IF THERE ARE EXCEPTIONS, IT IS STILL A RULE - A PROBABILISTIC UNDERSTANDING OF CONDITIONALS

Dissertation von Sonja Maria Geiger

Vorgelegt zur Erlangung des akademischen Grades der Doktorin der Philosophie.

Eingereicht an der Humanwissenschaftlichen Fakultät der Universität Potsdam
im Fach Psychologie Januar 2007.

Danksagung

Es war einmal ein alter chinesischer Bauer, der war sehr arm und hatte nur ein Pferd. Eines morgens, als er erwachte, war das Pferd war davongelaufen. Sein einziger Sohn brach in Tränen und Gezeter aus: „Das ist ja furchtbar..." Wie sollte er den jetzt das Feld bestellen, wie sollte er das Gemüse zum Markt tragen, wovon sollten sie leben? Der Bauer antwortete: „Furchtbar? Das kann man nie wissen, manchmal ein Segen, manchmal ein Fluch". Am nächsten Morgen kam das Pferd zurück, eine ganze Herde Wildpferde im Gefolge. Der Sohn brach in Jubel aus: „Das ist ja wunderbar, welch Reichtum!" Dar Bauer antwortete: „Wunderbar? Das kann man nie wissen, manchmal ein Segen, manchmal ein Fluch." Kurz darauf brach sich der Sohn beim Zureiten der Wildpferde ein Bein. Die Nachbarn liefen herbei: „Oh wie schrecklich, der einzige Sohn, wer sollte ihm jetzt beim Bestellen der Felder helfen und wie wollte er durch den Winter kommen?" Der Bauer sagte aber nur (ja genau): „Manchmal ein Segen, manchmal ein Fluch...". Etwas später kam die Armee des Königs vorbei und zog alle jungen Männer mit in den Krieg. Der Sohn des Bauers aber blieb verschont.
Was mich diese Doktorarbeit vor allem gelehrt hat, ist die Dinge etwas mehr dem chinesischen Bauer würdig anzugehen. Manchmal ein Segen, manchmal ein Fluch.

Abgesehen von dieser generellen Lehre habe ich auch noch viel Spezielles gelernt. Die Person, der ich dafür als erstes und über alles danken möchte ist mein Betreuer, Klaus Oberauer. Er hat mich nicht nur immer wieder mit seinem Humor, seiner Großzügigkeit und dem Beispiel an erfolgreicher, egofreier und sachlicher Wissenschaft beeindruckt, sondern auf den letzten Metern mit der Entdeckung, dass ein Lob Wunder wirkt, maßgeblich zur Fertigstellung der schriftlichen Arbeit beigetragen. Dass diese Promotion mit sehr lehrreichen Erfahrungen an ausländischen Universitäten und internationalen Konferenzen verbunden war, habe ich vor allem seiner Unterstützung zu verdanken. Herzlichsten Dank.

Ebenfalls herzlichen Dank möchte ich auch meinem Zweitgutachter Reinhold Kliegl aussprechen. Einer der ersten in Segen aufgegangenen Flüche verdanke ich ihm: „Man braucht für alles immer so lange, wie man Zeit hat". Recht hat er. Ebenfalls in Potsdam gelernt hab ich wohl, wie wissenschaftliches Arbeiten in einer erfolgreichen Abteilung aussieht und dadurch auch meinen eigenen Maßstab noch mal nachjustiert (ob das jetzt in die Segen oder Fluchkategorie gehört, will ich offen lassen). Vielen Dank.

Zwei Menschen, die meine Arbeit um ein Unglaubliches bereichert haben (nicht nur inhaltlich): Annekatrin Hudjetz, ehemals: die beste studentische Hilfskraft der Welt! Berry Claus, eine selten erlebte Mischung aus Geduld, Humor und fast hellseherischem Verständnis selbst komplett fragmentiert ausgedrückter Gedanken. Tausend Dank.

Allen meinen Kollegen, die mir die auch zum Teil anstrengende Zeit in Potsdam und vor allem die lange Fahrt dahin versüßt haben: Robin, Anja, Katrin, Jochen, Antje, Sarah, Thomas und in unserer Abteilung eben: u. v. m. Danke. Meinen internationalen Kollegen, die den zahlreichen Auslandreisen eine weitere Dimension gaben: Caren Frosch, Helen Neilens, Simon Handley, Shira Elquayam, Wim deNeys, Niki Verschueren – thanks heaps!

Meinen Eltern, die trotz der Einstellung „promovieren = immer noch nicht fertig sein" meinen Lebensweg meinen Lebensweg sein lassen. Danke für die Freiheit. Meinen Geschwistern, deren Anerkennung mir mehr bedeutet, als ich vielleicht zugeben will. Jutta, Heike, Andrea: Danke. Herrn Dengler, der seine letzten Worte an mich der Durchführung dieser Arbeit gewidmet hat. Vielen Dank für diesen Abschied.

Fürs immer wieder die Perspektive ver-rücken: Eti, Katrin, Joachim (auch fürs Zusammenwohnen aushalten). Der Dank ist so außerhalb von Kategorien wie die Hilfeleistung. Last, but not least mal wieder: für zahlreich angehörte (Probe-)Talks, für Aufbauarbeiten nach diversen Fehlschlägen, für gemeinsames Durchleben wissenschaftlicher Sinnkrisen, unterwegs in dieselbe Richtung, ein Fortsetzungsdank: Friedi, Elin, Corinna, ohne Euch? Wäre ich nicht nur nicht Doktorin der Philosophie. Ohne Worte.

Abstract

Numerous recent publications on the psychological meaning of "if" have proposed a probabilistic interpretation of conditional sentences. According to the proponents of probabilistic approaches, sentences like "If the weather is nice, I will be at the beach tomorrow" (or "If p, then q" in the abstract version) express a high probability of the consequent (being at the beach), given the antecedent (nice weather). When people evaluate conditional sentences, they assumingly do so by deriving the conditional probability P(q|p) with means of a procedure called the Ramsey test. This is a contradicting view to the hitherto dominant Mental Model Theory (MMT, Johnson-Laird, 1983), that proposes conditional sentences refer to possibilities in the world that are represented in form of mental models.

Whereas probabilistic approaches gained a lot of momentum in explaining the interpretation of conditionals, there is still no conclusive probabilistic account of conditional reasoning. This thesis investigates the potential of a comprehensive probabilistic account on conditionals that covers the interpretation of conditionals as well as conclusion drawn from these conditionals when used as a premise in an inference task.

The first empirical chapter of this thesis, Chapter 2, implements a further investigation of the interpretation of conditionals. A plain version of the Ramsey test as proposed by Evans and Over (2004) was tested against a similarity sensitive version of the Ramsey test (Oberauer, 2006) in two experiments using variants of the probabilistic truth table task (Experiments 2.1 and 2.2). When it comes to decide whether an instance is relevant for the evaluation of a conditional, similarity seems to play a minor role. Once the decision about relevance is made, believability judgments of the conditional seem to be unaffected by the similarity manipulation and judgments are based on frequency of instances, in the way predicted by the plain Ramsey test.

In Chapter 3 contradicting predictions of the probabilistic approaches on conditional reasoning of Verschueren et al (2005), Evans and Over (2004) and Oaksford & Chater (2001) are tested against each other. Results from the probabilistic truth table task modified for inference tasks support the account of Oaksford and Chater (Experiment 3.1). A learning version of the task and a design with every day conditionals yielded results unpredicted by any of the theories (Experiments 3.2-3.4). Based on these results, a new probabilistic 2-stage model of conditional reasoning is proposed.

To preclude claims that the use of the probabilistic truth table task (or variants thereof) favors judgments reflecting conditional probabilities, Chapter 4 combines methodologies used by proponents of the MMT with the probabilistic truth table task. In three Experiments (4.1 -4.3) it could be shown for believability judgments of the conditional *and* inferences drawn from it, that causal information about counterexamples only prevails, when no frequencies of exceptional cases are present. Experiment 4.4 extends these findings to every day conditionals. A probabilistic estimation process based on frequency information is used to explain results on all tasks. The findings confirm with a probabilistic approach on conditionals and moreover constitute an explanatory challenge for the MMT.

In conclusion of the evidence gathered in this dissertation it seems justified to draw the picture of a comprehensive probabilistic view on conditionals quite optimistically. Probability estimates not only explain the believability people assign to a conditional sentence in the present experiments, they also explain to what extend people are willing to draw conclusions from those sentences.

Content

List of Tables

List of Figures

Chapter 1: Introduction

Neal: "If you mean that there's an exception to every rule, then I am with you."

God: "If there's an exception to a rule, then it's not a rule."

(Walsh 1997, p.276)

There are proverbs in both English and German that express the human view in the above quotation. That "exceptions confirm the rule", which is the German version of the proverb, summarizes people's everyday experience that no general rule is beyond doubt and it should not be completely discarded in the face of some occasional exceptions.

That the notion of certainty coming in degrees could be applied to the interpretation of conditional sentences has first been proposed by philosophers (Adams, 1981; Edgington, 1991, 1995). From these authors comes the notion that the believability of a conditional of the form: "if p, then q" is based on the conditional probability of its consequence q, given the antecedent p, which is P(q|p). Recently, cognitive psychologists (Evans & Over, 2004) have formalized this notion of the "suppositional conditional" into a psychological theory of the interpretation of conditional sentence. The authors of this suppositional account claim that people interpret a conditional sentence by first supposing that p is true (hence the name of the theory). For a conditional like "If the weather is nice, then I'll be at the beach" people are thought to imagine situations where p is the case (the weather is nice) and evaluate the probability of q (going to the beach) within this suppositional frame. Non-p situations (where the weather is bad) do not play a role for the believability of the above statement. According to the suppositional account, the degree of confidence that people place in the conditional rule, is best reflected by their subjective estimation of this conditional probability.

This probabilistic approach on conditionals contrasts the hitherto main theory on conditional reasoning, the theory of mental models (Johnson-Laird, 1983; Johnson-Laird, 2001; Johnson-Laird & Byrne, 2002). The mental model theory is based on representations of possible states of the world, which can be either true or false in regard to the assertion of the conditional statement. Introducing a *principle of truth*, the authors assume that people only represent what is true or permissible, given the truth of the conditional statement. For the above example this would be situations with nice weather at the beach and also all situations in which the weather is bad, since rainy days (no matter what I'll do then) are not precluded by the conditional rule.

A recent probabilistic extension of the mental model theory (Johnson-Laird, Legrenzi, Girotto, Legrenzi, & Caverni, 1999) also allows people to have gradual beliefs in a conditional statement, according to the proportion of all represented models to all possible situations in the world, P(MM). However, this probability can never equal the probabilistic model's P(q|p), as it was shown in a famous proof by Lewis (1976). This can be intuitively understood if one considers the non-p instances (here: bad weather days); in probabilistic theories they do not play a role, in the mental model theory they do. Thus, regarding the believability of a conditional sentence the two different theories cannot both be right.

A yet to be answered question is, in how far the probabilistic account on interpretation of conditionals can be extended to explain the reasoning with conditional sentences. Over the last years there have been different formulations of theories that use subjective probabilities to explain reasoning with conditionals (Oaksford & Chater, 2001; Oaksford, Chater, & Larkin, 2000; Verschueren, Schaeken, & d`Ydewalle, 2005). One of the main differences in probabilistic accounts on reasoning compared to the mental model account is the representation and handling of counterexamples to inferences from conditional sentences. If somebody wanted to conclude from the fact that it is a nice day today, that I am at the beach (Example of a Modus Ponens inference, see Table 2), this could either be done by relying on the subjective probability P(q|p) or by representing the different possibilities using mental models. Whereas in the former alternative a couple of missed out sunny days would be permissible to still conclude that I would most likely be at the beach today, in the latter possibility it would not. Should the person come up with the model of a counterexample, a Modus Ponens inference is not longer justified, since there are two possibilities containing the minor premise p (sunny day): a model representing the beach and one representing other whereabouts, blocking one definite conclusion.

This dissertation provides more evidence for a probabilistic interpretation of conditionals, tests different probabilistic versions of conditional reasoning and tests the two opposing approaches on conditional reasoning in general in an overarching framework. In the next sections, the main aspects of both theories (or family of theories, respectively) will be presented on both issues, that is interpretation of conditionals and reasoning from them. Before moving to the empirical part of this dissertation, I will give a more detailed outline of the remaining chapters at the end of this introduction.

1.1 The Mental Model Theory of conditionals

In reasoning with conditionals the most prominent theory of the last two decades has been the Mental Model Theory (MMT) by Johnson-Laird (Johnson-Laird, 1983; Johnson-Laird, 2001; Johnson-Laird & Byrne, 1991; Johnson-Laird & Byrne, 2002). Mental models are used to explain a whole range of findings in deductive reasoning, such as syllogistic reasoning or reasoning about relations. In general, a mental model refers to an inner representation of an external situation or state of the world. A mental model is a compressed summary of whatever it represents, and most importantly retains critical features and properties of the represented external state. In the field of conditional reasoning, Johnson-Laird and collaborators formulated a theory that attempts to explain how people interpret conditional sentences and subsequently assign a certain degree of believability to it, as well as draw inferences from them.

1.1.1 Interpretation of the conditional

According to the MMT account (Johnson-Laird & Byrne, 1991), people use their knowledge about the world to represent the meaning of a conditional sentence in a set of mental models. A mental model here refers to the representation of a possible state of the world that is compatible with the truth conditions of the conditional sentence (Table 1). The authors distinguish between an initial model and an expanded set of models. Thus people initially form an incomplete representation of the conditionals by concentrating on the pq possibility only, though allowing for further possibilities to exist. For a conditional sentence such as "If I open the fridge the light goes on" or the abstract form "if p, then q", the initial model would look like this:

Open fridge	Light on	(pq case)
	…	

The three dots represent further possibilities that haven't yet been represented, and are thus called implicit models. Complying with the *principle of truth*, this model can be fleshed out to construct fully explicit models of what is true depending on the person's capability to hold more than one model at a time:

Open fridge	Light on	(pq case)
Closed fridge	Light on	(¬pq case)
Closed fridge	Light off	(¬p¬q case)

In this notation, "p" stands for the antecedent of the conditional, "q" for the consequent of the conditional and "¬" for the negation of either. The notation established by the authors allots a separate row for each mental model or possibility in the world. In a later specification of their theory (Johnson-Laird & Byrne, 2002) they introduce a principle called "pragmatic modulation". This

principle introduces the possibility that instead of the three model mentioned above, different combinations of models including more or less then three can be built depending on the content and context of the conditional in question.

The representation of what they call a "basic conditional", which is a conditional with abstract content to which no former knowledge can be connected, cannot be subject to pragmatic modulation. Therefore, a basic conditional is always represented by the initial model containing the mental footnote of further possibilities, or the fully explicit set of models, respectively. The core meaning of a fully represented basic conditional "If p, then q", is therefore:

$$p\,q$$
$$\neg p\ q$$
$$\neg p\ \neg q$$

and thus the represented possibilities correspond to the truth conditions of a material conditional (see Table 1). Findings with the classical truth table task led to a further refinement of this conceptualization of the "core meaning" of conditionals. In the classical truth table task people are presented with the four truth table cases and asked whether they make the conditional sentence true, false or are irrelevant for the truth or falseness of the conditional (Johnson Laird & Tagart, 1969). People seem to have a "defective truth table" for a conditional (Wason, 1966, see also Table 1), that is, they judge the $\neg p$ cases as being irrelevant for the truth of the conditional. This result has been replicated with different methods ever since (e.g. Wason & Johnson-Laird, 1972; Evans, Newstead & Byrne, 1993). Johnson-Laird and Byrne (1991) explain the defective truth table pattern with the inability of most people to flesh out the full set of explicit models, such judging the yet unrepresented possibilities as irrelevant.

Table 1: Truth table with example

Possible states of the world:		**(1) Material conditional**	**(2) Defective truth table**
p q	Fridge open, light on	True	True
p¬q	Fridge open, light off	False	False
¬pq	Fridge closed, light on	True	Irrelevant
¬p¬q	Fridge closed, light off	True	Irrelevant

Truth table for the conditional "If you open the fridge, then the light goes on" (="If p, then q"). Column (1) summarizes normative judgments about the truth or falsity of the four truth table cases according a material conditional, Column (2) summarizes judgments according to the defective truth table, a common response pattern in classic truth table task.

As mentioned above, the theory of mental models introduced a probabilistic extension to mental models, allowing gradual believability judgments of conditionals, (Girotto & Johnson-Laird, 2004; Johnson-Laird, Legrenzi, Girotto, Legrenzi, & Caverni, 1999). According to this extension, the probability of a conditional is the probability that one of the models representing the conditional refers to a true situation. Results of the extensive testing of predictions derived from this notion will be reviewed in the section on probabilistic theories on conditionals in Chapter 1.2.1.

No such probabilistic extension has been formulated for the reasoning with conditionals, as will be discussed in the next section.

1.1.2 Reasoning from conditionals

Drawing inferences from conditional sentences is conceptualized as a three staged process, including a model-construction phase, a conclusion-formulation phase and a conclusion-validation phase. If people are given an argument of Modus Ponens (MP) that consists of a major and a minor premise (see Table 2), they ideally will first construct the following models to represent the major premise "If p then q":

p q

¬p q

¬p¬q

The integration of the minor premise of the MP "p is given" rules out all models beside the first one:

p q

from which can putatively be concluded that q must also be the case. Likewise, the minor premise of Modus Tollens (MT) "¬q is given" is only compatible with what is represented in the third model,

¬p¬q

and thus the other two models are eliminated and "¬q" is putatively inferred. In a third stage, people search for alternative models of the premises that might falsify the putative conclusion and if none is found, the conclusion is accepted.

Table 2: Examples for the four inference patterns

	Major premise: "If you open the fridge, then light comes on" **(Conditional)** "If p then q"	
	Minor premise (observation)	**Default conclusion** (inference acceptance)
MP	The fridge is open (p).	The light is on (q).
AC	The light is on (q).	The fridge is open (p).
DA	The fridge is not open (¬p).	The light is not on (¬q).
MT	The light is not on (¬q).	The fridge is not open (¬p).

In a specification of the MMT, Schaeken, Schroyens and d`Ydevalle (2002) placed additional emphasis on the third stage which they formulate as a validation-by-falsification stage. According to the authors, reasoners explicitly search for counterexamples to the putative conclusion, not just any alternative. For the two inference patterns MP and MT a counterexample consists of the p¬q case. If people can think of a p¬q case, they will deny the inferences; if they can't, they should accept both of the inferences.

