, S. (2003). Polynesian tapu in the 'deontic square': a cognitive concept, its linguistic expression and cultural context. In R. Alterman & D. Kirsh (Eds.), *Proceedings of the Twenty-Fifth Annual Conference of the Cognitive Science Society* (pp. 133-138). Mahwah, NJ: Lauwrence Erlbaum.

Beller, S., Bender, A., & Kuhnmünch, G. (2005). Understanding conditional promises and threats. *Thinking & Reasoning, 11,* 209-238.

Braine, Martin D. S. (1978). On the relation between the natural logic of reasoning and standard logic. *Psychological Review, 85*, 1-21.

Bransford, J. D., Barcley, J. R. & Franks, J. J. (1972). Sentence memory: A constructive versus interpretative approach. *Cognitive Psychology, 3*, 193-209.

Broad, C. D. (1918/1920, 1968). *The relation between induction and probability, I, II.* Reprinted in a collection of his papers: *Induction, probability, and causation.* Dordrecht (NL): D. Reidel.

Bochenski, I. M. (1993). *Die zeitgenössischen Denkmethoden.* 10. Auflage. Tübingen, Basel: Francke Verlag.

Bucciarelli, M., & Johnson-Laird, P. N. (2005). Naïve deontics: A theory of meaning, representation, and reasoning. *Cognitive Psychology, 50*, 159-193.

Campbell, D. T (1974) Evolutionary Epistemology. In P. A. Schilpp (Ed.). *The Philosophy of Karl R. Popper* (pp. 413-463). La Salle, Ill.: Open Court.

Chalmers, A. F. (2001). *Wege der Wissenschaft. Einführung in die Wissenschaftstheorie* (translated from English by N. Bergemann & Ch. Altstötter-Gleich). Heidelberg: Springer.

Chater, N., & Oaksford, M. (1993). Logicism, mental models and everyday reasoning: Reply to Garnham. *Mind & Language, 8*, 72-89.

Chater, N., & Oaksford, M. (1996). Deontic reasoning, modules and innateness: A second look. *Mind & Language, 11,* 191-202.

Chater, N., & Oaksford, M. (1999a). Information gain and decision-theoretic approaches to data selection. Response to Klauer (1999). *Psychological Review, 106*, 223-227.

Chater, N., & Oaksford, M. (1999b). The probability heuristics model of syllogistic reasoning. *Cognitive Psychology, 38*, 191-285.

Chater, N., & Oaksford, M. (2000). The rational analysis of mind and behavior. *Synthese, 122*, 93-131.

Chater, N., Oaksford, M., Nakisa, R., & Redington, M. (2003). Fast frugal and rational: How rational norms explain behaviour. *Organizational Behavior and Human Decision Processes, 90,* 63-86.

Cheng, P. W. (1997). From covariation to causation: A causal power theory. *Psychological Review, 104*, 367-405.

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

Cheng, P. W., & Holyoak, K. J. (1989). On the natural selection of reasoning theories. *Cognition*, *33*, 285-314.

Cheng, P. W., Holyoak, K. J., Nisbett, R. E., & Lindsay, M. O. (1986). Pragmatic versus syntactic approaches to training deductive reasoning. *Cognitive Psychology, 18*, 293-328.

Cosmides, L. (1989). The logic of social exchange: Has natural selection shaped how humans reason? *Cognition, 31*, 187-276.

Cosmides, L., & Tooby, J. (1992). *Cognitive adaptations for social exchange*. In J. H. Barkow, L. Cosmides, & J. Tooby (Eds.), *The adapted mind* (pp. 163-228). New York: Oxford University Press.

Cosmides, L., & Tooby, J. (1994). Better than rational: Evolutionary psychology and the invisible hand. *American Economic Review, May*, 327-332.

Cox, J. R., & Griggs, R. A. (1982). The effect of experience on performance in Wason's selection task. *Memory and Cognition, 10*, 496-502.

Crott, H. W., Giesel, M., & Hoffmann, C. (1998). The process of inductive inference in groups: The use of positive and negative hypothesis and target testing in sequential rule-discovery tasks. *Journal of Personality and Social Psychology*, *75*, 938-952.

Cummins, D. D. (2000). How the social environment shaped the evolution of mind. *Synthese*, *122*, 3-28.

Dawkins, R. (1983/1982). *The extended phenotype. The gene as the unit of selection.* Oxford: Oxford University Press (first publ.: W. H. Freeman & Co. Ltd.).

Dennett, D. C. (1995). *Darwin's Dangerous Idea.* London: Penguin.

Dixit, A., & Skeath, S. (1999). *Games of strategy*. New York: W. W. Norton & Company.

Dove, A. (1996). *Perspektiveneffekte in der Selektionsaufgabe aus der Sicht der Theorie der Sozialen Verträge.* Diplomarbeit in Psychologie (supervisor: E. Erdfelder). Bonn: Rheinische Friedrich-Wilhelms-Universität.

van Duyne, P. C. (1974). Realism and linguistic complexity in reasoning. *British Journal of Psychology, 65,* 59-67.

Earman, J. (1992). *Bayes or bust?* Cambridge (MA): MIT.

Edgington, D. (2003). What if? Questions about conditionals. *Mind & Language, 18*, 380-401.

Erdfelder, E. & Bredenkamp. J. (1994). Hypothesenprüfung. In T. Herrmann & W. H. Tack (Eds.), *Enzyklopädie der Psychologie, Methodologische Grundlagen der Psychologie, Band B/1/1/1.* (pp. 604-648). Göttingen: Hogrefe Verlage.