The common finding that endorsement rates of MT are always lower than for MP can be explained within the construction phase. If people represent the major conditional premise with the initial *p q* model only, the categorical premise for MT, *¬q* cannot be integrated and people falsely infer that nothing follows. In order to draw the MT inference according to MMT, it is necessary that people flesh out all true possibilities in the model construction phase before they could potentially falsify the conclusion in the validation phase. For endorsement of MP it is sufficient to represent the initial model consisting of the *pq* model only.

Whereas the minor premises for Affirmation of the Consequent (AC), *q is given*, rules out the fourth model, the minor premise of Denial of the Antecedent (DA), *¬p is given* rules out the first model. This leaves two remaining models for either inference pattern after the construction phase, so the correct answer would be that nothing can be inferred. Again, if people represent the major premise with the initial *pq* model only, they will infer *p* from the minor premise *q* and thus erroneously endorse AC. To endorse DA, the *¬p¬q* has to be represented as well, which again explains the mostly higher endorsement rates of AC compared to DA.

1.2 Probabilistic accounts of conditionals

In recent research on conditionals cognitive psychologists have argued that people understand conditionals in a probabilistic way (e.g. Anderson, 1995; Evans & Over, 2004; Oaksford & Chater, 1994, 2001). These theories share the assumption that subjective conditional probabilities determine how much people are willing to believe a conditional sentence or to draw a given conclusion from it. Up to date, there has been no unified probabilistic theory advanced that coherently explains the interpretation of conditionals *and* drawing inferences from them. The next subchapters will first outline a probabilistic interpretation of the conditional and after that present different probabilistic approaches on conditional reasoning.

1.2.1 Interpretation of conditionals

The notion that people base their confidence on a conditional sentence according to their subjective estimation of a conditional probability was first advanced by philosophers (e.g., Adams, 1981; Edgington, 1995; Stalnaker, 1970) and has lately started to be considered by psychologists as well (Anderson, 1995; Evans and Over, 2004; Oaksford & Chater, 1994). The interpretation of conditionals as a function of this conditional probability originally goes back to Ramsey (1990, p.247), who argued that people evaluate a conditional by "...hypothetically adding p to their stock of knowledge and arguing on this basis about q. They are fixing their degree of belief in q given p."

The psychological specification of the Ramsey test by Over and Evans (2004) describes the process of interpreting a conditional sentence as follows. If people evaluate the credibility of a statement such as "If you open the fridge, then the light inside comes on", they assume in a first step that the antecedent p of the conditional in question is true. For the example above this would mean to focus on hypothetical cases of opening a fridge. In the second step people count all cases in which *p* and the consequent *q* are both true (opening the fridge and light on) and compare them with all cases of *p* and $\neg q$ (opening the fridge and light off). Translating this procedure into mathematical terms results in the conditional probability P(q|p) or P(light on| open fridge). If the number of pq cases exceeds the number of p$\neg$q cases, such that P(q|p) > 0,5, people are willing to believe the conditional (Over & Evans, 2003). Situations were p is not true (fridge stays closed) do not play a role for the believability of the above statement in this account. According to the hypothetical nature of the thought experiment involving these two steps, Evans and Over (2004) call this interpretation the *suppositional conditional*.

Extensive testing of this probabilistic view has first been provided by Evans et al. (2003) and Oberauer and Wilhelm (2003). Both research groups have independently designed a task that has come to be known as the probabilistic

truth table task. In this task participants are provided with a conditional statement and explicit frequency information about the four cases of the conditional's truth table, that is the conjunctions of *pq, p¬q, ¬pq*, and *¬p¬q*. They are then asked to evaluate their belief in the conditional statement considering the frequency information given. The initial studies with the probabilistic truth table task use basic conditionals, in a sense that no former knowledge might conflict with the frequencies assigned to the truth table task. An example for a task using a basic conditional with arbitrary content is given in Box 1.

Box 1: Probabilistic truth table task

Engineers from Earth try to benefit from the advanced technology displayed in air traffic on Noxus. Therefore they categorize the different flying objects that can be observed in the air traffic. Of the 2000 flying objects that were observed within the last 3 month the following records exist:

900 flying objects had invisible wings and more than two jet propulsions
100 flying objects had invisible wings and not more than two jet propulsions
500 flying objects had no invisible wings and more than two jet propulsions
500 flying objects had no invisible wings and not more than two jet propulsions

An expert from Earth claims that:

"If the flying object has invisible wings, then it has more than two jet propulsions."

The task in all of these studies is to rate how likely it is that a person stating the conditional sentence is right. For conditions like the one presented in Box 1, people assign the conditional a very high believability, as the conditional probability P(q|p) is 0.9 in this example. Replicated with different content (arbitrary cover stories, deck of playing cards), different presenting techniques (presentation of lists with explicit frequencies about items, natural sampling of all items) and different overall number of items (ranging from 100 – 2000) it could be shown, that it was always the conditional probability P(q|p) that had the largest influence on people's belief in the conditional (e.g. Evans et al., 2003; Oberauer, Geiger, Fischer, & Weidenfeld, in press; Oberauer & Wilhelm, 2003, Weidenfeld, Oberauer & Hörnig, 2005).

Over, Hadjichristidis, Evans, Handley and Sloman (in press) used a different methodology avoiding the critical representation of explicit frequencies. They had people assessing the probabilities of the four conjunctions of the truth table case for everyday conditionals, e.g. "If global warming continues, then London will be flooded" and the probability that the statement as a whole was true (or false respectively). They replicated the findings that P(q|p) is the strongest predictor for people's belief in a conditional statement with probability measures that were not based on explicit frequencies.

1.2.2 Reasoning from conditionals

In the filed of conditional inferences there are numerous approaches that explain the way of how people draw inferences with the help of subjective probabilities (e.g. Oaksford, Larkin & Chater, 2000, Verschueren, Schaeken & d`Ydevalle, 2005, Evans & Over, 2004). Three of the approaches will be outlined here.

Suppositional account by Evans & Over

When it comes to conditional reasoning, Evans and Over (2004) explain the endorsement of MP and MT in relation to the believability of the conditional, that is based on the conditional probability P(q|p) derived by running a Ramsey test. Specifically, MP can be directly inferred from running a Ramsey test on the conditional: suppose, p is true, which happens to be the minor premise of MP, the question: "how likely is it that q is true", directly translates into reading off the probability of the MP inference. MT on the other hand has to be derived by a suppositional instance of *reductio ad absurdum* in several steps: 1) supposing that p is true, 2) inferring from it via MP that q should be true, 3) realizing the contradiction of that conclusion with the actual minor premise ($\neg$q), and 4) concluding from this apparent contradiction that the initial supposition of "p" must be false. Obviously there is a potential of getting lost in the process and therefore concluding that nothing follows from the minor premise $\neg$q. Taken together, the theory of Evans and Over implies that both MP and MT inferences should be affected by people's degree of belief in the conditional premise, which in turn depends on its subjective conditional probability P(q|p), although for MT this is true to a lesser extent due to the additional processes described above.

This hypothesis is supported from studies showing that the higher the sufficiency of a conditional is, the more willingly do people accept the inferences of MP and MT drawn from it (e.g. Cummins, Lubart, Alksnis, & Rist, 1991; Liu, Lo, & Wu, 1996). Sufficiency of a conditional is the extent to which the occurrence of p alone guarantees the occurrence of q. Since both degree of belief in a conditional and perceived sufficiency of that conditional are based on the same parameter, P(q|p), they are strongly correlated and their effects on MP and MT should be comparable.

Regarding the so-called fallacies AC and DA, Evans & Over (2004) give only a vague explanation for why people endorse them even though they are not logically valid. They hypothesize that people might pragmatically add the converse and inverse conditional from the original conditional, that is "if q, then p" (converse) and "if not-p, then not q" (inverse) and draw AC and DA respectively, with a Modus Ponens inference from these invited additional premises.

Dual process account by Verschueren, Schaeken and d`Ydevalle

Verschueren, Schaeken and d'Ydevalle (2005) give a more elaborate account of the acceptance of conditional inferences in their dual process specification of conditional reasoning. They abandon the distinction between logically valid and logically invalid inferences and claim that the well justified endorsement of all four inferences just depends on different subjective probabilities (called 'likelihoods' by Verschueren et al.). For MP and MT this is, as shown above, the sufficiency of the conditional derived from P(q|p). Unlike Evans and Over, they claim that the endorsement of AC and DA depends on the necessity of the conditional. That is, the more necessary the antecedent p (here: fridge open) is for the consequent (here: light on), the higher the conditional probability of the antecedent p, given the consequent q, P(p|q) turns out and the more justified it is to endorse AC as well as DA. In our example, the antecedent (fridge open) is actually highly sufficient (in a well functioning fridge it is the one thing that surely causes the consequent) *and* highly necessary (without it, the consequent is not likely to occur), so from this conditional it seems sensible to derive all four conclusions.

Sufficiency and necessity of a conditional are assumed to be derived by a quick, heuristic estimation process that constitutes the first of Verschueren et al.'s two processes and thus the theory conforms to other dual process theories of cognition (Evans, 2003; Evans, in press; Sloman, 1996; Stanovich & West, 2000). A second process consists of searching for counterexamples that potentially falsify the putative conclusions derived by the fast heuristic process. This process is assumed to be analytic in nature, to take longer and yield categorical judgments. So if people can think of a counterexample to the conclusion, they are assumed to reject the conclusion and if they cannot, then they should accept the conclusion, given high enough probabilities. Counterexamples for MP and MT are called *disablers* (Byrne, 1989) and consist of p¬q cases (here: fridge open, light off). If people come up with the possibility that, for example, the bulb might be broken or there is an electricity failure, they might reject either of the inferences. On the other hand, counterexamples for AC and DA consist of *alternative causes*, or *¬pq* cases. If people can think of any (rather unlikely) circumstances that make the light go on in the fridge without the door being opened, they should reject AC and DA. General support for the dual process idea comes from Weidenfeld et al. (2005). In this study the authors found evidence for two distinct processes that affect conditional inferences akin to the processes described by Verschueren et al. One of the downside of Verschueren et al.'s account is that it has little to say about why the inferences containing negations are usually endorsed less often than their positive counterparts (MT vs. MP and, although somewhat less clear, DA vs. AC respectively, Evans & Over 2004)

Probabilistic Account by Oaksford & Chater

Oaksford and Chater (2001) explain the acceptance of the four inference patterns with the subjective conditional probability of the conclusion, given the minor premise of each inference. This approach largely ignores the believability of the major conditional premise and focuses directly on the conditional probability of the conclusion. The authors present a mathematical model with three parameters, P(p), P(q) and an exception parameter ε. From these parameters they can derive the probabilities of the conclusion, given the minor premise, for the four basic inferences (see Table 2). For our example sentence, the results of these conditional probabilities are as follows:

MP: P(q|p) = P(light on | fridge open)

AC: P(p|q) = P (fridge open | light on)

DA: P(¬q|¬p) = P (light off | fridge closed)

MT: P(¬p|¬q) = P (fridge closed | light off)

Hence, every inference in Oaksford and Chater's model relies on a different conditional probability, whereas in Verschueren et al.'s conceptualization MP and MT rely on the same probability, that is, P(q|p), and AC and DA rely jointly on another probability, namely P(p|q). It follows that the two accounts make agreeing predictions on the positive inferences MP and AC and diverge in their predictions for the inferences containing negations, MT and DA. There has been contradicting evidence on this issue and further testing of it will be presented in Chapter 3.

One of the models strength is, that it can accommodate for the findings that MP acceptance is always higher than MT acceptance. Given certain plausible conditions (e.g. the occurrence of exceptions, Oaksford & Chater, 2001), the model always predicts higher MP than MT acceptance. However, the precise predictions regarding the acceptance of all four inferences have not yet been directly tested, but only been indirectly supported via predictions regarding the model's three basic parameters' behaviour (Oaksford, Larkin and Chater, 2001). This procedure confirming some of the model's predictions resulted in an unreasonably high exceptions parameter that casts doubt on the model's viability.

Further criticism has been put forward by Oberauer, Weidenfeld and Hörnig (2004) who failed to replicate the model's main prediction, that a conclusion should be drawn depending on its prior probability. In a comparative approach by Oberauer (2006b) including models of conditional reasoning based on four different theories, two versions based on Oaksford and Chater's probabilistic model came last in fitting two large data sets and other characteristics, as high number of free parameters. A more general downside of Oaksford & Chater's

model is that the major premise of the argument, that is the conditional "if p then q", plays no direct role for the acceptance of the inferences and thus the theory has little to say about how people evaluate and interpret conditional sentences to start with.

1.3 Outline of this dissertation

1.3.1 Summary

The present situation of research on conditional reasoning might be one of a paradigmatic change (Kuhn, 1962). Over the last 20 years, the mental model theory has become the most prominent theory of conditional reasoning. One of its big achievements is that it offers a coherent framework in which explanations for the interpretation *and* the drawing of conclusions from conditionals can be fitted. With its amendments as the idea of initial vs. fleshed out models, or the proportional belief in conditional statements according to the ratio of true models to all models, it could accommodate a lot of experimental findings with tasks containing conditionals, e.g. the notorious "defective truth table" in a classic truth table task.

However, a body of empirical evidence that it struggles to explain, is the findings on the interpretation of conditional sentences obtained with a probabilistic version of the truth table task. It has been repeatedly shown using different varieties of task presentation and content (Evans et al., 2003; 2005; Oberauer, Geiger, Fischer, & Weidenfeld, in press; Oberauer & Wilhelm, 2003, Over et al. in press), that people assign a conditional sentence a believability according to P(q|p). Even the probabilistic amendment of the mental model theory cannot explain why this should be the case. If people represented an initial *pq* model of the conditional sentence, they should base their believability judgments on P(pq) (as actually a minority of people does in all the aforementioned studies). If people based their believability judgments on all fleshed out models, their judgments should reflect a probability of 1-P(p¬q), which is basically never observed, same studies as above). So in the field of interpreting a conditional, a lot of evidence has been accumulated against the predictions of the MMT and in favor of the suppositional account.

The suppositional account of conditionals has emerged as part of a family of theories that have one feature in common: they all explain the interpretation of conditionals *or* reasoning from conditional premises with the use of conditional probabilities, and thereby form a promising alternative to the MMT. They also have several shortcomings though. First, none of them explains the interpretation of conditionals *and* the reasoning from these conditionals within one coherent framework as the mental model theory does. Second, whereas

they all may more or less agree on how the conditionals are interpreted and thus perceived as believable or not, they disagree to quite an extent on how people reason from conditional premises. The most precisely formulated probabilistic theories in reasoning are those of Verschueren et al. and Oaksford et al. that make partly contradicting predictions. On the other hand, whereas Evans and Over have summoned compelling evidence for their suppositional account of the interpretation of conditionals they are not very precise (yet) of what a suppositional reasoning account would look like for all inferences and have yet to present evidence for a suppositional reasoning process.

To summarize the state of research on conditionals: on the one hand there is a theory (of mental models) that fails to explain a considerable body of evidence on a probabilistic interpretation of conditionals, but has a high explanatory power on specific phenomena concerning conditional inferences (e.g. the lower endorsement rates for MT), is very general (explains a lot of reasoning phenomena also outside of conditional reasoning) and coherent (based on one main concept), although not very parsimonious (see amendments, Johnson-Laird & Byrne, 2002). On the other hand we have a family of (probabilistic) theories that compellingly converge on the interpretation of conditionals, but have yet to establish a convincing and primarily coherent account of conditional reasoning. How this situation is further clarified and brought to the conclusion that the probabilistic theory of conditionals has the potential to overcome its biggest caveat and thus to become a serious competitor to the MMT will be clarified in the Outlook.

1.3.2 Outlook

To explore the potential of a general probabilistic approach on conditionals I will first focus on the interpretation of conditionals and what the probabilistic theories have come to explain in this domain. Since the main influence of the conditional probability P(q|p) on the interpretation of the conditional has been widely replicated, instead of testing probabilistic approaches against the MMT one more time, I will test the suppositional account against a further refined version of it in the first empirical chapter (Chapter 2). The idea of a refined Ramsey test put forward by Oberauer (submitted) combines findings of the general research on probability judgments with the findings on conditional probabilities in conditional reasoning. This result of a similarity sensitive version of the Ramsey test is tested against a Ramsey test purely operating on frequencies in Chapter 2.

As a next step after the interpretation of conditionals, any theory on conditionals has to explain how inferences are drawn from them. There is no convincing probabilistic account yet that can explain the whole range of empirical

findings in conditional inferences. In the next empirical part of this thesis I will therefore examine different probabilistic approaches on inferential reasoning. The existing accounts make different assumptions on how inferences are evaluated and on which probability they are supposedly based. These different predictions derived from probabilistic accounts of reasoning are tested against each other in Chapter 3. Solving the apparent contradictions between different probabilistic approaches on reasoning is an important step towards generalising the suppositional account of conditionals to inference evaluation from conditionals.

The approach in this dissertation of testing and integrating different probabilistic accounts on interpreting conditionals and drawing inferences from them works towards a general comprehensive probabilistic framework on conditionals. Nevertheless, since mainly the methodological framework introduced by proponents of probabilistic approaches is used in the first parts of this dissertation, it could be argued that predictions from MMT can not be fairly considered. Toward this end, the last part of this thesis compares different types of information on counterexamples used by proponents of either theory respectively. In Chapter 4, frequency information on exceptions (probabilistic accounts) is compared to information on counterexamples (here: disablers, MMT). Results on believability judgments *and* inference tasks show that judgments are made relying mainly on probabilistic information, without the use of an analytic process validating them, which should have been the case, if the MMT was correct.

Short overview over the remaining chapters

Chapter 2: The probabilistic approach of the interpretation of conditionals following the Ramsey test procedure is tested against a refined version of the Ramsey test. Instead of considering frequency information only to derive probability judgments, this refined version (called Ramsey ProbEx) incorporates similarity information as well.

Chapter 3: The three different probabilistic approaches on conditional reasoning (Oaksford et al., 2001; Verschueren et al., 2005 and Evans and Over, 2004) are tested against each other to resolve the contradictory predictions. Stemming from these results, a new, probabilistic 2- stage model of conditional inferences is proposed.

Chapter 4: Different types of information regarding counterexamples on conditional rules that are usually used in experiments run by proponents of either the MMT or the probabilistic accounts are tested against each other. Results on two sets of tasks, that is interpretation of the conditional and reasoning from it, have been obtained. A probabilistic estimation process based on frequency information is used to explain results on all tasks.