Erdfelder, E., & Dove, A. (1997). Effekte der natürlichen sozialen Perspektive bei der Bearbeitung der Selektionsaufgabe nach Wason. In E. van der Meer et al. (Eds.), *Experimentelle Psychologie. Abstracts der 39. Tagung experimentell arbeitender Psychologen* (pp. 185-186). Lengerich: Pabst.

Erdmann, B. (1892). *Logik. Erster Band: Logische Elementarlehre.* Halle a. S.: Max Niemeyer.

Evans, J. S. B. T. (1972). Interpretation and matching bias in a reasoning task. *British Journal of Psychology, 24,* 193-199.

Evans, J. S. B. T. (1989). *Bias in human reasoning: Causes and consequences.* Hove, UK: Lawrence Erlbaum.

Evans, J. S. B. T. (1996). Deciding before you think: Relevance and reasoning in the selection task. *British Journal of Psychology, 87*, 223-240.

Evans, J. S. B. T. (2002). Matching bias and set sizes: A discussion of Yama (2002*).* *Thinking and Reasoning, 8*, 153-163.

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

Evans, J. S. B. T., Over, D. E., & Handley, S. J. (2005). Suppositions, Extensionality, and Conditionals: A Critique of the Mental Model Theory of Johnson-Laird and Byrne (2002). *Psychological Review*, *112*, 1040–1052.

Evans, J. S. B. T., Legrenzi P., & Girotto, V. (1999). The Influence of Linguistic Form on Reasoning: The Case of Matching Bias. *The Quarterly Journal of Experimental Psychology, 52A*, 185-216.

Evans, J. S. B. T., & Lynch, J. S. (1973). Matching bias in the selection task. *British Journal of Psychology, 64*, 391-397.

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

Evans, J. S. B. T., Over, D. E., & Manktelow, K. I. (1993). Reasoning, decision making and rationality. *Cognition, 49,* 165-187.

Eysenck, M. W., & Keane, M. T. (1995). *Cognitive Psychology.* Hove (UK): Psychology Press.

Fairley, N., Manktelow, K. I., & Over, D.E. (1999). Necessity, sufficiency, and perspective effects in causal conditional reasoning. *Quarterly Journal of Experimental Psychology, 52A*, 771-790.

Feeney, A., & Handley, S. (2000). The suppression of q card selections: Evidence for deductive inference in Wason's selection task. *Quarterly Journal of Experimental Psychology, 53A*, 1224-1242.

Fehr, E., & Gächter, S. (2002). Altruistic punishment in humans. *Nature, 415,* 137-140.

Fehr, E., & Fischbacher, U. (2003). The nature of human altruism. *Nature, 425,* 785-791.

Fehr, E., & Fischbacher, U. (2004). Social norms and human cooperation. *Trends in Cognitive Science, 8*, 185-190.

Fiedler, K., & Hertel, G (1994). Content related schemata versus verbal-framing effects in deductive reasoning. *Social Cognition, 2,* 129-147.

Fiddick, L. (2003). Is there a faculty of deontic reasoning? A critical re-evaluation of abstract deontic versions of the Wason selection task. In D. E. Over (Ed.), *Evolution and the psychology of thinking: The debate* (pp. 33-60). Hove, UK: Psychology Press.

Fiddick, L. (2004). Domains of deontic reasoning. *Quarterly Journal of Experimental Psychology, 57*, 447-474.

Fiddick, L., Cosmides, L., & Tooby, J. (2000). No interpretation without representation: The role of domain-specific representations and inferences in the Wason selection task. *Cognition, 77*, 1-79.

Fiedler, K., & Hertel, G (1994). Content related schemata versus verbal-framing effects in deductive reasoning. *Social Cognition, 2*, 129-147.

Fodor, J. A. (1983). *The modularity of mind.* Cambridge, MA: MIT Press.

Forster, M. R. (1994). Non-Bayesian foundations for statistical estimation, prediction, and the ravens example. *Erkenntnis, 40*, 357-376.

Frege, G. (1998). *Begriffsschrift und andere Aufsätze*. (Including: *Begriffsschrift, eine der arithmetischen nachgebildete Formelsprache des reinen Denkens, 1879*) Halle: Verlag Louis Nebert.

Garnham, A. (1993). Is logicist cognitive science possible? *Mind and Language, 8*, 49-71.

Gebauer, G., & Laming, D. (1997). Rational choices in Wason's selection task. *Psychological Research, 60*, 284-293.

Gentzen, G. (1935). Untersuchungen über das logische Schliessen. *Mathematische Zeitschrift, 39,* 176-221.

Gigerenzer, G. (1996). On narrow norms and vague heuristics: A reply to Kahneman and Tversky (1996). *Psychological Review, 103, 3*, 592-596.

Gigerenzer, G., & Goldstein, D. G. (1996). Reasoning the fast and frugal way: Models of bounded rationality. *Psychological Review, 103,* 650-669.

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

Gigerenzer, G., Todd, P. M., & the ABC Research Group (1999). *Simple heuristics that make us smart*. New York: Oxford University Press.

Girgenzer, G. & Selten, R. (2001). *Bounded Rationality*. Cambridge, MA: MIT Press.

Gigerenzer, G. (2000). Adaptive Thinking. New York: Oxford University Press.

Goodman, N. (1955/1963/1973). *Fact, fiction and forecast*. Indianapolis (US): Bobbs Merrill.

Gould, S. J., & Lewontin, R. C. (1979). The spandrels of San Marco and the Panglossian paradigm: A critique of the adaptationist programme. *Proceedings of the Royal Society of London, B205,* 581-598.

Green, D. W. (1995). Externalization, counter-examples, and the abstract selection task. *Quarterly Journal of Experimental Psychology, 48 A,* 424-446.