Chapter 5 will give a short summary of all the results obtained in this dissertation and outline the conclusion that can be drawn from it for a probabilistic framework on conditionals.

Summarizing, this dissertation will give evidence for why it seems justified to draw an optimistic picture of a probabilistic framework on conditionals. Not only can it explain why people heavily draw on P(q|p) when interpreting a conditional sentences, there also seems to be good chances of an probabilistic explanation of how people draw inferences from conditionals. Taken together it seems that people have indeed a very human way (in reference to the opening quote) of thinking in eventualities as opposed to former theories of reasoning that still heavily draw on concepts like logical validity or truth.

Chapter 2: Evaluating a conditional premise

2.1 Introduction

Chapter 2 starts out comparing two different variants of a probabilistic approach on the interpretation of conditional sentences. The suppositional approach on conditionals assumes that people evaluate their belief in a conditional rule with a procedure called the Ramsey test (Evans & Over, 2004). According to the Ramsey test people suppose that p is the case in a mental simulation and relate the *pq* cases to $p \rightarrow q$ cases within this suppositional frame (Evans, Over, & Handley, 2005; Over & Evans, 2003). If the resulting ratio surpasses a certain threshold (at least 0.5), they tend to belief the conditional sentences. When evaluating the belief in a rule as "if you open the fridge, then a light inside goes on", people thus suppose p (in which the fridge door is opened), and compare the number of rule-confirming cases (cases where the fridge was opened and the light went on) to exceptions (fridge opened and the light did not go on). Accordingly, people process frequency information about confirming and rule violating instances when facing to evaluate a conditional. The effect of explicit frequency information about *pq* and $p \rightarrow q$ cases has been shown in many studies so far (Evans, Handley & Over, 2003; Oberauer & Wilhelm, 2003).

Consider a conditional sentence as: "If an animal is a bird, then it can fly". According to a strict version of the Ramsey test, people derive their belief in the sentence by relating all birds that fly to all birds they know to be not able to fly in a simple frequentist manner. With an everyday conditional like the example above they do this by comparing the relative frequencies they can derive from long term memory.

From the research tradition of categorization and induction (Lopez, Gelman, Gutheil, & Smith, 1992; Osherson, Smith, Wilkie, Lopez, & Shafir, 1990; Rips, 1975; Sloman, 1998) comes the notion, that for strength of inductive arguments not only frequency information, but also the similarity of premise and conclusion category plays an important role. A key finding in this research is that "Robins have a property P, therefore sparrows have a property P" is usually judged to be a stronger argument than "Robins have a property P, therefore geese have a property P" (Heit, 1997; Osherson et al., 1990). The effect is explained with reference to the concept of similarity: people are more willing to transfer a property of robins to the more similar category of sparrows than to the less similar category of geese. If argument strength benefits from the similarity of categories or situations mentioned in the premise and in the conclusion in inductive reasoning, it seems justified to suspect that similarity of exemplars may also play a role in the evaluation of the convincingness of a conditional sentence. The effect of additional features like similarity has recently started to be considered also in the related field of subjective probability judgments.

This research is focused on how exactly people arrive at a probability judgment. Juslin and Persson (2002) have proposed a model they called "ProbEx", from "probabilities through exemplars", in which they claim that people derive conditional probability judgments through the consideration of exemplars reasonably similar to a given probe exemplar. Their model assumes that exemplars are taken into account according to their relative similarity to the feature patterns of previously learned exemplars. The more similar a specific exemplar is to a probe exemplar, the higher is its multiplicative weight with which it is considered. For example, if the symptoms of a former patient strongly resemble the symptom pattern of a current patient, the doctor is thought to take this patient more into account than another one with fewer of the same symptoms to derive a diagnose.

This way, the model is sensitive to the frequency of exemplars in a certain category (as proposed by the Ramsey test) as well as the similarity of those exemplars to a certain probe exemplar. The question used by Nilsson et al. (2005) concerned the conditional probability of belonging to a specific category, given a specific feature pattern. They put their model into test against other explanations of probability judgments as the representative heuristic (Kahneman & Frederick, 2002; Kahneman & Tversky, 1972) or cue validity (Gigerenzer, Hoffrage, & Kleinbölting, 1991) of probability judgments with categorization tasks (Nilsson, Olsson, & Juslin, 2005). In this study, people had to judge how likely it was that an exemplar with a specific feature pattern belonged to a specific category A or B after they learned about the category membership of many exemplars with different feature patterns. In several experiments ProbEx had a high predictive power for the probability judgments (Nilsson et al., 2005).

For conditional statements such as "If the patient has symptom pattern X, then he has the flu", it was widely shown that their believability is evaluated by the conditional probability of the consequent (e.g. having the flu), given the antecedent (e.g. symptom pattern X: fever, headache, muscle pain, fatigue). The findings of the probabilistic truth table task (see Chapter 1) relates the believability judgments of conditionals closely to the conditional probability judgment task used by Juslin et al. (2003; Nilsson et al., 2005). If people take more similar exemplars more strongly into account for the latter judgments, they might quite likely do the same when evaluating a conditional sentence. An example for the conditional probability judgment task Juslin et al. used would be:

> *Consider a patient with symptom pattern X. How likely is it, that he has the flu?*

The corresponding evaluation of the conditional in a probabilistic truth table task would look like this:

Claim: If the patient has symptom pattern X, then he has the flu. How likely do you think this sentence is true?

The question of whether similar processes are involved when deciding about which situations should be taken into account in the evaluation of a conditional premise with an antecedent p, has been addressed by Oberauer (2006a). He raises the theoretical argument, that in a suppositional account of conditionals it has to be specified how reasonably close the p-situations should be to the situation mentioned in the antecedent, on which the mental simulation of "is q the case?" is run. Oberauer calls this specification the "relevant set of p-cases". Evaluating the conditional above, doctors should refrain from thinking of patients that share no symptoms with our patient X, since they are not part of the relevant set. Oberauer suggests that retrieval of relevant exemplars from long term memory proceeds until a stopping criterion is reached, which in our case might be "shows at least three symptoms of pattern X", but could vary from doctor to doctor. Within the boundaries of retrieved exemplars, he expects the exemplars with a higher similarity to the mentioned p-cases ("sharing more symptoms with the pattern X patient" in our conditional) to be weighted more heavily, comparable to the idea specified in ProbEx for conditional probability judgments.

The formal specification by Oberauer (2006a) is presented in Equation 1. To compute P(q|p) people retrieve a certain number of situations s_i (here: exemplars) until a stopping criterion is reached. This criterion is based on the similarity S of s_i to the present situation s_0 in which the antecedent p holds. Each of the retrieved s_i in the relevant set (denominator) is then evaluated in regard of whether the consequent q holds in this situation (or for this exemplar) or not and Q(si) is either set to 1 or 0 (numerator).

Equation 1:

$$P(q \mid p) = \frac{\phi P(q) + \sum_{i=\alpha}^{N} S(s_0 + p, s_i) Q(s_i)}{\phi + \sum_{i=\alpha}^{N} S(s_0 + p, s_i)} + \varepsilon$$

ϕ is a dampening parameter for the effect of retrieved exemplars when N is small, ε is a normally distributed error term and P(q) is taken for the prior probability. α is set to 0 for conditionals with a true antecedent and to 1 for conditionals with a false or uncertain antecedent, to ensure that the present situation s_0 is only considered when it is one in which p holds.

For the example with the antecedent "if the patient has symptom pattern X" the present situation s_0 in which the conditional is uttered draws our attention to patient who have a reasonable similar symptom pattern to patient X. In a first step it is assumed that a doctor starts to retrieve a number of patients (s_i) that

according to her stopping criterion showed e.g. at least three symptoms of pattern X. Of all patients considered (e.g. ones with a fever, headache and muscle pain or all four symptoms), the weighted ratio of those who satisfy the consequent ("had influenza") is computed. The higher the similarity S (number of shared symptoms) is, the higher the weight of the exemplar (patient) is.

More specifically, as can be seen regarding the equation, similarity works in two ways: similar exemplars are in a first step more likely to be retrieved and included in the relevant set, and in a second step among those retrieved they are the ones that are weighted more.

To test the idea of a similarity graded evaluation of conditional premises I conducted two experiments. I used categories that had a clear defined boundary (a disadvantage to investigate subjective stopping criterion), but therefore allowed precise manipulation of similarity measures, the main concept under investigation here. The similarity of given exemplars to the probe mentioned in the conditional was varied by using typicality measures. The underlying assumption for this procedure is that people use the prototype of a given probe as retrieval cue and retrieve items more similar to it with higher probability. The less similar a specific exemplar is to the prototypical category member, the less typical this exemplar is for its category. According to this rationale, *similarity* (to prototypical member) and *typicality* (for the category) refer to the same characteristic of the exemplar in this chapter. In all our materials (see Appendix 2.1), we used unspecified category members in the antecedent of the conditional as a probe (e.g. a bird, fish or vegetable) assuming that people think of a prototypical category member when reading the conditional and relate all subsequent specific exemplars (e.g. penguin, sharks, artichokes) to that prototypical member.

Experiment 2.1 contrasts a set of typical exemplars with atypical exemplars over a range of different categories within a factorial design, Experiment 2.2. takes a closer look at the grading of similarity over a whole range of differently similar exemplars belonging to one category ("birds").

2.2 Experiment 2.1: typicality of exemplars

In Experiment 2.1 we varied the typicality of the p¬q exemplars belonging to the category mentioned in conditional rule. These exceptional cases were either very typical, very untypical or not further specified exemplars of the category mentioned in the antecedent of the conditional, yielding a factor *typicality of exceptions* with 3 levels.

The typicality variation of counterexamples was embedded in a variation of the original form of the probabilistic truth table task, presenting explicit frequencies in the cover stories with either a high or low conditional probability of p, given q, $P(q|p)$. The 2 x 3 design is summarized in Table 3. We will refer to

the six conditions as HL, HH, HU; LL, LH, LU, with the first letter referring to few versus many exceptions and the second letter referring to the typicality of the exceptional category member as either low, high or unspecified.

Table 3: Experimental manipulation in Experiment 2.1.

Conditions:	**HL**	**HH**	**HU**	**LL**	**LH**	**LU**
P(q|p)	high	high	high	low	low	low
Typicality of exceptions	low	high	un-specified	low	high	un-specified

Legend: First letter of the condition code represents P(q|p), the second letter represents the typicality of p¬q cases.

To emphasize the information on the exceptional cases, we used a reduced array of frequency information introducing the number of pq and exceptional p¬q cases only. According to the plain version of the Ramsey test, the typicality of the exceptional cases should not matter for the evaluation of the conditional premise and should subsequently also have no influence on the inference tasks. If probability ratings are indeed graded according to perceived similarity of the situation in question we should observe an effect of typicality such that more typical p¬q cases should lead to a lower belief in the conditional and subsequently to a lower acceptance of MP and MT. Since p¬q cases are only indirectly relevant for the inference of AC and DA, manipulation of those cases should not affect these inferences.

2.2.1 Method

Participants

Participants were 40 high school and University students from the University of Potsdam (age range: 18-27 years). Order of tasks was varied between subjects, yielding a factor *task order* with a group of 20 subjects that gave probabilistic judgment first and then solved reasoning tasks and 20 subjects working on the tasks in reversed order.

Material and Procedure

The experiment was a computerized study realizing the design in Table 1 within subjects. For the probabilistic judgments, participants received 12 tasks presented on a monitor, 2 for each condition in Table 1. For each task we designed a cover story describing differently typical exemplars of a certain category. A full list of the categories and their members used in Experiment 2.1 is listed in Appendix 2.1. An electronic version of the experiment (as all other experiments conducted for this thesis) is available on CD-Rom on the electronic appendix of this thesis located on the inside of the back cover.

Every site presented a short story about a fictional situation concerning one of 12 different categories. The cover stories introduced a conditional statement made by an "expert" of the matter and gave information about the total number of p cases and the relative distribution of p¬q to pq cases. The overall number of cases varied between 350 and 1250, depending on the cover story. The conditional probability P(q|p) was either .95 or .55, equalling a proportion of 5% or 45% p¬q cases respectively.

The exceptional cases either concerned very typical, atypical or non-specified members of the according category and were always described as lacking a specific feature (that made them exceptional). The features specified in the consequent of the conditional always concerned an arbitrary or even fictional characteristic and have been used in studies of inductive reasoning (c.f. 'blank predicate', Lopez et al., 1992; Osherson et al., 1990; Sloman, 1998). These predicates are supposed to not draw on specific knowledge people might have.

Category members were taken from the handbook of German word norms (Mannhaupt, 1994), depending on their generation frequency in a category generation task. Here is an example of a cover story (for the condition HL):

> *Researchers in a biology institute examine the blood supply of animals. The researchers assume that the rule holds.*
>
> *"If the animal is a bird, then it has pulnar arteries"*
>
> *In the last week they examined 750 birds, among them penguins, black birds, raven, partridges and eagles. Of the 750 birds, 413 had pulnar arteries and 337 did not have pulnar arteries.*
>
> *The 337 birds that did not have pulnar arteries were penguins and partridges."*[1]

For the two conditions with unspecified exceptional cases (LU and HU) the last sentences was omitted. Subjects were asked to rate the probability that the expert was right on a scale from 0 ("absolutely impossible") to 100 ("absolutely certain").

Either following or preceding the probabilistic judgments of all the conditional statements, participants had to solve the four types of inference tasks for each cover story. Random order of cover stories and conditions was the same as in the probability-judgment part of the experiment. The cover stories and conditionals were presented again, this time followed by MP, MT, DA and AC on one of four specific cases of the sample. Order of the inference tasks were randomised anew for each cover story. Here is an example for Modus Ponens:

---

[1] We changed the predicate of "ulnar arteries" (as used by e.g. Lopez et al. (1992) into "pulnar arteries" to make them comparable to our other fictional predicates in the remaining 11 cover stories.

Expert's statement: "If the animal is a bird, then it has pulnar arteries".
1st case: This animal is a bird.
Conclusion: It has pulnar arteries.

Participants had to rate their confidence in the conclusion on a 6 point scale ranging from "certain that I *can* draw the conclusion" to "certain that I *cannot* draw the conclusion", as introduced by Cummins et al. (1991) and since used in many studies investigating inference tasks (e.g. DeNeys, Schaeken, & D'Ydewalle, 2003; Oaksford, Chater, & Larkin, 2000). No emphasis on the logical validity of the conclusion was made.

2.2.2 Results

Since the factor *task order* did not have a general effect nor interacted with any of the other factors, all data were collapsed over this factor and submitted to a 2 x 3 ANOVA with *P(q|p)* (high-low) and *typicality of exceptions* (low-high-unspecified) as factors. In this and all other subsequent experiments, the 6 answer options for the inference tasks were coded from +5 ("certain that I can draw this conclusion") to -5 ("certain that I cannot draw this conclusion") in steps of 2 to generate equal numerical distances between the six options.

Probability of the conditional

Figure 1[2] shows the probability ratings of the conditional P(cond). The ratings were largely influenced by *P(q|p)*, $F(1,39) = 178.5$, $p<0.001$, $\eta_p^2 = 0.82$. The typicality of exceptions also exerted an overall effect, even if small, $F(1,38) = 4.0$, $p<0.05$, $\eta_p^2 = 0.09$. Planned contrasts revealed a difference between typical and unspecified exceptions, $F(1,39) = 5.4$, $p<0.05$, $\eta_p^2=0.12$ and between typical and atypical exceptions $F(1,39) = 3.7$, $p=0.06$, $\eta_p^2=0.09$ (one-tailed testing $p=0.03$). This effect of typicality of exceptions was larger when P(q|p) was high, as the interaction between number and typicality of exceptions revealed $F(1,38) = 3.6$, $p<0.05$, $\eta_p^2 = 0.08$.

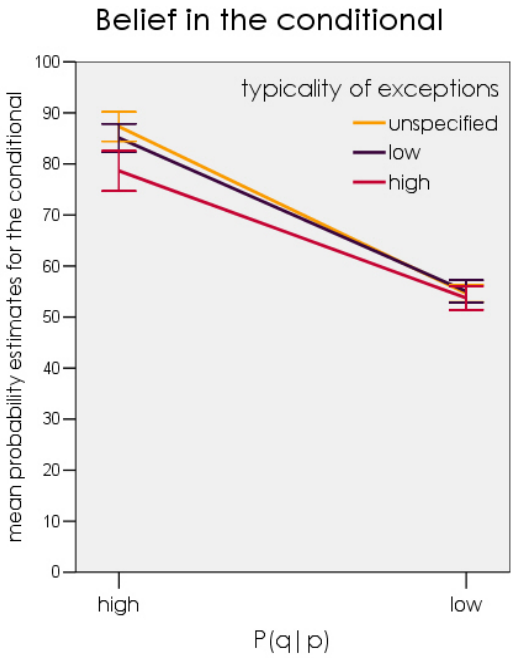

Figure 1: Believability of the conditional

Probability estimates (on a scale from 0 to 100) for the conditional. P(q|p) is grouped on the x-axis, typicality is indicated by different patterns.

[2] Error bars in this figure and in all consecutive figures in this dissertation indicate the 95% confidence interval around the mean. Within-subject design confidence intervals have been corrected according to the Bakeman-McArthur procedure for standard errors.

Reasoning Tasks

Figure 2 shows participants´ confidence in the four inference tasks. P(q|p) had a medium effect on Modus Ponens, $F(1,39) = 30.0$, $p<0.001$, $\eta_p^2 = 0.44$ and a smaller effect on Modus Tollens, $F(1,39) = 13.0$, $p<0.001$, $\eta_p^2 = 0.25$.