Green, David W. (2000). Review of 'Oaksford & Chater 1998. Rationality in an uncertain world'. *The Quarterly Journal of Experimental Psychology, 53A*, 281-283.

Green, D. W., & Over D. E. (1997). Causal inference, contingency tables and the selection task. *Current Psychology of Cognition, Cahiers de Psychologie Cognitives*, *16*, 459-487.

Green, D. W., & Over, D. E. (1998). Reaching a decision: A reply to Oaksford. *Thinking and Reasoning, 4,* 231-248.

Green, D. W., Over, D. E., & Pyne, R. A. (1997). Probability and choice in the selection task. *Thinking and Reasoning, 3*, 209-236.

Grice, H. P. (1975). Logic and conversation. In P. Cole & J. L. Morgan (Eds.), *Syntax and semantics: Vol. III. Speech acts* (pp. 41-58). New York: Seminar Press.

Griggs, R. A., & Cox, J. R. (1982). The elusive thematic-materials effect in Wason's selection task. *British Journal of Psychology, 73*, 407-420.

Griggs, R. A., Platt, R. D., Newstead, S. E., & Jackson, S. L. (1998). Attentional Factors in a Disjunctive Reasoning Task. *Thinking and Reasoning*, *4*, 1-14.

Gummelt, A. (2005). *Cheater detection, cooperator detection und altruist detection bei präskriptiven reziproken Regeln*. Unveröffentlichte Diplomarbeit in Psychologie (Betreuer: M. v. Sydow) am Institut für Psychologie. Göttingen: Universität Göttingen.

Güth, W., Schmittberger, R., & Schwarze, B. (1982). An experimental analysis of ultimatum bargaining. *Journal of Economic Behavior and Organization, 3,* 367-388.

Haegeman, L. (2003). Conditional clauses: External and internal syntax. *Mind & Language, 18,* 317-339.

Hagmayer, Y., & Waldmann, M. R. (in press). Kausales Denken. In J. Funke (Ed.), *Enzyklopädie der Psychologie. Denken und Problemlösen: Band C/II/8*. Göttingen: Hogrefe Verlag.

Handley, S. J., Feeney, A, & Harper, C. (2002). Alternative antecedents, probabilities, and the suppression of fallacies in Wason's selection task. *The Quarterly Journal of Experimental Psychology, 55A*, 799-818.

Hardman, D. (1998). Does reasoning occur on the selection task? A comparison of relevance-based theories. *Thinking and Reasoning, 4,* 353-376.

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

Hajek, P. (2002). Fuzzy Logic. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy.* <http://plato.stanford.edu/archives/fall2002/entries/logic-fuzzy/>.

Hegel, G. W. F. (1994 / 1816). Wissenschaft der Logik. Die Lehre vom Begriff. Hamburg: Felix Meiner Verlag.

Hempel, C. G. (1945a, b). Studies in the logic of confirmation. *Mind, 54*, 1-25, 7-121.

Hempel, C. G. (1958). Empirical statements and falsifiability. *Philosophy, 33*, 342-348.

Hertwig, R., & Hoffrage, U. (2001). Eingeschränkte und ökologische Rationalität: Ein Forschungsprogramm. *Psychologische Rundschau, 52*, 11-19.

Hintikka, J., & Suppes, P. (Eds.). (1966). *Aspects of inductive logic.* North-Holland Publishing Company: Amsterdam.

Hilpinen, R. (Ed.). (1970). *Deontic logic: Introductory and systematic readings.* Dordrecht: Reidel.

Hilpinen, R. (Ed.). (1981). *New Studies in Deontic Logic. Norms, actions, and the foundation of ethics.* Dordrecht: Reidel.

Hiraishi, K., & Hasegawa, T. (2001). Sharing-rule and detection of free-riders in cooperative groups: Evolutionarily important deontic reasoning in the Wason selection task. *Thinking and Reasoning, 7,* 255-294.

Holland, J.H., Holyoak, K.J., Nisbett, R.E., & Thagard, P.T. (1986). *Induction: Processes of Inference, Learning, and Discovery*. Cambridge, MA: The MIT Press.

Holyoak, K. J., & Cheng, P. W. (1995a). Pragmatic reasoning about human voluntary action. In St. E. Newstead, & J. St. B. T. Evans (Eds.), *Perspectives on Thinking and Reasoning* (pp. 67-89). Hove, UK: Erlbaum.

Holyoak, K. J., & Cheng, P. W. (1995b). Pragmatic reasoning with a point of view. *Thinking and Reasoning, 1,* 289-313.

Holyoak, K. J, & Thagard, P. (1995). *Mental Leaps*. Cambridge, MA: MIT Press.

Hooker, C. A., & Stove, D. (1967). Relevance and the Ravens. *British Journal of the Philosophy of Science, 18,* 305-315.

Hosiasson-Lindenbaum, J. (1940). On Confirmation. *Journal of Symbolic Logic, 5*, 133-148.

Howson, C., & Urbach, P. (1993). Scientific Reasoning: The Bayesian Approach (2nd ed.; 1st ed.: 1989). Chicago: Open Court.

Hull, D. L., & Ruse, M. (Ed.). (1998). *The Philosophy of Biology.* Oxford, etc.: Oxford University Press.

Humberstone, I. L. (1994). Hempel meets Wason. *Erkenntnis, 41,* 391-402.

Hume, D. (1739/1888/1978). A Treatise of Human Nature (L. S. Selby-Bigge, Ed.; 2nd edition, P. N. Nidditch, Ed.). Oxford: Clarendon Press.

Husserl, E. (1900/1993). *Logische Untersuchungen. Prolegomena zur reinen Logik.* Tübingen: Max Niemeyer.

Hussy, W. (1984; 1986). *Denkpsychologie. Band 1 und 2*. Stuttgart: Kohlhammer.

Iwin, A. A. (1975). Grundlagen der Logik von Wertungen (Ed. by H. Wessel, transl. from Russian by K. Wuttich & W. Stelzner). Berlin: Akademie-Verlag.

Jackson, S. L., & Griggs, R. A. (1990). The elusive pragmatic reasoning schema effect. *Quarterly Journal of Experimental Psychology, 42A*, *353-373.*

Johnson-Laird, P. N., & Byrne, R. M. J. (1991). *Deduction.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Johnson-Laird, P. N., & Byrne, R. M. J. (1992). Modal reasoning, models, and Manktelow and Over. *Cognition, 43*, 173-182.

Johnson-Laird, P. N., & Byrne, R. M. J. (1995). A model point of view. *Thinking and Reasoning, 1*, 339-350.

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

Johnson-Laird, P. N., Legrenzi, P., Girotto, P., & Legrenzi, M. S. (2000). Illusions in reasoning about consistency. *Science, 288*, 531-532.

Johnson-Laird, P. N., Legrenzi, P., Girotto, V., Legrenzi, M. S., & Caverni, J.-P. (1999). Naive probability: A mental model theory of extensional reasoning. *Psychological Review*, *106*, 62-88.

Johnson-Laird, P. N., Legrenzi, P., & Legrenzi, M. (1972). Reasoning and a sense of reality. *British Journal of Psychology, 63,* 395-400.

Johnson-Laird, P. N., & Wason, P. C. (1970a). A theoretical analysis of insight into a reasoning task. *Cognitive Psychology, 1,* 134-148.

Johnson-Laird, P. N., & Wason, P. C. (1970b). Insight into a logical relation. *Quarterly Journal of Experimental Psychology, 22*, 49-66.

Jonas, E., Schulz-Hardt, St., & Frey, D. (2001). Konfirmatorische Informationssuche bei simultaner vs. sequentieller Informationssuche. *Zeitschrift für Experimentelle Psychologie, 48, 3,* 239-247.

Kahneman, D., & Tversky, A. (1996). On the Reality of Cognitive Illusions. *Psychological Review, 103,* 582-591.

Kant, I. (1990, 1781/1787). *Kritik der reinen Vernunft* (Ed.: R. Schmidt). Hamburg: Felix Meiner.

Kant, I. (2003/1787). *Kritik der praktischen Vernunft.* Hamburg: Felix Meiner.

Kao, S.-F., & Wassermann, E. A. (1993). Assessment of an information integration account of contingency judgment with examination of subjective cell importance. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 19*, 1363-1386.

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

Kirby, K. N. (1994b). False alarm: a reply to Over and Evans. *Cognition, 52*, 245-250.

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

Klayman, J., & Ha, Y.-W. (1987). Confirmation, disconfirmation and information in hypothesis testing. *Psychological Review, 94,* 211-228.

Kroger, J. K., Cheng, P. W., & Holyoak, K. J. (1993). Evoking the permission schema: The impact of explicit negation and a violation-checking context. *The Quarterly Journal of Experimental Psychology, 46A*, 615-635.

Krynski, T. R., & Tenenbaum, J. B. (2004). Causal structure in conditional reasoning. *Proceedings of the Twenty-Sixth Annual Conference of the Cognitive Science Society* (pp. 744-749). Mahwah, NJ: Erlbaum.

von Kutschera, F., & Breitkopf, A. (1992). *Einführung in die moderne Logik.* Freiburg: Karl Alber Verlag.

Laming, D. (1996). On the analysis of irrational data selection: A critique of Oaksford and Chater (1994). *Psychological Review, 103,* 364-373.

Leibniz, G. W. (2000 / 1683-1698). Grundlagen des logischen Kalküls. Hamburg: Felix Meiner Verlag.

Liberman, N., & Klar, Y. (1996). Hypothesis testing in Wason's selection task: Social exchange, cheating detection or task understanding. *Cognition, 58*, 127-156.

Lien, Y., & Cheng, P. W. (2000). Distinguishing genuine from spurious causes: A coherence hypothesis. *Cognitive Psychology, 40*, 87-137.

Lloyd, E. A. (1999). Evolutionary psychology: The burdens of proof. *Biology and Philosophy, 14*, 211-233.

Love, R. E., & Kessler, C. M. (1995). Focusing in Wason's selection task: Content and instruction effects. *Thinking and Reasoning, 1*, 153-182.

Lowe, E. J. (1993). Rationality, deduction and mental models. In K. I. Manktelow & D. E. Over (Eds.), *Rationality. Psychological and philosophical perspectives* (pp. 211-230). London: Routledge.

Lütkemeier, E., Westermann, R., & Gerjets, P. (2003). Induktive Vermutungen über die Anwendbarkeit von Theorien und Methoden nach erfolgreichen und gescheiterten Anwendungsfällen. *Zeitschrift für Psychologie, 211*, 38-56.

Lüer, G., & Spada, H. (1992). Denken und Problemlösen. In H. Spada (Ed.), Lehrbuch Allgemeine Psychologie (2nd. ed., pp. 189-280). Bern: Hans Huber.

Lumer, Ch. (1990). Induktion. In H. J. Sandkühler (Ed.), *Europäische Enzyklopädie zu Philosophie und Wissenschaften* (pp. 659-676). Hamburg: Felix Meiner.

MacIntyre, A. (1985/1981). *After virtue: A study in moral theory.* 2nd ed. London: Gerald Duckworth & Co.

Mackie, J. L. (1963). The Paradox of Confirmation. *The British Journal for the Philosophy of Science, 13*, 265-277.