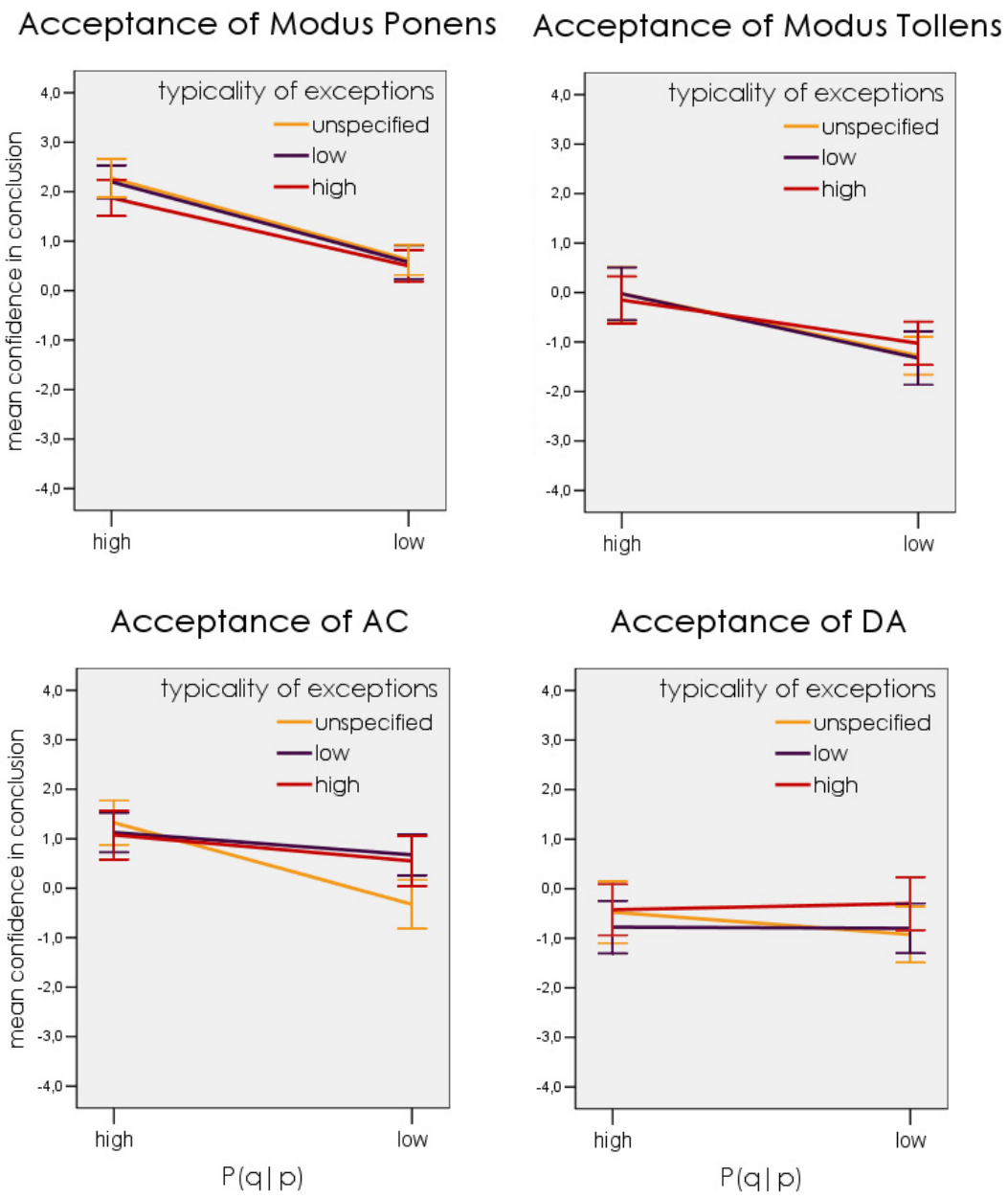

Figure 2: Acceptance of the inference task.
Confidence ratings in the conclusions (on a scale from -5 to +5).

Typicality of exceptions did not have an effect on these two inference forms, nor on any of the other inferences, all $F<1.2$.

P(q|p) did not have an effect on DA, F<1. For AC, there was a small overall effect of number of exceptions F(1,39) = 9.1, p<0.01, η_p^2= 0.18 and a trend towards an interaction with typicality, F(1,38) = 2.9, p=0.06, η_p^2= 0.06, such that number of exceptions only had an influence in the "unspecified" condition.

2.2.3 Discussion

By now corroborating a large number of studies (Evans et al., 2003; Oberauer, Geiger, Fischer, & Weidenfeld, in press; Oberauer & Wilhelm, 2003; Over et al., in press), results on the believability of the conditional were largely effected by the conditional probability. Additional to the overarching effect of P(q|p), we found a small effect of the typicality manipulation of exceptional p¬q cases. If exceptional exemplars were very typical to the category the conditional referred to, the conditional was judged less believable than if exceptions were highly atypical. This can be interpreted as first evidence for a similarity graded evaluation of conditional premises as postulated by an exemplar based theory of probability judgment. In the sense, that counterexamples were more or less typical to the category mentioned in the antecedent of the conditional (e.g. penguin and black birds for "If the animal is a bird, …") they were more or less similar to the situation mentioned in p (cf. Equation 1 from Oberauer, 2006). That more similar situations (here: more typical birds) were taken into account more heavily supports the idea of a similarity graded version of the Ramsey test.

Against the intuitive notion that non-specified exemplars (in our example: "birds") should convey a higher similarity to the general category in the antecedent than the atypical ones ("penguins"), the believability ratings just showed the opposite: non-specified exemplars led to similar believability judgments as atypical ones. This rather surprising result might be due to the exceptional nature of the exemplars in question. If we hear something about birds that lack a specific bird-like predicate, we might be prone to envision those birds as rather atypical ones. This explanation could be easily tested by varying the typicality of rule-confirming pq-cases in a further experiment. If the explanation is correct, probability judgments for the non-specified pq-cases should closely match the judgments for typical pq cases in this experimental setting.

Moreover, the effect of typicality was attenuated by the number of exceptions. The typicality of exceptional exemplars played a larger role when P(q|p) was high. This is explainable by the fact that if there are only a few exceptions to a rule, the specific characteristics of these exceptions might weigh more heavily compared to a situation where there are so many exceptions (and P(q|p) is so low), that the rule itself is not believable anymore.

For the inferences, an effect of number of exceptions were obtained for MP and MT, in a way that more exceptions lead to suppression of the inference acceptance, although the effects were smaller than for the believability of the conditional. More results on the effect of counterexamples to inferences from conditional premises will be reported in chapter 3. There it was also found, that effects of counterexamples were always smaller on the inference tasks than on the believability of the conditional ratings. Typicality of exceptions did not have an additional effect on the inference task, also pointing towards a more material-insensitive nature of the tasks. Effects of counterexamples in form of $p\neg q$ cases are not suggested for AC and DA by any theoretical stance. The effects obtained for AC are due to an exceptional low acceptance in the unspecified-many exceptions condition only and thus will not be discussed here further.

2.3 Experiment 2.2: graded similarity of exemplars

Experiment 2.1 revealed a general effect of similarity of judged situations (respectively exemplars) to the actual situation stated in the antecedent of the conditional. Coming from these results, Experiment 2.2 attempts to refine these results to a graded way of similarity weighting. For this undertaking I concentrated on one of the categories from Experiment 2.1 and specified further members regarding to their gradual similarity to a typical category member. The category of "birds" was chosen for its relative ease of testing a multitude of different members. To test the hypothesis of a similarity graded weighting of exemplars, a design slightly different from that in Experiment 1 was chosen. 16 different category members in 4 different similarity classes were identified in a pretest and then provided with a randomly assigned feature mentioned in the consequent of the conditional. To the extent that similarity to the antecedent specification plays a role in the evaluation of conditionals, more typical exemplars should be weighted gradually more for the evaluation of the conditional. To test this hypothesis, a design similar to the one of Nilsson, Olsson and Juslin (2005) was implemented, that varied the typicality of exemplars belonging to a specific category. The experiment was implemented with the software MOUSELAB (Willemsen & Johnson, 2004) that hides and displays information user driven and thus allows to track successive information acquisition. This additional feature allows to track which exemplars are considered relevant by the participant thus provide a clue for the stopping criterion employed. If an exemplar is considered completely irrelevant for the task, information about this exemplar should not even be retrieved.

2.3.1 Pretest

The material for Experiment 2.2 consisted of tables with information about 20 animals, 16 of which were more or less typical birds to a varying degree. To establish the typicality of the birds used, a pretest was run in German, in which 40 different birds were rated regarding to their similarity to a prototype bird. The survey was run in the internet (N=40, age range 21-58). The instructions were as follows:

How similar are the following kinds of birds to a typical bird?

Please give your answer on a scale from 0 ("not similar at all") to 6 ("very similar") by clicking on the according button."

There was also an answer option for the case that participants didn't know the bird. Results for the similarity ratings of the 16 birds chosen within 4 different similarity categories are shown in Appendix 2.2.

2.3.2 Method

Participants

Participants in the main experiment were 32 last-year high school students (age range: 17-20 years).

Material and Procedure

The experiment was a computer based study realizing the design in Table 2. The independent variable was the variation of feature values (coded as 0 respectively 1 in Table 2) across exemplars who were more or less typical birds. The assignment of a specific feature (see Appendix 2.4) to the according factor level was varied across participants according to a Latin square yielding 16 different combinations. Two participants were assigned to each specific feature-exemplar distribution combination, resulting in 32 participants. Each participant worked trough 16 tasks in random order, one for each factor level.

Table 4: Experimental manipulation in Experiment 2.2.

Factor levels:	Category 1 similarity=5.1 (e.g. sparrow) 4 exemplars	Category 2 Similarity=4.1 (e.g. eagle) 4 exemplars	Category 3: similarity=3.1 (e.g. duck) 4 exemplars	Category 4: similarity=2.1 (e.g. penguin) 4 exemplars	Category 5: Non birds (e.g. bat) 4 exemplars
1	1	1	1	1	1 0
2	1	1	1	0	1 0
3	1	1	0	1	1 0
4	1	0	1	1	1 0
5	1	1	0	0	1 0
6	1	0	1	0	1 0
7	1	0	0	1	1 0
8	1	0	0	0	1 0
9	0	1	1	1	1 0
10	0	1	1	0	1 0
11	0	1	0	1	1 0
12	0	0	1	1	1 0
13	0	1	0	0	1 0
14	0	0	1	0	1 0
15	0	0	0	1	1 0
16	0	0	0	0	1 0

Parameter value for a specific exemplar is coded "0", if the exemplar does not posses a specific feature and "1" if it does posses the feature. For the non-birds feature distribution was varied so that half of the exemplars possessed a feature in each trial.

The experimental material was designed with the software MOUSELAB (Willemsen & Johnson, 2004), a device that allows to investigate information acquisition by the participant (by e.g. clicking on a box). For each task, the feature information about the 20 animals was initially hidden in the cells of a 4 x 5 table in random arrangement labelled with the animals´ names (cf. Figure 3).

Information about whether any given animal possessed the specific feature or not, could be retrieved by clicking on the according box. Once the information was viewed it stayed uncovered until the next trial. This way, data can be analyzed regarding two different aspects: which of the animals are considered relevant at all, and if considered relevant, to what extent did they influence the probability judgments of the conditional.

A short cover story introduced a scientific context, according to which 20 different animals (16 of which were birds) were examined regarding 16 different features. Each task introduced one conditional sentence as an experts´ assumption about the parameter value of one of the 16 features in birds, e.g. "If it is a bird, then the animal has gotagan in its stomach." Participants were asked to rate the probability that the expert was right on a scale from 0 ("absolutely impossible") to 100 ("absolutely certain").

Figure 3: Original example task for Experiment 2.2.
Boxes show the information about the 20 animals. Once clicked on a box, the according information about the feature in question is displayed.

2.3.3 Results

Information acquisition

Figure 4 shows in percentages how often feature information about a specific exemplar was retrieved. Information about the non-birds was retrieved in less than 20% of instances, whereas information about the birds was retrieved between 50.6% (penguin) and 71.3% (pigeon) of times.

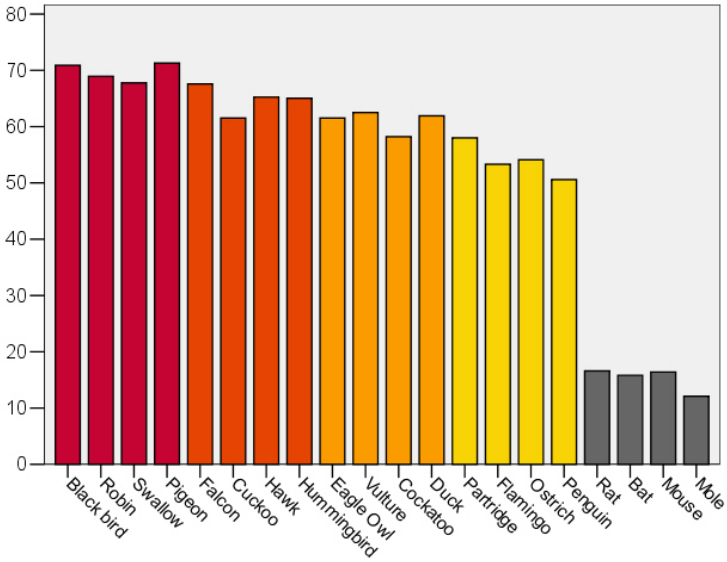

Figure 4: Mean information acquisition rates
Bars indicate the percentage of information retrieval about a specific exemplar. Exemplars are ordered according to their typicality rating in the pretest (Black bird = most typical bird, penguin = least similar bird).

A 1 x 4 ANOVA (4 typicality categories indicated by different colours) showed an overall tendency to consider more typical exemplars more often $F(3, 29) = 8.93$, $p<0.01$, $\eta_p^2 = 0.22$. Planned contrasts revealed that this difference already occurs between the first and second class, $F(1,31) = 4.2$, $p<0.05$, $\eta_p^2 = 0.12$, as well as all subsequent typicality classes (class 1 vs. 3: $F(1,31) = 5.2$, $p<0.05$, $\eta_p^2 = 0.14$, class 1 vs. 4: $F(1,31) = 12.1$, $p<0.001$, $\eta_p^2 = 0.28$).

Regression Analyses

To test the hypothesis, whether the similarity of an exemplar is taken into account when evaluating conditional premises referring to them we computed two different predictors:

a) predictor according to the Ramsey Test plain version (*RamseyPlain*)

$$RamseyPlain = \frac{\sum_i^{16} (F_i * IR_i)}{\sum_i^{16} IR_i}$$

with F= feature parameter value (0 or 1), IR = Information retrieved (0 or 1, for participants retrieving information about the exemplar in question or not) and a running index i for the 16 bird exemplars. For the test between the two different

versions of the Ramsey test, the feature distribution of the non-birds should be irrelevant.

b) Predictor according to a similarity graded Ramsey test (*RamseyProbEx*)

$$RamseyProbEx = \frac{\sum_{i}^{16}(F_i * IR_i * S_i)}{\sum_{i}^{16} IR_i}$$

The *RamseyProbEx* predictor only differs from the *RamseyPlain* predictor in an additional parameter S_i that is set to the typicality rating of the specific exemplar i and functions as a weight for this exemplar according to its similarity to the category mentioned in p.

For the linear regression all trials in which information for only 4 exemplars or less were viewed were excluded from the analysis. If people only viewed 4 or less animals, the likelihood that animals of different typicality categories were chosen is very low and people must have mainly guessed. From the 512 trials thus only 438 were included.

In Table 3 the results the two regressions are shown. Both predictors have a comparable high (or low) beta weight and explain 14% and 13% percent of the variance of the believability judgments of the conditional. The two predictors are highly correlated, r=.95, p<0.001. For a further analysis linear regression for the 16 trials on participant level were computed. A list of individual beta weights for all 32 subjects is shown in Appendix 2.5. A big individual difference was revealed with that analysis. Beta weights varied from -.96 to +.99. The beta weights were only significant for 15 of the participants, for another 16, the regression equation did not explain a significant portion of the dependent variable. Participant 24 never actively retrieved information for more than 4 exemplars and thus no regression parameters could be computed for that person.

Table 5: Results of the linear regression

P(conditional)		**Beta**	**t(437)**	**p**
	R^2 = .14			
	RamseyPlain	.37	8.38	.000
	R^2 = .13			
	RamseyProbEx	.36	8.19	.000
	R^2 = .11			
	P(Conjunction)	.33	7.20	.000

Table shows the beta weights of three predictors in a linear regression on P(conditional) = belief in the conditional.

As reported in the introduction section, a substantial number of studies (Evans et al., 2003; Oberauer et al., in press; Oberauer & Wilhelm, 2003) showed, that there are individual differences in interpreting conditional sentences and that a minority of people (app. 30%) interprets the conditional according to the conjunctive probability P(qp). Referring to these results, a third predictor P(conjunction) was computed:

c) Predictor according to the conjunctive probability (P(conjunction))

$$P(\text{conjunction}) = \frac{\sum_{i}^{16}(F_i * IR_i)}{20}$$

Again, this predictor did account for only 11% of the variance o believability judgments and more importantly did not explain the individual differences. The predictor was substantially correlated with the RamseyPlain predictor, $r=0.78$, $p<0.000$ ad with RamseyProbEx predictor, $r=.69$, $p<0.000$. Thus, participants that had unsubstantial beta weights in the two Ramsey version predictors also obtained unsubstantial beta weights in the conjunction predictor and participants with high beta weights on the two Ramsey version predictors obtained high conjunction beta weights in turn.

2.3.4 Discussion

The results of Experiment 2.2 can be discussed on two levels. Since the experiment was designed with a software tool that allows tracking information acquisition, on a first level, it could be analyzed which information people considered at all. On a second level it was analyzed to what extent they took the acquired information into account.

Concerning the acquired information, Experiment 2.2 revealed that exemplars were considered more often, the more typical they were for the category specified in the antecedent of the conditional ("birds"). Information about the most typical birds (category 1) was retrieved in 70% of all cases, information about the least typical birds (category 4) was retrieved in only 54% of all cases. Exemplars that were clearly not members of the category, thus the $\neg$p-cases, were only viewed in 15% of all cases.

These results first of all speak clearly for a suppositional account of conditional premise evaluation. According to this account, when facing a conditional sentence, people make a supposition about p being the case and then reason about q within this supposition. For our example, "if the animal is a bird, then…" participants concentrated on the information about the birds and

widely discarded information about the non-birds. Within the considered information about the bird exemplars, there was a clear trend of retrieving information about more typical exemplars more often, which speaks for a similarity graded version of the Ramsey test called 'RamseyProbEx' in chapter 2.3.

Regarding the question in how far the information about more typical exemplars was given priority in the probability judgments, the results are less clear. As the regression analyses showed, neither of the predictors (RamseyPlain or RamseyProbex) could outperform the other. This means that people might as well apply a similarity measure when deciding which exemplars to consider at all to evaluate a conditional premise (i.e. including the appropriate stopping criterion, c.f. Oberauer, 2006), but once an exemplar has made it into the relevant set of p-cases they are probably weighted the same.