Mahler, P. (1999). Inductive Logic and the Ravens Paradox. *Philosophy of Science, 66,* 50-70.

Mally, E. (1926). *Grundgesetze des Sollens: Elemente der Logik des Willens*. Graz: Leuschner & Lubensky.

Manktelow, K. I., & Evans, J. St. B. T. (1979). Facilitation of reasoning by realism: Effect or non-effect. *British Journal of Psychology, 70,* 477-488.

Manktelow, K. I., & Fairley, N. (2000). Superordinate principles in reasoning with causal and deontic conditionals. *Thinking and Reasoning, 6*, 41-65.

Manktelow, K. I., & Over, D. E. (1990). Deontic thought and the selection task. In K. J. Gilhooly, M T. G. Keane, R. H. Logie, & G. Erdos (Eds.), *Lines of thinking, reflections on the psychology of thinking: Vol. 1.* (pp. 153-164). Chichester, UK: Wiley.

Manktelow, K. I., & Over, D. E. (1991). Social role and utilities in reasoning with deontic conditionals. *Cognition, 43*, 183-186.

Manktelow K.I., & Over, D.E. (1992). Utility and deontic reasoning: some comments on Johnson-Laird and Byrne. *Cognition, 43,* 183-188.

Manktelow, K. I., & Over, D. E. (1995). Deontic reasoning. In St. E. Newstead & J. St. B. T. Evans (Eds.), *Perspectives on thinking and reasoning. Essays in honour of Peter Wason* (pp. 91-114). Hove, UK: Erlbaum.

Manktelow, K. I., Sutherland, E. J., & Over, D. E. (1995). Probabilistic factors in deontic reasoning. *Thinking & Reasoning*, *1*, 202-220.

McKenzie, C. R. M., Feerreira, C. S., Mikkelsen, L. A., McDermott, K- J., & Skrable, R. P. (2001). Do conditional statements target rare events? *Organizational Behavior & Human Decision Processes, 85*, 291-309.

McKenzie, C. R. M., & Mikkelsen, L. A. (2000). The psychological side of Hempel's paradox of confirmation. *Psychonomic Bulletin Review, 7*, 360-366.

Meder, B. (2006). *Seeing versus doing: Causal Bayes nets as psychological models of causal reasoning.* Unpublished doctoral dissertation, Universität Göttingen.

Metzner, N. (2006). *Cheater- und Cooperator-Detection bei Verbots- und Verpflichtungsregeln*. Unveröffentlichte Diplomarbeit in Psychologie (Betreuer: M. v. Sydow) am Institut für Psychologie. Göttingen: Universität Göttingen.

Moore, G. E. (1903). *Principia ethica.* Cambridge, etc.: Cambridge University Press.

Moshman, D., & Geil, M. (1998). Collaborative reasoning: Evidence for collective rationality. *Thinking and Reasoning, 4,* 231-248.

Nicod, J. (1924). *Le problem logique de l'induction.* Paris: Alcan.

Nickerson, R. S. (1996). Hempel's paradox and Wason's selection task: Logical and psychological puzzles of confirmation. *Thinking and Reasoning, 4,* 231-248.

Nida-Rümelin, J. (2001). *Strukturelle Rationalität.* Stuttgart: Reclam.

Novick, L. R., & Cheng, P. W. (2004). Assessing Interactive Causal Influence. *Psychological Review, 111*, 455-485.

Nortmann, U. (1989). *Deontische Logik ohne Paradoxien- Semantik und Logik des Normativen* (PhD Thesis, Univ. Göttingen, 1985). München: Philosophia Verlag.

Oaksford, M. (1998). Task demands and revising probabilities in the selection task: A Comment on Green, Over, and Pyne. *Thinking and Reasoning, 4,* 179-186.

Oaksford, M. (2001). Language processing, activation and reasoning: A reply to Espino, Santamaría, and García-Madruga (2000). *Thinking and Reasoning, 7*, 205-208.

Oaksford, M. (2002). Predicting the results of reasoning experiments: Reply to Feeney and Handley (2000). *The Quarterly Journal of Experimental Psychology, 55 A*, 793-798.

Oaksford, M., & Chater, N. (1991). Against logicist cognitive science. *Mind and Language, 6,* 1-38.

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. (1995a). Information gain explains relevance which explains the selection task. *Cognition, 57,* 97-108.

Oaksford, M., & Chater, N. (1995b). Theories of reasoning and the computational explanation of everyday inference. *Thinking and Reasoning, 1,* 121-152.

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

Oaksford, M., & Chater, N. (1998a). *Rationality in an uncertain world. Essays on the cognitive science of human reasoning.* Hove, UK: Psychology Press.

Oaksford, M., & Chater, N. (1998b). A revised rational analysis of the selection task: Exceptions and sequential sampling. In M. Oaksford & N. Chater (Eds.), *Rational models of cognition* (pp. 372-398). Oxford: Oxford University Press.

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

Oaksford, M., & Chater, N. (2003). Optimal data selection: Revision, review, and reevaluation*. Psychological Bulletin & Review, 10*, 289-318.

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

Oaksford, M., Chater, N., Grainger, B, & Larkin, J. (1997). Optimal data selection in the reduced array selection task (RAST). *Journal of Experimental Psychology: Learning, Memory, and Cognition, 23,* 441-458.

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

Oaksford, M., & Sellen, J. (2000). Paradoxical individual differences in conditional inference. *Behavioral and Brain Sciences, 23*, 691-692.

Oaksford, M. & Wakefield, M. (2003). Data Selection and Natural Sampling: Probabilities Do Matter. *Memory & Cognition, 31,* 143-154.