From the results on the individual level (see Appendix 2.5) it must be concluded that almost half of the participants have based their judgment of the conditional on some unknown parameter or have simply guessed. This explanation must definitely be true for the 15% of all trials where people retrieved information about only 4 or less exemplars. The question why for 15 participants none of the two predictors could explain any of the variance remains unsolved. The explanation due to individual differences according to which some people evaluate the conditional via the conjunctive probability has to be ruled out as well. The 15 participants in question scored low beta-weights for the conjunctive predictors as well.

This brings us to possible caveats of the design used. The RamseyPlain predictor was substantially correlated with the conjunctive predictor. Considering the observed individual differences in evaluating conditional premises, the design should have been set up in a way that allows the distinction of these two predictions. More so, the RamseyPlain predictor correlated almost perfectly with the RamseyProbEx predictor, which makes a distinction in explanatory power factually impossible. The correlation of these two predictors depends on the amount of information retrieved. The less information is retrieved, the less the typicality weights multiplied with the retrieved bits of information can unfold their effect. In the experiment here, of all birds, information for only 62% was retrieved. This might have been too few for a possible distinction to be made between the *RamseyPlain* and the *RamseyProbEx* predictor.

2.4 Summary and conclusions of chapter 2

In Chapter 2 the issue on how people come to believe in a conditional sentence is addressed. The underlying theoretical approach that has been tested is based on the suppositional account of conditionals as proposed by Evans and Over (2004). A large set of studies (Evans, 2003; Oberauer et al., in press;

Oberauer & Wilhelm, 2003) has proven its main prediction, that probability judgments of the conditional are mainly based on the conditional probability of p, given q. Accordingly, predictions from the MMT have been widely disproved, and thus were neither tested once more nor will be discussed here.

In the experiments here, results on the believability of the conditional replicate the basic findings on the suppositional account of conditionals. The large effects of P(q|p) on he believability of the conditional were even obtained with the reduced array of frequency information used in Experiment 2.1. In Experiment 2.2, where information about all cases of the truth table was potentially available, approximately half of the sample also based their believability judgments on the conditional probability P(q|p), as reflected in large beta weights of the *RamseyPlain* predictor for 15 of the 32 participants. There is no obvious explanation for differing results on the other half of the sample that based their judgment on an unidentified process, which might be due to the difference in the material presentation. Since this result has not been replicated with using the MOUSELAB presentation version of the task so far, the differences will not be discussed further.

The main goal in Chapter 2 was to refine the notion of a suppositional account of conditionals based on the Ramsey test to a similarity weighted variant. The idea, first proposed by Oberauer (2006a) comprises the notion of a relevant set of p-situations that is taken into account when evaluating an indicative conditional of the form "if p, then q". If people base their judgments on relevant p-cases they retrieve from memory or simulate in hypothetical thought, the question arises, which situations are deemed sufficiently similar to be considered as relevant. The second question concerns the importance each relevant situation is granted. The idea of a similarity graded Ramsey test would claim that the more similar a situation to the antecedent p is, the stronger it will be considered for the evaluation of the conditional.

In our examples, the antecedent of the conditional always referred to members of a specific category (e.g. "birds" in Experiment 2.2). To answer the first question, we can draw on the results on information acquisition in Experiment 2.2. As people almost exclusively retrieved information on the p-cases, it can be argued, that the relevant set for judging conditionals about birds in the experimental material simply consisted of all the birds in the sample. In the sense that the conditionals in question referred to a category with clear boundaries, these results might seem rather trivial. It could be considered as an alternative way of showing that people have a defective truth table downrightly ignoring information about $\neg$p cases.

An interesting question would be the definition for the relevant set of p-cases in conditionals with a less clear antecedent as e.g. "If the weather is better tomorrow, then my friend will come along for a hike". We might know that our

friend despises rain and would never go for a hike in bad weather and that today is a rainy day. The antecedent of the conditional makes it less clear, which situations should be considered as relevant enough to be taken into account. Considering the principle of closeness, we should only be considering days, where anything else but the weather is kept constant to today, since considering a day with better weather in e.g. another country might not help evaluating the likelihood of the friend coming along for a hike. As Oberauer (2006) points out with reference to Bennett (2003), this closeness has to be reasonable in a sense that it also should not be too close. Considering a day with only a few raindrops less than today would also not help to evaluate the conditional. Regarding the relevant set defining p-situations to even enter the Ramsey test, the numerous experiments run so far in favour of the Ramsey test all tested the trivial case of conditionals with a very clear defined antecedent (e.g. "if the card shows a triangle, then…"). As the consideration above shows, more research is needed on conditionals with less clearly defined antecedents. These are the interesting cases more likely to be uttered in everyday life.

Regarding the second question, how strongly the considered exemplars are taken into account, Experiment 2.1 showed a clear, although small effect of typicality of $p\neg q$ -cases. This result speaks for the stronger emphasis on more similar p-situations. It has to be noted though, that exemplars in this Experiment varied extremely in their typicality measures and that it was always the most typical exemplars that were contrasted against the least typical exemplars available. In Experiment 2.1 it is also unclear, whether people considered the very untypical exemplars to be part of the relevant set at all. It might be that believability measures for conditionals with highly atypical exceptions were higher because those exceptions were not considered as relevant per sé. The results on information acquisition in Experiment 2.2 also speak for a small graded effect of similarity on exclusion in the relevant set. There, more information was retrieved from exemplars more similar to the category mentioned in the antecedent. Regarding the question on how the retrieved information influenced the probability judgment, these results were indifferent to the similarity variation.

In sum, it could be generalized that similarity indeed does have an effect if it comes to deciding which situations to consider for the evaluation of a conditional, but that this effect is less clear in the actual probability judgments for the conditional. The rather small effect sizes for the extreme comparisons in Experiment 2.1 in combination with indifferent regression analysis results in Experiment 2.2 could lead to the conclusion that the evaluation of a conditional probability according to the Ramsey test is a rather dichotomous process where exemplars or situations are either considered relevant to p or not, but if they are

considered they are all weighted the same, leaving frequency as the main source of information.

However, it could also be the case that the similarity variation on Experiment 2.2. was just not strong enough to exert an effect on the actual probability judgments. Since only approximately 60% of available information was retrieved, a weighting of information according to its similarity was potentially hard to detect. To enhance the usage of all information available, it might be promising to run a similar experiment with all the information given in list form. In this way, it is more likely that information over a large similarity range of relevant exemplars is considered and that similarity variations are more likely to exert an effect.

Conclusions

The suppositional account of conditionals based on the Ramsey test was tested against an extended similarity graded version in two experiments. Results confirm the major role of the conditional probability of p, given q, for the interpretation of the conditional. When it comes to decide whether a certain instance is considered as relevant for the evaluation of this probability, similarity of this instance to the situation described in the antecedent p of the conditional seems to still play a minor role. Thus more similar instances are slightly more likely to be considered as relevant. Once the decision about relevance is made, similarity does not seem to play a role and subsequent probability judgment is based on frequency of instances alone.

Chapter 3: Reasoning from conditional premises

3.1 Introduction

There are three theoretical approaches that have assigned a central role to subjective conditional probabilities in reasoning with conditionals: the suppositional theory of conditionals by Evans, Handley and Over (2003, Evans & Over, 2004), the dual process theory of Verschueren (Verschueren, 2004;Verschueren, Shaeken & d`Ydevalle, 2005) and the probabilistic account of Oaksford and Chater (2001, Oaksford, Chater & Larkin 2000). Chapter 3 will discuss the predictions of these three theories for two inferences from a conditional premise, together with a minor premise: MP uses the affirmation of the antecedent as minor premise, as in: "If the fridge door is opened, then the light the light goes on. The fridge is open. Therefore, the light is on". MT uses the negation of the consequent as minor premise, as in "If the fridge door is opened, then the light the light goes on. The light is not on. Therefore, the door must be closed" (see Table 2, Chapter 1). We will show that whereas the three theories make the same predictions for conditions under which people accept MP, they diverge in their predictions for MT. The divergence is clearest when contrasting Verschueren et al. (2005) with Oaksford et al. (2000), and therefore our experiments will focus on the predictions of these two theories.

Suppositional account by Evans & Over

In their suppositional account Evans and Over (2004) state that people come to believe a conditional statement by evaluating the conditional probability P(q|p) as described in the Ramsey test (see Chapter 2). When it comes to conditional reasoning, Evans and Over (2004) draw on evidence showing that MP and MT are endorsed less for conditionals with lower perceived sufficiency (e.g., Cummins, Lubart, Alksnis & Rist, 1991). Mathematically, sufficiency can be defined as a direct function of P(q|p), such that a high conditional probability of the consequent, given the antecedent, corresponds to a high degree of sufficiency of the antecedent for the consequent. Hence, both perceived sufficiency of a conditional and degree of belief in that conditional are based on the same parameter, the subjective conditional probability P(q|p). Whereas MP can be directly derived from this probability, the suppositional strategy for MT involves supposing that p is true, inferring from it via MP that q is true, and then realizing the contradiction of that conclusion with the actual minor premise ($\neg$q). From this contradiction it is inferred that supposition must be false and $\neg$p must be the case. In summary, the theory of Evans and Over implies that both MP and MT inferences should be affected by people's degree of belief in the conditional

premise, which in turn depends on their subjective conditional probability P(q|p). People's willingness of accepting MP should be closely linked to their subjective conditional probability of q, given p, whereas their willingness to accept MT, however, would be less closely associated with their subjective P(q|p) because additional reasoning processes must mediate between believing the conditional premise and arriving at the conclusion that ¬p must be true.

Dual process account by Verschueren, Schaeken and d`Ydevalle

The connection between conditional probabilities and inferences from conditionals is more explicitly elaborated in the dual-process account of Verschueren and colleagues (Verschueren et al., 2005). They assume two processes that determine people's inferences from conditionals at different time intervals after reading the premises. A fast heuristic process makes use of quick automatic estimates of two subjective conditional probabilities associated with a given conditional, P(q|p) and P(p|q). These subjective probabilities (called 'likelihoods' by Verschueren) determine people's initial willingness to endorse inferences from the conditional. The first one, P(q|p), reflects the perceived sufficiency associated with the conditional and is assumed to determine endorsement of MP and MT. The second, P(p|q), reflects the perceived necessity associated with the conditional and determines acceptance of the other two forms of conditional reasoning, "denial of the antecedent" (DA) and "acceptance of the consequent" (AC), see Table 1. In our example, the conditional "if the fridge is opened, then the light comes on" has a high sufficiency and therefore MP and MT should be endorsed to a high degree. It also has a high necessity, so people should also willingly accept AC and DA from it. For instance, given the minor premise that the light is on, most people would be ready to make the AC inference that the door has been opened.

The initial tendency to endorse or reject a conclusion based on the relevant likelihoods is modulated by a second, slower process resting on the assessment of how many counterexamples to a given conditional come to mind. Once again, different counterexamples are relevant for the evaluation of MP and MT on the one hand, AC and DA on the other hand. A case of p and not-q (e.g. open fridge door, dark fridge) counts as a counterexample to the sufficiency of the conditional and thereby blocks acceptance of MP and MT. Thinking of a case of q in the absence of p (e.g. closed fridge with light inside), in contrast, provides a counterexample to the necessity associated with a conditional and thereby blocks acceptance of AC and DA.

To conclude, Verschueren (2004) assumes that people's willingness to endorse MP and MT depend on how sufficient they perceive the conditional's antecedent p to be for the consequent q to occur. Perceived sufficiency equals

the subjective conditional probability P(q|p). Sufficiency estimates on the one hand and the availability of counterexamples on the other hand are partially independent and influence the evaluation of conclusions on separate paths. For an integrated model estimating the effects of both paths see Weidenfeld, Oberauer and Hörnig (2005). Here we focus only on the predictions from the first path: Acceptance rates of both MP and MT should increase as P(q|p), that is the sufficiency of the conditional premise, increases.

Probabilistic Account by Oaksford & Chater

Oaksford and Chater (2001) take an approach that focuses on reasoning from conditionals while bypassing subjective estimates of believability of the conditional. In their probabilistic account they present a mathematical model assuming that inferences are drawn in proportion to the probability of the conclusion, given the minor premise. Their model has three parameters, the two marginal probabilities P(p) = a and P(q) = b, and an exception parameter ε that is defined as the conditional probability that the consequent is not true given that the antecedent is true, P(¬q|p). From these parameters they can derive the probabilities of the conclusion, given the minor premise, for the four basic inferences to be drawn from a conditional together with a minor premise (see Table 2).

MP: $P(q|p) = 1 - \varepsilon$

AC: $P(p|q) = a(1- \varepsilon) / b$

DA: $P(\neg q|\neg p) = (1 - b - a\ \varepsilon) / (1-a)$

MT: $P(\neg p|\neg q) = (1 - b - a\ \varepsilon) / (1-b)$

In their model, the major premise of the argument, that is the conditional "if p then q", plays no direct role in the inference. It could play an indirect role, however, because stating the conditional as a premise could be argued to reduce the exception parameter ε relative to an argument lacking that premise (cf. Liu, 2003 for discussion of the role of the major premise in the framework of Oaksford et al.). A shortcoming of this account is that it has little to say about how people interpret conditionals, that is, factors that affect the degree of belief they have in a given conditional.

To summarize, all three theories agree that acceptance rates of MP depend on people's estimate of P(q|p). This conditional probability determines the degree of belief in the conditional premise in Evans' and Over's theory, the perceived sufficiency in Verschueren's theory, and it equals 1- ε in the model of Oaksford and colleagues. The theories differ, however, in their predictions for MT (Verschueren & Schaeken, 2006). According to Verschueren and colleagues,

acceptance of MT should, like MP, depend on P(q|p). Oaksford and Chater, in contrast, predict that acceptance of MT is only a function of people's estimates of P(¬p|¬q). In a study using conditionals with everyday contents, Verschueren and Schaeken (2006) could show that P(q|p) was a better predictor for the endorsement of MT than P(¬p|¬q). In that study, the conditional probabilities were obtained from people's ratings on the conditionals used in the experiment, and the ratings of the two conditional probabilities were positively correlated across conditionals. After controlling for P(q|p), the correlation between P(¬p|¬q) and endorsement of MT was not significant.

The purpose of the present experiments is to evaluate the contrasting predictions of Verschueren et al. (2005, 2006) on the one hand, and Oaksford et al. (2000) on the other hand, by experimentally manipulating P(q|p) and P(¬p|¬q). Although Evans and Over might generally agree with Verschueren et al. that there should be an influence of sufficiency of the conditional on acceptance of MT, this influence could be blurred due to the assumed suppositional processes to derive MT according to their account. In the present paper we will therefore mainly discuss the results in the light of predictions of the two aforementioned theories.

Experiment 3.1 used pseudo-naturalistic conditionals and manipulated the two conditional probabilities through explicit information about the frequency of the four logically possible combinations of p or ¬p with q or ¬q (c.f. Oberauer and Wilhelm, 2003). Experiments 3.2 and 3.2 also used arbitrary conditionals but manipulated the conditional probabilities through a learning phase in which participants could acquire subjective probabilities through natural sampling (Gigerenzer & Hoffrage, 1995, Oberauer, Weidenfeld, & Hörnig, 2004). Experiment 3.4 used conditionals with everyday content, for which the two critical conditional probabilities were measured by independent ratings.

3.2 Experiment 3.1: probabilistic truth table task employed on reasoning

In Experiment 1 we manipulated P(q|p) and P(¬p|¬q) independently by providing participants with explicit information about the frequencies of the four cases of the conditional's truth table, that is the conjunctions of p & q, p & ¬q, ¬p & q, and ¬p & ¬q. Different cover stories introduced a conditional statement concerning a population of 2000 instances, which were distributed over the four truth-table cases in varying frequencies. The frequencies assigned to each truth-table case in the four conditions are summarized in Table 2.

We will refer to the four conditions as HH, HL, LH, and LL, with the first letter referring to a high or low level of P(q|p), and the second letter referring to a high or low level of P(¬p|¬q). Verschueren et al. predict that the acceptance

rates of both MP and MT will depend on P(q|p), and will be unaffected by P(¬p|¬q). Oaksford and colleagues predict that acceptance rates of MP will depend on P(q|p) only, whereas acceptance rates of MT will depend only on P(¬p|¬q).

Table 6 : Experimental manipulation in Experiment 3.1.

Conditions:	**HH**	**HL**	**LH**	**LL**
pq	900	900	100	500
p¬q	100	100	100	500
¬pq	100	900	900	500
¬p¬q	900	100	900	500
P(q|p)	**.9**	**.9**	**.5**	**.5**
P(¬p|¬q)	**.9**	**.5**	**.9**	**.5**

First letter of the condition code represents the conditional probability P(q|p) = sufficiency, the second letter represents the conditional probability P(¬p|¬p) = the probability of the conclusion, given the minor premise for MT.

Participants were asked to rate their degree of belief in a given conditional and to evaluate the four basic inference forms MP, AC, DA, and MT. Thereby we are able to directly test the relationship between degree of belief in the conditional and acceptance of inferences in the reasoning tasks. As shown in other studies before (Evans et al., 2003, Oberauer & Wilhelm, 2003), high objective P(q|p) should result in a higher degree of belief in the conditional. There is no empirical evidence so far that P(¬p|¬q) should influence believability ratings of the conditional, so we expect no effect of this factor on the degree of belief in the conditional.

3.2.1 Method

Participants

Participants were 60 psychology students at the University of Potsdam (age range: 19-36 years). Order of tasks was varied between participants, yielding a factor *task order* with two groups of 30 participants that gave probabilistic judgment first and then solved reasoning tasks, and 30 participants working on the tasks in reversed order.

Material and Procedure

For both tasks, judging the probability of the conditional and judging the four inference forms, participants received eight trials presented on a monitor, two trials for each condition in Table 2. Each trial was based on a short story of pseudo-naturalistic circumstances and a conditional statement made by an

"expert" of the matter. The eight conditionals and their cover stories were taken from Weidenfeld et al. (2005). The pseudo-naturalistic content was chosen so that the conditional was semantically rich but the connection between antecedents and consequents was arbitrary and not influenced by common world knowledge. For each participant the eight conditionals were assigned at random to the eight trials of the probability judgment task, and the same assignment was used for the inference tasks. The presentation order of cover stories and conditions was also determined at random for each participant, and was repeated for the two tasks (probability judgment and inferences).