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

Oberauer, K., Weidenfeld, A., & Hörnig, R. (2004). Logical Reasoning and probabilities: A comprehensive test of Oaksford & Chater (2001). *Psychonomic Bulletin & Review, 11*, 521-527.

O'Brien, D. (1995). Finding logic in human reasoning requires looking in the right places. In St. E. Newstead & J. St. B. T. Evans (Eds.), *Perspectives on thinking and reasoning. Essays in honour of Peter Wason* (pp. 189-216). Hove, UK: Erlbaum.

Osherson, D. N., Smith, E. E., Wilkie, O., López, A., & Shafir, E. (1990). Category-based induction. *Psychological Review, 97*, 185-200.

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

Ossa, M., & Schönecker, D. (2004). Ist keine Aussage sicher? Rekonstruktion und Kritik der deutschen Fallibilismusdebatte. *Zeitschrift für philosophische Forschung, 58,* 54-79.

Over, D. E. (Ed.). (2003). Evolution and the psychology of thinking: the debate. Hove (GB): Psychology Press.

Over, D. E. (2004). Naïve Probability and Its Model Theory. In V. Girotto, & P. N. Johnson-Laird (Eds.), *The shape of reason: Essays in honour of Paolo Legrenzi.* Hove: Psychology Press.

Over, D. E., & Manktelow, K. I. (1993). Rationality, utility and deontic reasoning. In K. I. Manktelow & D. E. Over (Eds.), *Rationality. Psychological and philosophical perspectives* (pp. 231-259). London: Routledge.

Over, D. E., & Evans, J. S. B. T. (1994). Hits and misses: Kirby on the selection task. *Cognition, 52,* 235-243.

Over, D. E., & Jessop, A. L. (1998). Rational analysis of causal conditionals and the selection task. In M. Oaksford & N. Chater (Eds.), *Rational models of cognition* (pp. 399-414). Oxford: Oxford University Press.

Over, D. E., & Evans, J. S. B. T. (2003). The probability of conditionals: The psychological evidence. *Mind & Language, 18,* 340-358.

Pearl, J. (1988*). Probabilistic Reasoning in Intelligent Systems*. San Francisco: Morgan Kaufman.

Pearl, J. (2000). *Causality. Models, reasoning, and inference.* Cambridge, UK: Cambridge University Press.

Perham, N., & Oaksford, M. (2005). Deontic reasoning with emotional content: Evolutionary psychology or decision theory? *Cognitive Science, 29,* 681-718.

Piaget, J. (1932). *Le jugement moral chez l'enfant.* Paris: Alcan.

Platt, R. D., & Griggs, R. A. (1993). Facilitation in the abstract selection task: The effects of attentional and instructional factors. *Quarterly Journal of Experimental Psychology, 46A ,* 591-613.

Politzer, G., & Nguyen-Xuan, A. (1992). Reasoning about conditional promises and warnings: Darwinian algorithms, mental models, relevance judgments or pragmatic schemas? *Quarterly Journal of Experimental Psychology, 44A,* 401-421.

Pollard, P. (1990). Natural selection for the selection task: Limits to social exchange theory. *Cognition, 36*, 195-204.

Popper, K. R. (1934/1994/2002). *Logik der Forschung.* Jubiläumsausgabe (10th ed.). Tübingen: Mohr Siebeck (engl. ed. in 1959. *The logic of scientific discovery.* London: Hutchingson).

Popper, K. R. (1972/1991). *Objective knowledge: An evolutionary approach*. Oxford, UK: Oxford University Press.

Popper, K. R. (1974). Replies to my critics (Book II). In P. A. Schilpp (Ed.), *The philosophy of Karl Popper* (pp. 961-1197). Illinois: Open Court.

Popper, K. R. (1977). On hypothesis (Excerpts from K. R. Popper, 1976, *Unended quest.* Glasgow: Fontana). In P. N. Johnson-Laird & P. Wason (Eds.), *Thinking* (pp. 264-273). Cambridge: Cambridge University Press.

Popper, K. R. (1996/2004). Alles Leben ist Problemlösen. München: R. Piper.

Putnam, H. (1974). The 'corroboration' of theories'. In Schilpp, P. A. (Ed.), *The philosophy of Karl Popper* (pp. 221-240). La Salle, Ill.: Open Court Publishing.

Quelhas, A. C., & Byrne, R. M. J. (2003). Reasoning with deontic and counterfactual conditionals. *Thinking and Reasoning, 9*, 43-65.

Quine, W. V. O. (1974). On Popper's negative methodology. In Schilpp, P. A. (Ed.), *The philosophy of Karl Popper* (pp. 218-220). La Salle, Ill.: Open Court Publishing.

Quine, W. V. O. (1969). *Ontological relativity and other Essays.* New York: Columbia University Press.

Rawls, J. (1999/1971). *A Theory of Justice.* (Revised edition). Oxford: Oxford University Press.

Rips, L. J. (1990). Reasoning. *Annual Review of Psychology, 41*, 321-353.

Rips, L. J. (1994). *The psychology of proof.* Cambridge, MA: MIT Press.

Rosenkrantz, R. D. (1982). Does the philosophy of induction rest on a mistake? *Journal of Philosophy, 79*, 78-97.

Russell, B. (1991/1961). *History of Western philosophy* (2nd. ed.; first edition published 1946). London: Routledge.

Santamaria, C., Espino, O., & García-Madruga, J. A. (2001). Theories of reasoning and the representational level: A reply to Oaksford. *Thinking and Reasoning, 7,* 209-213.

Schroyens, W. & Schaeken, W. (2003). A critique of Oaksford, Chater, and Larkin's (2000) conditional probability model of conditional reasoning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29*, 140-149.

Siebel, M. (2004). Der Rabe und der Bayesianist. *Journal for General Philosophy of Science - Zeitschrift für Wissenschaftstheorie, 2*, 313-329.

Sloman, S. A., Over, D, Slovak, L., & Stibel, J. M. (2003). Frequency illusions and other fallacies. *Organizational Behavior and Human Decision Processes, 91*, 296-309.

Sober, E., & Wilson, D. S. (1998). *Unto others. The evolution and psychology of unselfish behaviour.* Cambridge, MA: Harvard University Press.

Sperber, D., Cara, F., & Girotto V. (1995). Relevance theory explains the selection task. *Cognition, 57*, 31-95.

Sperber, D., & Girotto, V. (2002). Use or misuse of the selection task? Rejoinder to Fiddick, Cosmides, and Tooby. *Cognition, 85*, 277-290.

Sperber, D., & Girotto, V. (2003). Does the selection task detect cheater-detection? In J. Fitness & K. Sterelny (Eds.), *New directions in evolutionary psychology. Macquarie monographs in cognitive science.* Sidney: Psychology Press.

Staller, A., Sloman, S. A. & Ben-Zeev, T. (2000). Perspective effects in nondeontic versions of the Wason selection task. *Memory & Cognition, 28*, 396-405.

Stanovich, K.-E., & West, R. F. (2000). Individual differences in reasoning: Implications for the rationality debate? *Behavioral and Brain Sciences, 23,* 645-726.

Stone, V. E., Cosmides, L., Tooby, J., Kroll, N., & Knight, R. T (2002). Selective impairment of reasoning about social exchange in a patient with bilateral limbic system damage. *Proceedings of the National Academy of Sciences, 99,* 11531-11536.

Suppes, P. (1966). A Bayesian Approach to the Paradoxes of Confirmation. In J. Hintikka, P. Suppes (Eds.), *Aspects of inductive logic* (pp. 198-207). Amsterdam: N-Holland PC.

Staller, A., Sloman, S. A., & Ben-Zeev, T. (2000). Perspective effects in nondeontic versions of the Wason selection task. *Memory & Cognition, 28,* 396-405.

Stanovich, K.-E., & R. F. West (2000). Individual differences in reasoning: Implications for the rationality debate? *Behavioral and Brain Sciences, 23,* 645-726.

von Sydow, M. (2001). *Sociobiology, universal Darwinism and their transcendence.* Doctoral Dissertation, Department of Philosophy, University of Durham, UK.

von Sydow, M. (2002). *Probabilistisches Prüfen von wenn-dann-Hypothesen.* Diplomarbeit Institut für Psychologie. Supervisor: E. Erdfelder, J. Bredenkamp. Bonn: Universität Bonn.

von Sydow, M. (2004a). 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* (pp. 1411-1416). Mahwah, NJ: Erlbaum.

von Sydow, M. (2004b). Experimentelle Befunde zur bayesianischen Aufhebung des Rabenparadoxons. In D. Kerzel, V. Franz, & K. Gegenfurtner (Eds.), *Beiträge zur 46. Tagung experimentell arbeitender Psychologen in Gießen* (S. 270). Lengerich: Pabst.

von Sydow, M. (2004c). Strukturmodelle des Hypothesentestens. Arbeitsgruppe Kausalität. In T. Rammsayer, S. Grabianowski, S. Troche (Eds.), *44. Kongress der Deutschen Gesellschaft für Psychologie. 100 Jahre DGP* (S. 279). Lengerich: Pabst.

von Sydow, M. (2005a). Cooperation Detection and the Flexible Deontic Logic Theory of the WST. In B. G. Bara, L. Barsalou, & M. Bucciarelli (Eds.), *Proceedings of the Twenty-Seventh Annual Conference of the Cognitive Science Society* (p. 2573). Mahwah, NJ: Erlbaum.

von Sydow, M. (2006). *‚Bayes-Logik'. Flexibles bayesianisches Testen logischer Hypothesen.* (Förderantrag bei der DFG; Bewilligungsdatum 3.3.2006).

von Sydow, M. (submitted). Cooperation detection and the deontic logic of the Wason selection task.

von Sydow, M., Hagmayer, Y., Meder, B., & Waldmann, M. R. (2005). Flexibles bayesianisches Hypothesentesten. In K. W. Lange et. al. (Eds.), *Beiträge zur 47. Tagung experimentell arbeitender Psychologen* (S. 213). Lengerich: Pabst.

von Sydow, M. & Hagmayer, Y. (2006). Deontic Logic and Deontic Goals in the Wason Selection Task. In R. Sun (Ed.). *Proceedings of the Twenty-Eighth Annual Conference of the Cognitive Science Society* (pp. 864-869). Mahwah, NJ: Erlbaum.

von Sydow, M., Hagmayer, Y., Metzner, N., & Waldmann, M. R. (2005). Cooperation detection and deontic reasoning in the Wason selection task. In K. Opwis & I.-K. Penner (Eds.), *Proceedings of KogWis05. The German Cognitive Science Conference 2005* (pp. 195-200). Basel: Schwabe.

Trivers, R. L. (1971). The evolution of reciprocial altruism. *Quarterly Review of Biology, 46,* 35-57.

Vranas, P. B. M. (2004). Hempel's Raven Paradox. Lacuna in the standard Bayesian solution. *British Journal for Philosophy of Science, 55,* 545-560

Waldmann, M. R. (1996). Knowledge-based causal induction. In D. R. Shanks, K. J. Holyoak, & D. L. Medin (Eds.), *The psychology of learning and motivation, Vol. 34: Causal learning* (pp. 47-88). San Diego: Academic Press.

Waldmann, M. R., & Hagmayer, Y. (in press). Categories and causality. The neglected direction. *Cognitive Psychology*.