Each trial was presented on a single screen, which first displayed the cover story, followed by information about the distribution of 2000 instances over the four truth-table cases according to the experimental condition (see Table 2). Next, the conditional was introduced as a claim by the expert. For the probability judgment task, participants were then asked to rate the probability that the expert was right. For the inference task, they were instead presented with the four inferences, MP, AC, DA, and MT in random order. Depending on the task order variable, participants either completed the eight trials of the probability judgment task first and then moved on to the eight trials of rating the four inferences, or completed the tasks in reverse order. The relevant cover story, including the frequency information, was displayed at the beginning of each trial of each task. In the following, we give an example of a conditional with its cover story:

In the year 4000, astrophysicists discovered a new inhabited planet in a foreign galaxy. Scientists are engaged in resolving the biophysical characteristics. In a lot of places the planet's atmosphere contains philoben gas unknown to terrestrial atmosphere. Of 2000 probes of this planet's atmospheric particles it is known that:

900 probes were rich of philoben gas and warmer than 22° centigrade.

100 probes were rich of philoben gas and not warmer than 22° centigrade.

100 probes were not rich of philoben gas and warmer than 22° centigrade.

900 probes were not rich of philoben gas and not warmer than 22° centigrade.

An expert claims that:

"If the probe is rich in philoben gas, then it is warmer than 22° centigrade".

For the probability judgment task, participants were asked at this point to judge whether the expert was right on a scale from 0 ("absolutely impossible") to 100 ("absolutely certain"). For the inference task, the cover story and conditional were instead followed by the four inference tasks in random order. The following is an example for the Modus Tollens (MT) inference task:

Expert's statement: "If the probe is rich in philoben gas, then it is warmer than 22° centigrade".
Observation: The probe is not warmer than 22° centigrade.
Conclusion: "The probe is not rich in philoben gas".

Participants had to rate their confidence in the conclusion on a 6 point scale ranging from "certain that I *can* draw the conclusion" to "certain that I *cannot* draw the conclusion", as introduced by Cummins et al. (1991) and later also used by Oaksford et al. (2000).

3.2.2 Results

Data of all dependent variables were submitted to a 2 x 2 x 2 ANOVA with *task order* (probability judgments first vs. last)*, P(q|p)* (high vs. low) and *P(¬p|¬q)* (high vs. low) as factors.

Probability of the conditional

There was no significant main effect of task order, and no interaction involving this variable (all F < 2.6). Figure 5 shows participants´ evaluation of the believability of the conditional statement for factor levels of P(q|p) and P(¬p|¬q). The main effect for P(q|p) was significant with F(1, 58) = 255.0, p<0.001, ε^2= 0.82. The main effect of P(¬p|¬q) was not significant (F < 1), but P(¬p|¬q) interacted with P(q|p), F (1, 58) = 20.6, p<0.001, ε^2= 0.26. Figure 5 shows that ratings of believability of the conditional highly depend on the conditional probability P(q|p). In addition, when P(q|p) was high, a higher level of P(¬p|¬q) made the conditional more believable. The opposite effect of P(¬p|¬q) was observed with low levels of P(q|p). This interaction can in part be explained by the difference in the frequency of pq cases between conditions LH and LL. Previous studies (Evans et al., 2003, Oberauer & Wilhelm, 2003) have shown that a minority of participants base their judgment of the probability of the conditional not on P(q|p), but on the relative frequency of the pq conjunction, which was higher in the LL condition than in the LH condition (cf. Table 6).

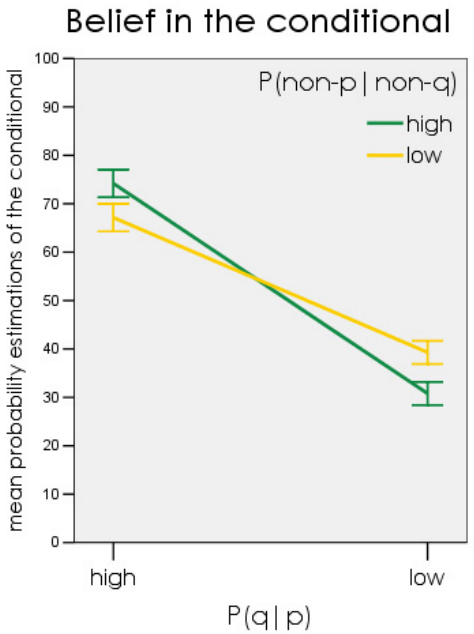

Figure 5: Believability of the conditional
Probability estimates (on a scale from 0 to 100) for the conditional. P(q|p) is grouped on the x-axis, P(¬p|¬q) is indicated by different colours.

Reasoning Tasks

In all experiments, answers for the inference tasks were coded from +5 ("certain that I can draw this conclusion") to -5 ("certain that I cannot draw this conclusion") in steps of 2 to generate equal numerical distances between the six answer options.

The mean acceptance ratings of Modus Ponens are shown in the left panel of Figure 6. There were two significant main effects, for P(q|p), F (1, 58) = 36.8, $p<0.001$, $\eta_p^2= 0.39$ and for task order, F(1, 58) = 10.9, $p<0.01$, $\eta_p^2= 0.16$. None of the other factors or their interactions reached significance (all F<1).

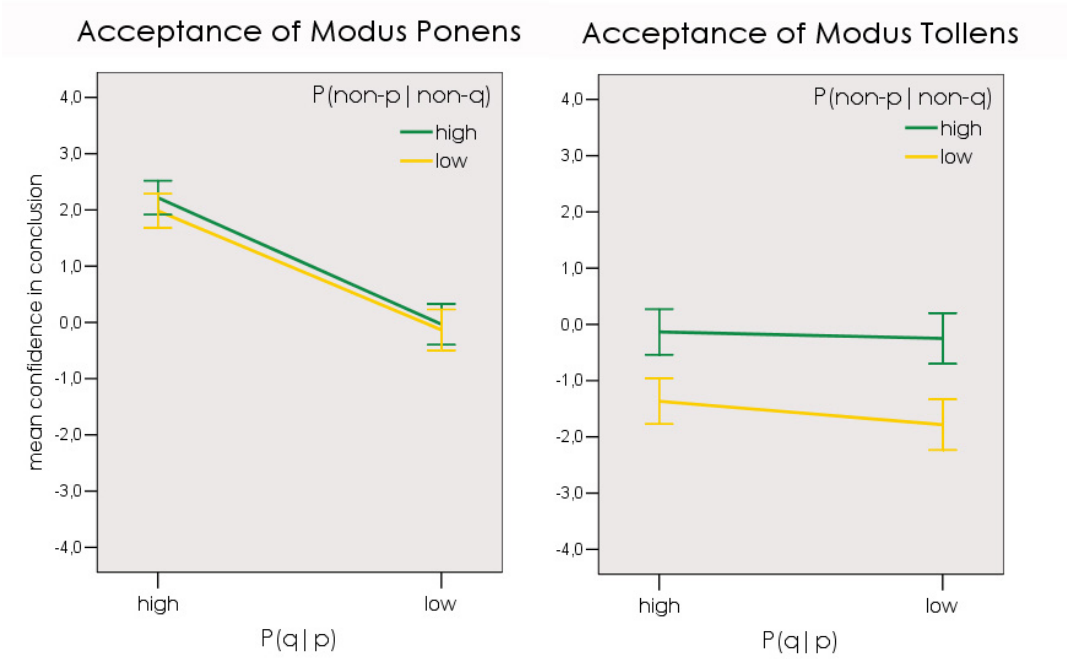

Figure 6: Acceptance of Modus Ponens and Modus Tollens.
Confidence ratings in the conclusions (on a scale from -5 to +5).

Modus Ponens increased with the conditional probability P(q|p), as predicted by all three probabilistic theories of conditionals discussed above. In addition, MP was accepted more willingly when probability judgements of the conditional were made first.

The right panel of Figure 6 shows acceptance data for Modus Tollens. As predicted by the theory of Oaksford et al. (2000), there was a significant main effect of P(¬p| ¬q), F(1, 58) = 14.1, $p<0.001$, $\eta_p^2= 0.20$ and no main effect of P(q|p), F(1,58)=1.9. Further, the interaction between P(q|p) and task order was significant, F(1, 58) = 6.3, $p<0.05$, $\eta_p^2= 0.10$; P(q|p) had a smaller effect on MT when the conditional was rated first compared to when it was rated last. None of the other effects reached significance (all F < 1.9).

In Figure 7, left panel, the data for AC acceptance are presented. The main effect of P(q|p), $F(1, 58)= 26.9$, $p<0.001$, $\eta_p^2= 0.31$ and the interaction between P(q|p) and P(¬p|¬q), $F(1, 58)= 7.4$, $p<0.01$, $\eta_p^2= 0.11$ were significant. For DA, right panel, a significant main effect of P(q|p), $F(1, 58) = 6.4$, $p<0.05$, $\eta_p^2= 0.10$, P(¬p|¬q), $F(1, 58)= 8.1$, $p<0.01$, $\eta_p^2= 0.12$ and an interaction of both factors, P(¬p|¬q), $F(1, 58)= 8.5$, $p<0.01$, $\eta_p^2= 0.13$, was observed.

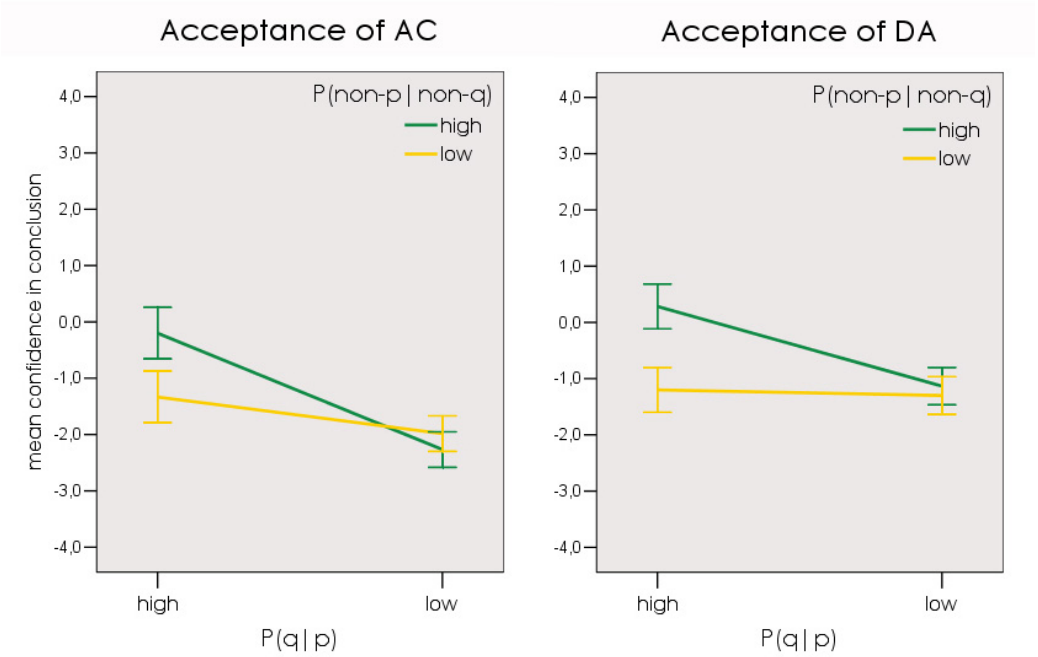

Figure 7: Acceptance of AC and DA.
Confidence ratings in the conclusions (on a scale from -5 to 5).

3.2.3 Discussion

Both theories under consideration predicted that acceptance of MP will depend on P(q|p), and this was the case. The theories diverge in their predictions on MT. The results confirmed the prediction of Oaksford et al. (2000) that acceptance of MT depends only on P(¬p|¬q), and not on P(q|p). This result contradicts the prediction of Verschueren et al. (2005). It also contradicts the assumption of Evans and Over (2004) that acceptance of MT depends on the belief in the conditional, because people's degree of belief in the conditional was a function of P(q|p), and not of P(¬p|¬q).

The experimental manipulation was not designed to compare predictions of the theories regarding AC and DA, but the results for those inferences will still be discussed. For AC, both theories would predict that acceptance should depend on P(p|q), which is the perceived necessity of the conditional and likewise the

conditional probability of the conclusion p, given the minor premise q. This conditional probability was high (0.9) in the HH condition and low (0.5 in HL & LL, and 0.1 in LH) in the other three conditions, which is roughly reflected in the actual answer patterns for AC.

For DA, Verschueren et al. would predict the same answer pattern as for AC, whereas according to Oaksford et al. DA should depend on $P(\neg q|\neg p)$. The latter conditional probability is also high in the HH condition (0.9) and low in the other three conditions (0.5 respectively 0.1 in the HL lo condition). Considering the assumption that a conditional premise (or a conclusion from it) should at least exceed a 0.5 threshold to gain some support of believability it could be argued, that any probability of 0.5 or below would lead to comparable low acceptance, as was observed for AC and DA. The two theoretical approaches are therefore not distinguishable in their predictions for the two inferences AC and DA.

Turning to the results on MP and MT, it could be, that providing people with numerical frequency information biases reasoners towards computations along the lines assumed by Oaksford et al. (2000) because the information required for calculating the relevant conditional probabilities is so easily available. The next two experiments were designed to test whether the results of Experiment 3.1 generalize to situations in which the relevant frequency or probability information is not given together with the problem but instead has to be retrieved from memory. This change arguably moves the experimental situation closer to most real-life uses of conditionals, in which explicit numerical information about the frequency of truth-table cases is rarely available.

3.3 Experiment 3.2 and 3.3: a learning version of the probabilistic truth table task employed on reasoning

Experiments 3.2 and 3.3 repeated the design of Experiment 3.1 with probabilities manipulated through a probability learning task. Instead of directly presenting the frequency distributions of truth table cases (cf. Table 7) to participants, they had a chance to learn them in a task where they guessed for 100 exemplars to which of the four truth-table cases each of them belonged, and receiving feedback on the truth-table category of each exemplar. This sequential acquisition of information is also referred to as "natural sampling" and advocated by Gigerenzer and Hoffrage (1995) for its relative approximation to acquisition of frequency and probability information in everyday life. It could be shown in studies using probability judgment tasks that the use of a natural sampling procedure reduces, for example, the otherwise often observed base rate neglect (Betsch, Biel, Eddelbüttel & Mock, 1998, Gigerenzer & Hoffrage, 1995). Oaksford and Wakefield (2003) also advocated the natural sampling procedure as a more adequate way of manipulating subjective probabilities for a test of Oaksford and Chater's (2003) theory of the Wason selection task. Therefore, we sought to

Table 7: Experimental manipulation in Experiment 3.2 and 3.3.

Conditions:	**HH**	**HL**	**LH**	**LL**
pq	45	45	5	25
p¬q	5	5	5	25
¬pq	5	45	45	25
¬p¬q	45	5	45	25
P(q|p)	**.9**	**.9**	**.5**	**.5**
P(¬p|¬q)	**.9**	**.5**	**.9**	**.5**

Legend: First letter of the condition code represents the conditional probability P(q|p) = sufficiency, the second letter represents the conditional probability P(¬p|¬p) = the probability of the conclusion, given the minor premise for MT.

reduce effects of potentially artificial reasoning processes evoked through the explicit presentation of numerical frequencies in Experiment 1 by using a natural-sampling presentation format for the frequency information.

To avoid interference, each participant learned only one of the four frequency distributions of our design. Therefore, we used a between-subjects design in Experiment 3.2 and 3.3. Moreover, the population size of instances was reduced from 2000 to 100 to keep the learning time within a reasonable range. The frequencies used for each condition are shown in Table 7.

3.3.1 Method

Experiment 3.2 was a computerized study run in a university lab. Experiment 3.3 was a parallel experiment run in the internet, accessible through the Website of the Experimental Weblab of the Department of Psychology at University of Potsdam.

Participants

Participants of Experiment 3.2 were 80 high school students and first-year university students from Potsdam (age range: 17-23 years). Order of tasks was varied between participants, with two subgroups in each of the condition groups, resulting in 4 subgroups giving probabilistic judgment first and then solved reasoning tasks (n= 40) and another 4 subgroups of participants (n= 40) working on the tasks in reversed order. Participants were assigned randomly to one of the subgroups of N=10.

Participants of Experiment 3.3 were 108 internet users. To prevent multiple participation of the same person, the IP-address of participants was collected and data sets from IP addresses appearing more than once removed. Participants in the final sample were 102 internet users aged between 19 and 53 years. Procedure was exactly as in Experiment 2A, resulting in 8 subgroups of 10 to 15 participants each.

Material and Procedure

For the learning phase, participants were told that they were going to take part in a quiz were they could earn points by guessing the correct characteristics of 100 playing cards. The cards could be threes or nines of either spades or hearts. The instruction stressed that there were different numbers of cards in each of the four categories, and participants were alerted that they could benefit from this information over time if they used the underlying frequency distribution to maximise their outcome of points. Participants then worked through 100 screens, each showing the back side of a playing card. They entered their guess on which of the 4 categories the card belonged to by clicking one of four buttons assigned to the four categories (e.g. "9 of spades"). The front side of the card was then shown and participants received feedback as to whether or not they guessed correctly and thus earned a point. The order of the 100 cards was randomised for each participant individually but had the same frequency distribution for all participants in one group according to the design in Table 3.

In the assessment phase half of the participants in each design level group answered the question on the probability of the conditional first and then assessed the four inferences tasks, and the other half answered the questions in reversed order. The probability judgment task was presented on one screen and the four inference tasks MP, MT, DA and AC were presented together on a further screen. The four inferences were presented in randomised order for each participant. Instructions for the probability of the conditional were as follows:

A person who has seen the same cards as you claims:

´If the card is spades, then it is a nine´

How likely do you think this person is right? Please enter a number between 0 ("absolutely impossible") and 100 ("absolutely certain").

The instructions for the inference tasks were as follows (e.g., for Modus Tollens):

In the following you will read two statements that refer to the cards you have just seen. From these two statements a conclusion is drawn. Please indicate how certain you are that you can draw this conclusion.

Assumption: "If the card is spades, then it is a nine".

Observation: The card is not a nine.

Conclusion: "The card is not spades".

As in Experiment 3.1, confidence in the conclusion had to be rated on a 6 point scale ranging from ("certain that I can draw the conclusion") to ("certain that I cannot draw the conclusion").