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

Waldmann, M. R., & Holyoak, K. J. (1992). Predictive and diagnostic learning within causal models: Asymmetries in cue competition. *Journal of Experimental Psychology: General, 121*, 222-236.

Waldmann, M. R., & von Sydow, M. (2006). Wissensbildung, Problemlösen und Denken (pp. 217-229). In K. Pawlik (Ed.), *Springer Handbuch Psychologie.* Kapitel 15. Berlin: Springer Verlag.

Warneken, F. & Tomasello, M. (2006). Altruistic Helping in Human Infants and Young Chimpanzees. *Science, 311,* 1301-1303.

Wason, P. C. (1966). Reasoning. In B. M. Foss (Ed.), *New horizons in psychology* (pp. 135-151). Harmondsworth, Middlesex, UK: Penguin.

Wason, P. C. (1968). Reasoning about a rule. *Quarterly Journal of Experimental Psychology, 20*, 273-281.

Wason, P. C. (1977). Self-contradictions. In B. Foss (Ed.), *New horizons in psychology* (pp. 135-151). Harmondsworth, UK: Penguin.

Wason, P. C., & Johnson-Laird, P. N. (1969). Proving a disjunctive rule. *Quarterly Journal of Experimental Psychology*, *21*, 14-20.

Wason, P. C., & Shapiro, D. (1971). Natural and contrived experience in a reasoning problem. *Quarterly Journal of Experimental Psychology, 23*, 63-71.

Watkins, J. W. N. (1957). Between analytic and empirical. *Philosophy, 32,* 112-131.

Watkins, J. W. N. (1958). A rejoinder to Professor Hempels reply. *Philosophy, 33,* 349-355.

Westermann, R, & Gerjets, P. (1994). Induktion. In Th. Herrmann & W. H. Tack (Eds.), *Enzyklopädie der Psychologie. Methodologische Grundlagen der Psychologie: Band B/1/1/10* (pp. 429-471). Göttingen: Hogrefe Verlag.

White, P. A. (2000). Causal judgement from contingency information: relation between subjective reports and individual tendencies in judgement. *Memory and Cognition, 28*, 415-426.

Whitehead, A. N., & Russell, B. (1999/1925). *Principia mathematica* (translated from English by H. Mokre; originally published by Cambridge University Press). Frankfurt am Main: Suhrkamp.

Wiener, N. (1948). *Cybernetics.* New York: Wiley.

Wittgenstein, L. (1990/1921). *Traktatus logico-philosophicus*. Leipzig: Reclam.

von Wright, G. H. (1966). The Paradoxes of Confirmation. In J. Hintikka & P. Suppes (Eds.), *Aspects of inductive logic* (pp. 208-218). Amsterdam: N-Holland PC.

von Wright, G. H. (1981). *On the logic of norms and actions.* In R. Hilpinen (Ed.), *New studies in deontic logic. Norms, actions, and the foundation of ethics* (pp. 3-35). Dordrecht: Reidel Publishing.

von Wright, G. H. (1994). *Normen, Werte und Handlungen.* Frankfurt am Main: Suhrkamp.

Yama, H. (2001). Matching versus optimal data selection in the Wason selection task. *Thinking and Reasoning, 7*, 295-311.

Yama, H. (2002). Context, goal, utility, and relevance: A reply to Evans (2002) considering Oaksford (2002), *Thinking and Reasoning, 8*, 225-230.

Curriculum Vitae

Dr. Momme v. Sydow

Akademischer Werdegang

2006	Projektleiter und wissenschaftlicher Mitarbeiter in eigenem DFG-Projekt „Bayes-Logik“, Georg-August-Universität Göttingen, Georg-Elias-Müller-Institut für Psychologie
2006	Zweit-Promotion in Psychologie (im Mai), Universität Göttingen, Doktorvater: Professor Dr. Michael Waldmann
2003-2006	Wissenschaftlicher Mitarbeiter (bei Prof. Dr. Michael Waldmann) im DFG-Projekt „Kategorisierung und induktives Lernen“, Schreiben der Doktorarbeit in Psychologie, Universität Göttingen
2002	Abschluss des Psychologiestudiums (Universität Bonn), Diplomarbeit: Professor Dr. Edgar Erdfelder
2001	Verleihung des Doktorgrades an der University of Durham (PhD) im Fach Philosophie. Doktorvater: Professor Dr. David Knight
1996-1999	University of Durham (UK), Studium der Philosophie
1992-1996	Doppelstudium der Philosophie (Magisterstudiengang) und Psychologie (Diplomstudiengang) an der Rheinischen Friedrich-Wilhelms-Universität Bonn
	Studentische Hilfskraftstelle in einem DFG Projekt in Psychologie
	Referententätigkeit AStA der Universität Bonn, Bundessprecher der Bundesfachschaftentagung Philosophie

Stipendien und Forschungsprojekte

2006	Eigenes DFG-Projekt „Bayes-Logik – Flexibles bayesianisches Testen logischer Hypothesen“, DFG Az.: Sy 111/1/1-1
1996-2002	Stipendium der Heinrich-Böll-Stiftung; DAAD-Kurzstipendium; Stipendium Kölner Gymnasial- und Stiftungsfonds

Sonstiges

2004	„Handbuch Studium und Praktikum im Ausland“ (Eichborn Verlag)
2002-2003	Projekttätigkeiten beim Deutschen Studentenwerk (DSW),
	Projekttätigkeiten beim studentischen Dachverband (fzs)
1996 bis heute	Herausgeber Studienführer Philosophie (Akademia Verlag), aktualisierte Internetversion: *www.philos.de*
1992-1995	Deutsche Presseagentur, Bundesbüro, Assistenz