3.3.2 Results

Because the control factor *task order* did not reach significance for any of the dependent variables, all data analyses were collapsed over the two task order groups. The two experiments were exact replications of each other, differing only in target population (students vs. internet users) and testing method (university lab vs. internet). We analyzed the data of both experiments together, introducing *data source* as an independent variable, to assess to what degree the lab-based experiment and the internet experiment yielded significantly different results. Hence, data of all dependent measures from all 182 participants were submitted to a 2 x 2 x 2 ANOVA with *P(q|p)* (high vs. low), *P(¬p| ¬q)* (high vs. low) and *data source* (Lab vs. Web) as factors.

Probability of the conditional

Figure 8 shows the main effect of P(q|p) for the believability of the conditional, $F(1,174) = 45.3$, $p<0.001$, $\eta_p^2 = 0.21$. Neither the main effect of P(¬p| ¬q) nor of *data source*, nor any of their interactions were significant (all $F < 1.6$).

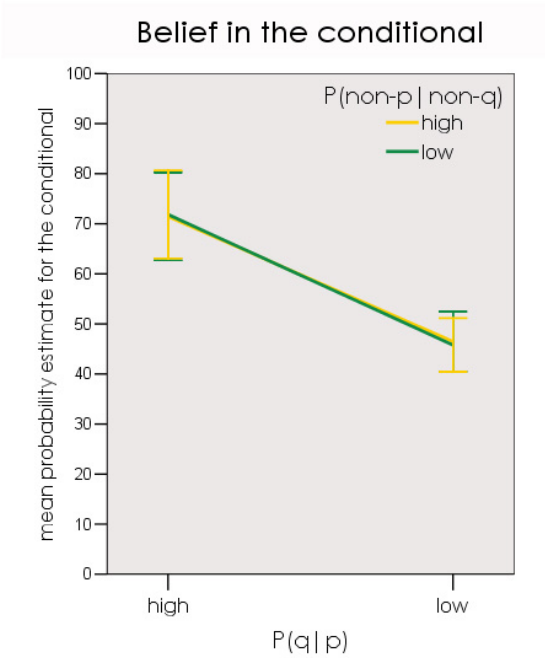

Figure 8: Believability of the conditional
Probability estimates (on a scale from 0 to 100) for the conditional. P(q|p) is grouped on the x-axis, P(¬p|¬q) is indicated by different patterns.

Reasoning Tasks

The acceptance rates for Modus Ponens and Modus Tollens are shown in Figure 9. The left panel shows the main effect of P(q|p) on acceptance ratings of Modus Ponens, $F(1,174) = 5.8$, $p < 0.05$, $\eta_p^2 = 0.03$. Acceptance of MP increased with P(q|p), showing the same pattern as the probability ratings of the conditional. None of the other factors or their interactions reached significance (all $F < 2.0$).

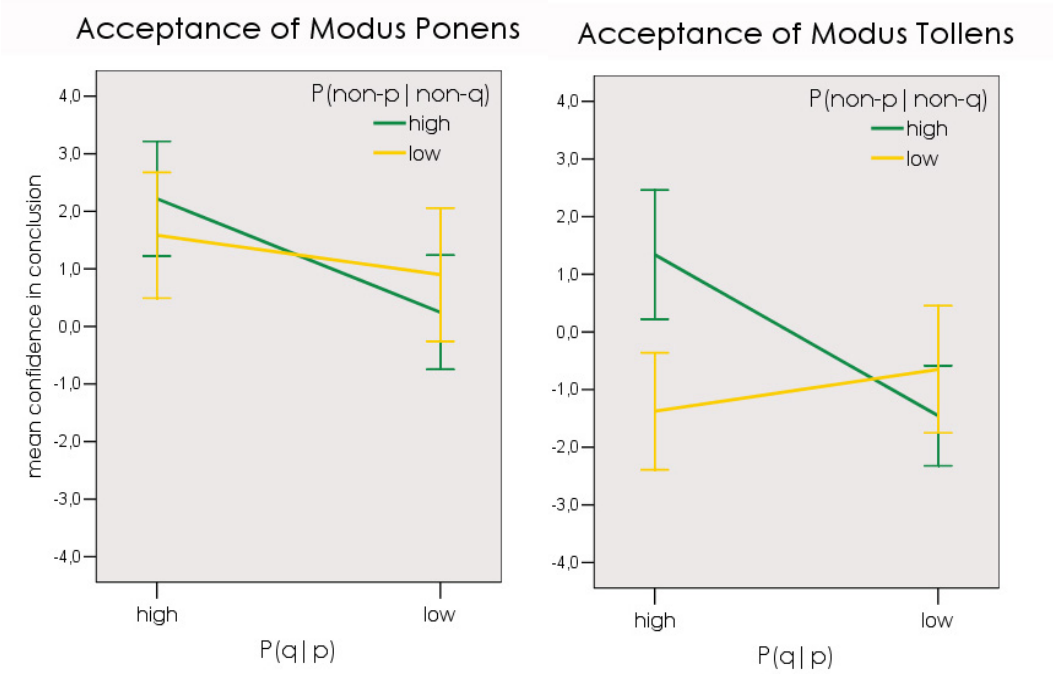

Figure 9: Acceptance of Modus Ponens and Modus Tollens
Confidence ratings in the conclusions (on a scale from -5 to +5).

The right panel of Figure 9 shows acceptance data for Modus Tollens. The analysis yielded a significant main effect of P(¬p|¬q), $F(1,174) = 4.0$, $p < 0.05$, $\eta_p^2 = 0.03$, and a trend towards an effect of P(q|p), F= (1,174) =3.4, p=0.067, $\eta_p^2 = 0.02$ (with one-tailed testing, p=0.034). Unlike in Experiment 1, the interaction of P(q|p) and P(¬p| ¬q) was significant, $F(1,172) = 11.7$, p=0.001, $\eta_p^2 = 0.06$. As can be seen in Figure 9, acceptance of MT increased with a high level of P(¬p| ¬q) only when P(q|p) was also high. A post-hoc t-test revealed a difference for P(¬p|¬q) only when P(q|p) was high, $t(87) = 3.6$, $p < 0.001$. When P(q|p) was low, it did not matter how probable the conclusion, given the premise was, $t(91) = -1.1$, $p > 0.2$; endorsement of MT was low in either case. Again, *data source* had no significant main effect and was not involved in any interaction (all $F < 1.6$).

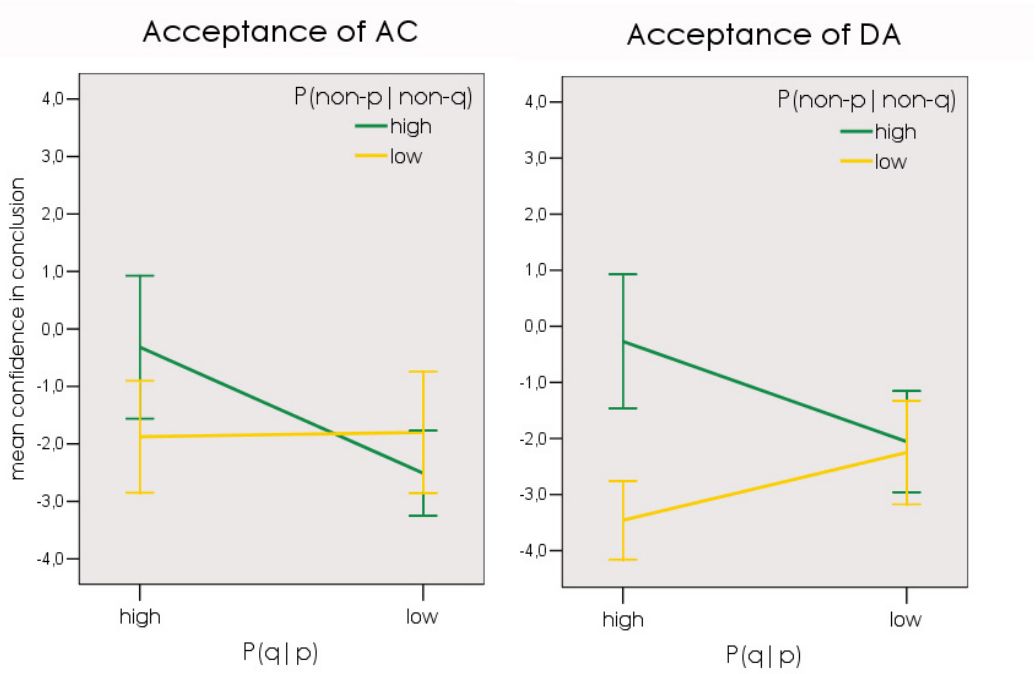

Figure 10: Acceptance of AC and DA.
Confidence ratings in the conclusions (on a scale from -5 to +5).

Figure 10, left panel, shows the data for AC acceptance. The main effect of P(q|p), $F(1,174)= 5.7$, $p<0.05$, $\eta_p^2= 0.03$ and the interaction between P(q|p) and P(¬p|¬q), $F(1, 172)= 4.8$, $p<0.05$, $\eta_p^2= 0.03$ were significant. There was also a small main effect for *data source* $F(1,174)= 5.2$, $p<0.05$, $\eta_p^2= 0.03$. Internet users gave AC lower ratings than laboratory participants. For DA, right panel, a significant main effect of P(¬p|¬q), $F(1,174)= 11.6$, $p<0.01$, $\eta_p^2= 0.06$ and an interaction of P(q|p) and P(¬p|¬q), $F(1,172)= 9.9$, $p<0.01$, $\eta_p^2= 0.05$, was observed. None of the other effects reached significance (all $F < 1.4$).

3.3.3 Discussion

Over all, besides a small effect in acceptance of AC, the findings from Experiments 3.2 and 3.3 were not reliably different, demonstrating once more that, at least in the field of reasoning, internet based experiments and lab experiments yield comparable results (Klauer, Musch, & Naumer, 2000).

As in Experiment 3.1, P(q|p) led to a higher degree of belief in the conditional, and a higher confidence in the conclusion of Modus Ponens, in line with the predictions of all three theories. Contrary to Experiment 1 neither of the two probabilities P(q|p) and P(¬q|¬p) could account for the acceptance rates for Modus Tollens alone. Rather, it was the interaction of both probabilities that described the acceptance of MT in the present experiment.

We propose a two-stage inference process to explain these results. In a first step, the believability of the conditional premise is estimated. Only if the result of this first step is positive (i.e., the believability of the conditional is high), participants proceed to the second step, which consists of assessing the conditional probability of the conclusion, given the minor premise. If any of the probabilities is low, the conclusion is rejected. Consistent with this interpretation, the post-hoc comparison of conditions for acceptance of MT revealed an effect of $P(\neg q|\neg p)$ only when $P(q|p)$ was high.

Acceptance rates for AC and DA, as discussed in Experiment 3.1 can only be summarized descriptively. Both theories tested, as well as our two step proposal would lead to comparable predictions for these two inferences. Only in the HH condition is the believability of the conditional (0.9) and the conclusion, given the minor premise (0.9) high. In all other conditions either the believability of the conditional or the conclusion is low (0.5, respectively 0.1 in the LH condition for AC and in the HL condition for DA). Thus results for AC and DA do not differentiate between the three different theoretical explanation and are consistent with all three of them.

It is important to note, that the 2-stage evaluation process is different from the dual process account by Verschueren et al (2005). The first process in their account is evaluating the sufficiency of the conditional, that is, $P(q|p)$. Their first step, hence, is equivalent to the first step in our proposal, assessing the believability of the conditional upon an estimate of $P(q|p)$. The second process in Verschueren et al. consists of looking for disabling information that would falsify the conclusion in question. For both MP and MT, disabling conditions are defined as conjunctions of $p \,\&\, \neg q$, and the conclusions are rejected to the degree that instances of $p \,\&\, \neg q$, or reasons for their possible occurrence, are found. Our second step, in contrast, consists of the evaluation of the probability of the conclusion, given the minor premise, as assumed by Oaksford and Chater. Hence, the idea of a two-step process of inferential reasoning put forward here includes assumptions of both theories tested, with a chronological priority of the process described in Verschueren et al (2005).

To put the two-stage inferential process to a further test, we conducted Experiment 3.4 using everyday, semantically rich conditional sentences. In using such conditionals, we could completely refrain from using arbitrary frequency information, neither explicitly mentioned nor previously learned during the experiment. In tasks where only the conditional statement (and a minor premise) is given explicitly and the probabilities associated with it must be added from world knowledge, people should be less prone to use a shortcut bypassing the conditional premise, as observed in Experiment 3.1. If the two-stage assumption is correct, we should again find an interaction of $P(q|p)$ and $P(\neg p|\neg q)$ as determinants of the acceptance of MT conclusions.

Also, since the results of Verschueren and Schaeken (2006) were derived from a study using everyday conditionals, in order to substantiate our alternative explanation of inferential reasoning, the third study used the same kind of material as Verschueren and Schaeken (2006).

3.4 Experiment 3.4: everyday conditionals

The third study used conditionals with everyday content. In a pre-test, 100 conditionals were rated for P(q|p) and P(¬p|¬q), and we selected conditionals on the basis of these ratings to fill the four cells of the design presented in Table 4.

Table 8: *Experimental design used in Experiment 3.4 with example items*

		P(q|p)	
		high	**low**
P(¬p|¬q)	**high**	HH: If you see molehills in a yard, then there are moles.	LH: If somebody is a millionaire, then he prefers coffee over tea.
	low	HL: If a soccer team plays out of town, then a referee opens the game.	LL: If somebody finds a four leaves clover, then the day is warmer then 15 degrees.

Using this method we designed experimental material that is comparable to that of Verschueren and Schaeken (2006). They also used everyday conditionals rated for P(q|p) and P(¬p|¬q). Different from their study, we made an attempt to find sets of conditionals that vary independently and to an equal degree with regard to the two critical conditional probabilities.

3.4.1 Pre-Test

A set of 100 conditionals in German with everyday content was constructed with an eye on obtaining a large range of both conditional probabilities, P(q|p) and P(¬p|¬q). The conditionals in this set were rated for the two conditional probabilities of interest by 206 participants who filled in an internet based questionnaire. Instructions for the two rating questions were as follows (using one example conditional):

Consider the following statement:

"If you see molehills in a yard, then there are moles."

Question 1 (for P(q|p)):
You see molehills in a yard. How probable is it, that there are moles? Please give a number between 0 (completely impossible) and 100 (absolutely certain).

Question 2(for P(¬p|¬q)):
There are no moles in a yard. How probable is it, that you do not see molehills there? Please give a number between 0 (completely impossible) and 100 (absolutely certain).

Each of the 206 participants (aged between 17 and 66) answered the two questions for 20 randomly chosen conditionals. Data that were produced by people who did not complete the whole survey were nevertheless included in the analysis. Ratings for all 100 conditionals tested for Experiment 3.4 are shown in Appendix 3.2. On the basis of these ratings five conditionals were chosen to fit in each cell of the 2 x 2 design of high vs. low mean ratings on the two dimensions P(q|p) and P(¬p|¬q). Ratings for the 20 conditionals used in Experiment 3.4 are shown in Appendix 3.3. The ratings for the chosen conditionals differ in the two dimensions on an absolute level, but the differences between conditions were comparable for both independent variables. There was no correlation between the two conditional probabilities (r= 0.13, p=0.57) over the 20 conditionals used in the main experiment.

3.4.2 Method

Participants

Participants were 30 last-year high school and first-year university students studying different subjects (age range: 17 – 27 years). Order of tasks was varied between subjects such that 15 participants gave probabilistic judgments first and then solved reasoning tasks, and 15 participants worked on the tasks in reversed order.

Material and Procedure

The experiment was a computer based study realizing the design in Table 8 within subjects for probability judgements and reasoning tasks. The procedure used to assess these variables was the same as in Experiment 3.1, 3.2. and 3.3. Again, participants rated the believability of the conditional on a scale from 0 to 100 and solved all four inference tasks for each conditional either before or after probability judgments of all conditionals. Participants rated their confidence in the conclusion on the same 6- point scale as in the previous experiments.

3.4.3 Results

Data of all dependent variables were submitted to a 2 x 2 x 2 ANOVA with *task order* (probability judgements first vs. last), *P(q|p)* (high vs. low) and *P(¬p|¬q)* (high vs. low) as factors.

Probability of the conditional

Figure 11 shows participants´ evaluation of the believability of the conditional statement. P(q|p) again had a large effect on the probability of the conditional with F(1,28) = 215.5, p<0.001, η_p^2 =0.89. Different from previous studies, we also obtained a significant main effect of P(¬p|¬q) on the probability of the conditional with F(1,28)= 24.5, p<0.001, η_p^2 =0.47. In addition, the two conditional probabilities interacted, F=(1,28) = 39.6, p<0.001, η_p^2 =0.59, As can be seen in Figure 11, high levels of P(q|p) and P(¬p|¬q) supported the belief in the conditional, although the positive effect of P(¬p|¬q) only prevailed when P(q|p) was high. A post-hoc comparison revealed that in this case the lack of P(¬q|¬p) lowers its perceived believability, t(29)= 5.98, p<0.001, but there was no effect of P(¬p|¬q) when P(q|p) was already low, t(29)= -.62, p= .5. None of the other effects reached significance (all F<2.5).

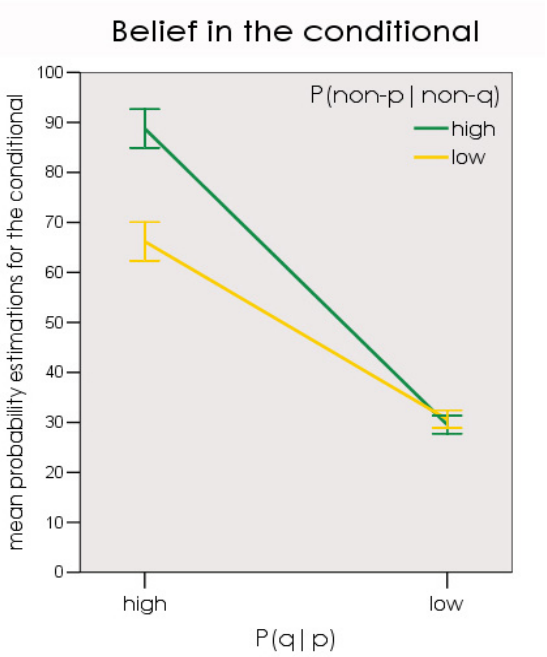

Figure 11: Believability of the conditional
Probability estimates (on a scale from 0 to 100) for the conditional. P(q|p) is grouped on the x-axis, P(¬p|¬q) is indicated by different patterns.

Reasoning tasks

As shown in the left panel of Figure 12, effects on acceptance rates of Modus Ponens were similar to those on the probability of the conditional. There was a significant effect of P(q|p) with F(1,28)= 34,7, p<0.001, η_p^2 =0.56, a smaller main effect of P(¬p|¬q) with F(1,28)=6,0, p<0.05, η_p^2 =0.18, and a significant interaction of P(¬p|¬q) with P(q|p), F(1,28)= 8,8, p<0.01, η_p^2 =0.24. A post-hoc comparison of conditions showed that if P(q|p) was high, P(¬q|¬p) did have an effect on endorsement of MP, t(29) = 4.77, p<0.001. If P(q|p) was low, this effect did not occur, t(29) = -1.97, p=0.14.

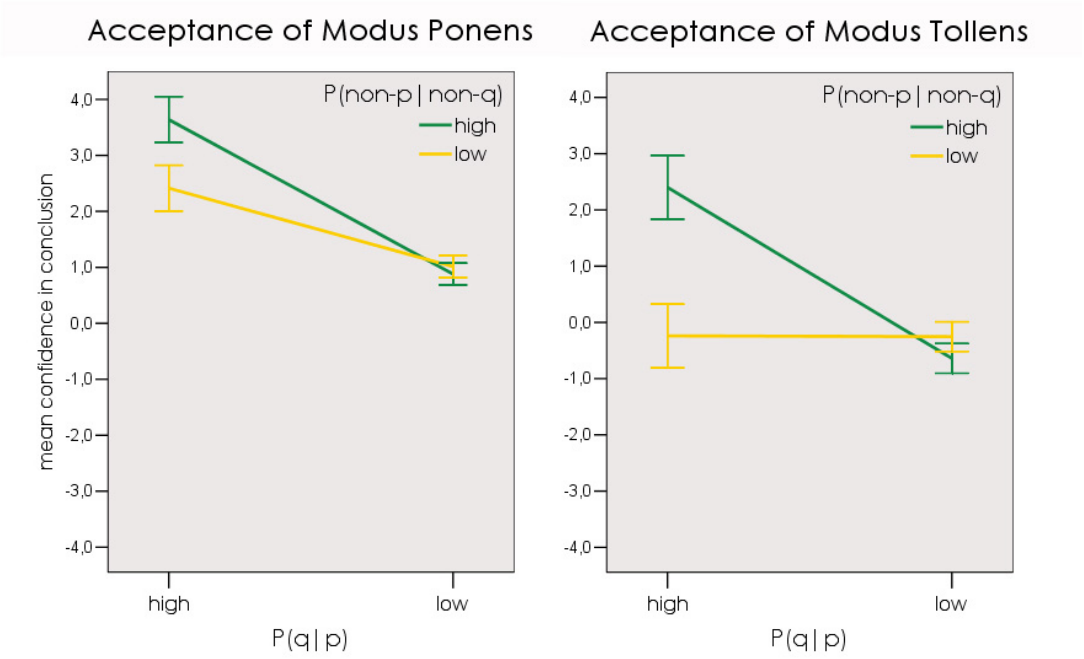

Figure 12: Acceptance of Modus Ponens and Modus Tollens
Confidence ratings in the conclusions (on a scale from -5 to +5).

Additionally, the factor task order had an effect on MP, $F(1,28)= 6,1$, $p<0.05$, $\eta_p^2 =0.2$ in such a way that when the conditional was rated first, MP was less endorsed. Also, task order interacted with P(q|p), $F(1,28)= 8,0$, $p<0.01$, $\eta_p^2 =0.22$. P(q|p) had a greater effect on MP acceptance when the conditional was rated first, as opposed to the condition where inference tasks were answered first. None of the other effects reached significance (all $F<1$).

The right panel of Figure 12 shows acceptance data for Modus Tollens. There was a large main effect of P(q|p) with $F(1,29) = 36,5$, $p<0,001$, $\eta_p^2 =0.57$, a smaller effect of $P(\neg p|\neg q)$, $F(1,29) = 10,9$, $p<0.01$, $\eta_p^2 =0.28$, and an interaction of these two variables, $F(1,29) = 34,8$, $p<0,001$, $\eta_p^2 =0.55$. Task order was not involved in any significant effect (all $F < 2.6$).

In Experiment 3, I refrained from analysing results of AC and DA, since two important dimension necessary to test predictions of the two theories and the two-stage inferential process in question were not assessed for the 20 conditionals used. This is firstly P(p|q), the necessity of the conditional, that according to Verschueren should affect AC and DA, and secondly $P(\neg q|\neg p)$, the probability of the conclusion $\neg q$, given the minor premise $\neg p$, that according to Oaksford should affect DA. This probability would also affect acceptance of DA in the second step of the two-stage interferential process suggested here.

Multiple Regression

In Experiment 3.4, conditions of the factor levels of P(q|p) and P(¬p|¬q) were realized through specific conditional sentences. Since each of them was rated in both of the probability dimension, we ran multiple regression analyses with the 20 conditionals as cases to determine the effects of P(q|p), P(¬p|¬q), and the rated probability of the conditional, P(cond), on acceptance rates of MP and MT. Separate regressions were run using the three predictors to predict MP and to predict MT. Results are shown in *Table 5*.

Table 9: Results of the multiple regression analysis in Experiment 3.4

Modus Ponens	**R^2 = .88**	**Beta**	**t(19)**	**p**
	P(cond)	.81	3.59	.002**
	P(q|p)	.13	.57	.577
	P(¬p|¬q)	.06	.64	.530
Modus Tollens	**R^2 = .68**	**Beta**	**t(19)**	**p**
	P(cond)	1.08	2.97	.009**
	P(q|p)	-.41	-1.16	.265
	P(¬p|¬q)	.31	2.24	.039*

For Modus Ponens 88% of the variance could be explained by one single predictor, the belief in the conditional, P(cond). P(q|p) could not explain any additional variance, since it was highly correlated with P(cond), r = .93.

For Modus Tollens two predictors were found to jointly account for 68% of the variance. Variations in P(cond), reflecting the belief in the conditional sentence, explained the larger amount of variance in MT acceptance. P(¬p|¬q), the probability of the conclusion, given the minor premise, accounted for a smaller, but still substantial additional amount of variance.

3.4.4 Discussion

Experiment 3.4 showed, like Experiment 3.1-3.3, a large effect of P(q|p) on the believability of the conditional. In this experiment we used natural conditionals with a semantically rich content and thus could corroborate findings of Over, Hadjichristidis, Evans, Handley and Sloman (in press), who also found that the subjective probabilities of conditionals with everyday contents were

largely determined by P(q|p). These findings are important because they show that people assess the probability or believability of a conditional on the basis of their estimates of P(q|p) not only when the conditional probability was provided by the experimenter but also when the relevant probabilistic information must be retrieved from world knowledge associated to the contents of the conditionals under evaluation.

In addition to the large main effect of P(q|p) on the subjective probabilities of conditionals we found a smaller main effect of P(¬p|¬q) and an interaction. This additional effect for everyday conditionals can be interpreted as a residual impact of the causal contrast on judgements for believability of the conditional. The causal contrast (or delta-p rule, P(q|p) - P(q|¬p)) is higher in the condition were P(q|p) and P(¬q|¬p) are both high and thus might have led to higher believability judgements. Over et al. (in press) also found an additional small (in their case negative) effect of P(q|¬p) on believability judgements for the conditional, which equals the reverse effect of P(¬q|¬p) in our study.

Results for acceptance of Modus Ponens closely matched the results of the believability ratings of the conditional. Moreover, in the regression analysis the ratings of the probability of the conditional emerged as the only significant predictor of MP acceptance ratings. This finding confirms the assumption of Verschueren and Schaeken (2005) and Evans and Over (2004) that the acceptance of MP is directly driven by people's assessment of the probability of the conditional premise.

With the results for Modus Tollens in Experiment 3.4 we see our dual stage assumption confirmed. Again, there was a significant decrease of acceptance of MT with a low P(¬p|¬q) = P(conclusion|minor premise) in the conditions were P(q|p) was high, which wasn't apparent when P(q|p) was low.

3.5 General discussion

A two-stage inferential reasoning process

Four experiments were reported testing predictions from two probabilistic theories of reasoning from conditionals. We focused in particular on Modus Tollens because the three probabilistic theories we investigated make diverging predictions on this inference form. Whereas according to Verschueren et al. (2005), as well as Evans and Over (2004), the acceptance of both MP and MT should be affected by P(q|p) and not by P(¬p|¬q), Oaksford et al. (2000) predict a main effect of the probability of the conclusion, given the minor premise, which for MT translates into P(¬p|¬q). Experiment 3.1 yielded results in accordance with the predictions of Oaksford et al. (2000). In Experiment 3.2 to 3.4, however, we obtained an interaction of both conditional probabilities on the

acceptance of MT, which cannot readily be explained by any of the three theories. To explain these results we suggest a two-step inferential process. The first step consists of evaluating the believability of the conditional premise on the basis of an estimate of P(q|p). If the believability of the conditional is sufficiently high, reasoners proceed to step 2. If it is not, they reject any conclusion from this premise. The second step consists of evaluating the probability of the conclusion, given the minor premise. On a continuous rating scale such as ours, acceptance of the conclusion is rated proportional to that probability. For a categorical decision, reasoners use a threshold as in the first step: If the probability of the conclusion, given the premise, is sufficiently high, the conclusion is accepted, otherwise it is rejected.

In Experiments 3.2, 3.3 and 3.4 we only observed a significant effect of P(conclusion|minor premise) on acceptance of MT when the believability of the conditional was high. When the believability of the conditional was low it did not matter whether the probability of the conclusion, given the minor premise was high. This pattern is exactly what would be expected if the assessment of the conditional probability of the conclusion, given the minor premise, is conducted only if the probability of the conditional premise surpasses a threshold below which no conclusion from that premise is deemed acceptable. Verschueren et al. would not be able to explain the results with the help of a second, slower analytical phase in which people explicitly consider the number of counterexamples, that is, for MT, the p¬q cases. This should lead to the same acceptance in the HH and HL conditions (over all experiments) and the lowest acceptance in the LL condition which wasn't observed. If a probabilistic amendment of the MMT postulates a more likely construction of a certain model according to its presented frequency, the MMT could explain the results of the Experiment 3.1, but not results of the other three. According to the MMT, it would be especially in the conditions 1 and 3, where the ¬p¬q model should be constructed, and since not many counterexamples are available in those conditions, MT should be endorsed quite readily, whereas in the other two conditions endorsement should be low, in condition 2, due to few ¬p¬q models, in condition 4 due to relatively many counterexamples. The interaction effect obtained in Experiments 3.2-3.4 can not be explained by the MMT.

An explanation for why Verschueren and Schaeken (2006) did not find any relation between P(¬p|¬q) and the acceptance of MT may be found in the material used. For natural conditionals P(q|p) and P(¬p|¬q) are usually highly correlated (Verschueren et al. 2006). For instance, in the set of 100 conditionals used for pre-testing in the present experiment, the correlation was still $r = 0.42$, $p<0.001$, although they were preselected to maximize independence in these two dimensions. Moreover, when constructing the conditionals for the pre-test we noticed that it was difficult to come up with a set in which P(¬p|¬q) varies as

much as P(q|p). If the amount of variation was not equated for the two variables in the study of Verschueren and Schaeken (2006), P(¬p|¬q) might have had lower variance, and thereby a lower chance of emerging as a powerful predictor.

The difference in the results on MT between Experiment 3.1 and all the following experiments is best explained by assuming that, in Experiment 3.1, people used a mental short cut to assess MT (and probably also MP), bypassing an evaluation of the conditional premise and estimating the conditional probability of the conclusion, given the minor premise, from the explicitly given frequency information. Assuming that the two-stage process is the default way to solve inferences, the first step is likely to be jumped only if the direct assessment of the conclusion, given the minor premise is made so salient (as e.g. through presenting explicit frequencies) that the default combination of both steps is discharged.

Further implications for the suppositional account of conditionals

As Experiment 3 has shown, acceptance rates for inference tasks (here: MP) are primarily affected by the believability of the conditional as opposed to P(q|p) per sé, which slightly differed in this experiment. Although the two concepts (sufficiency of the conditional and belief in it) are closely related to each other, as has been shown in many studies by now (Evans et al., 2003, Oberauer et al., in press, Oberauer and Wilhelm, 2003), it seems important to distinguish them on a theoretical level: one (sufficiency) is a necessary, but not sufficient condition for the other (belief), as shown with the results on believability of the conditional in Experiment 3.

With the three studies reported here we confirm the view that believability of the conditional (conveyed through P(q|p)) has a fundamental effect on the acceptance of the classic inference tasks. This general effect has been shown by many studies using other means of varying the believability of the major premise, as e.g. with a suppression paradigm (Byrne, 1989, Byrne and Santamaria, 1999) or manipulating expertise status of the speaker (Stevenson and Over, 2001). As Weidenfeld et al. (2005), we manipulated the believability of the conditional by varying P(q|p) in our studies, thus linking them to Verschueren et al.'s account, who investigated the role of perceived sufficiency of the conditional on the acceptance of MP and MT.

Acceptance of MP showed a very close relation to the belief in the conditional in all three experiments, thus confirming the view that a Modus Ponens inference is a special case among the inference tasks, solved by very similar processes as those for assessing the believability of a conditional. This result confirms a suppositional account of conditionals on further grounds than the recent studies concentrating on P(q|p). Whereas in a Modus Ponens task the minor premise p is

given as a fact and the probability that q holds (in our probabilistic version) should be evaluated, there's no such additional information in the judgment of the conditional task. Assuming that people run a Ramsey test when evaluating a conditional sentence, they "hypothetically add p to their stock of knowledge", on which grounds they also assess the probability of q. This procedure should lead them to comparable answers in both tasks, which it did.

Conclusions

Results from comparing three different probabilistic approaches on conditional reasoning led to a new dual stage proposal of conditional reasoning. When frequencies of truth tables were learned in a learning paradigm or had to be retrieved from memory for everyday conditionals, people seem to engage in 2fold process of conditional reasoning. First, they evaluate the believability of the conditional major premise by assessing the conditional probability of q, given p, P(q|p). If this is low, the process is stopped and the conclusion is rejected, if the believability of the conditional is high, they then assess in a second step the probability of the conclusion, given the minor premise. If this is also high, people accept the conclusion, if this is low, they reject it. This two stage process might be bypassed if frequencies for computing the second probability are made so readily available that it invites people to jump the first stage and proceed to the evaluation of the conclusion, given the minor premise, right away.

Chapter 4: Counterexample information: The combination of the probabilistic and mental model based methodological approaches

4.1 Introduction

In the last empirical chapter of this dissertation the role of counterexample information on interpreting conditionals sentences *and* drawing inferences from them is investigated. Specifically, this chapter will compare the roles of different types of counterexample information that have so far been separately used in the two different theoretical accounts introduced in the opening chapters.

As has been reviewed in Chapter 2, according to the probabilistic approach, people interpret a conditional sentence with the usage of the conditional probability of p, given p. For any given conditional (e.g. "If you open the fridge, then the light comes on") this comes down to comparing rule-confirming pq-cases (cases where the fridge was opened and the light went on) to exceptional p¬q- cases (fridge opened and the light did not go on). The higher this ratio is, that is the fewer exceptions there are relative to confirming cases, the higher people's belief in the conditional rule will turn out. As discussed in Chapter 3, where a probabilistic truth table task was employed for the inference tasks, an effect of number of exceptional cases was also found for the acceptance of MP and MT.

In the related field of reasoning with conditional statements, a large amount of research has been conducted to investigate the role of counterexamples on inferences drawn from conditional statements (e.g. Byrne, 1989; Cummins, 1995; DeNeys, Schaeken, & D'Ydewalle, 2003b; Markovits & Potvin, 2001). Byrne (1989; Byrne, Espino, & Santamaria, 1999) was one of the first researchers to systematically investigate effects of counterexamples on inference tasks such as MP and MT. In several experiments she could show that, if a conditional rule is accompanied by a further necessary precondition, the acceptance of otherwise valid inferences declines. For example, if people are presented with the following information:

(1) "If you open the fridge, then the light inside will go on."

(2) "If the light bulb is working, then the light inside will go on."

"Somebody opens the fridge,"

they tend to reject the conclusion "the light inside goes on" (Example for a MP inference) more often than in a condition where the second statement is omitted. Byrne explains this so-called "MP suppression" effect with the enhanced availability of counterexamples for the conditional. In our example this is the

disabling condition of a broken bulb that leads to the p¬q situation (open door, dark fridge) defying MP.

This suppression effect of counterexamples has been widely replicated since (Cummins, 1995; Cummins, Lubart, Alksnis, & Rist, 1991; Markovits & Potvin, 2001; Thompson, 1995). For instance, Cummins (1995) could show that the suppression effect also prevails if disabling conditions are not explicitly mentioned but could be retrieved from everyday experience. DeNeys, Schaeken and d'Ydewalle (2003b) demonstrated that the retrieval of counterexamples is a continuous process in which every additional counterexample lowers the acceptance of inferences proportionally.

To our knowledge there's been only one study so far that tested the effect of disabling conditions on the interpretation of the conditional. Dieussaert, Schaeken and d`Ydevalle (2002) could show that conditionals for which people retrieved many disabling conditions received lower ratings of the conditional probability that the consequence of conditional occurs (the light goes on) given that the antecedents of the conditional holds (door is opened).

A shared feature of the aforementioned studies on the inference suppression and the one study by Dieussaert et al. (2002) is that counterexamples are given to participants or asked from them in the form of *reasons*, *explanations* (Verschueren, Schaeken, & d`Ydewalle, 2005), *conditions* (Weidenfeld, Oberauer, & Hörnig, 2005) or *factors*, (Cummins, 1995; DeNeys, Schaeken, & D'Ydewalle, 2002; 2003b) that allow or cause exceptions to occur. For example, Cummins (1995) asked participants to "please write down as many factors as you can that could make this situation [the p¬q case] possible". The number of different disabling conditions needs to be distinguished from the number of instances that are expected to occur that violate a given conditional. For example, there might be just one disabling condition that can cause a fridge to stay dark when the door is opened (i.e., a broken light bulb), but the probability of this happening might be large when the light bulbs used in fridges are unreliable. We will call an individual instance of p¬q an exception, and a potential cause leading to the occurrence of one or many such instances a disabling condition.

So far, the two different lines of research on interpreting conditionals and reasoning from them have mainly been using one *or* the other type of information, that is, frequency information about p¬q cases in the former case or counterexample information relating to (causal) schemata stored in semantic memory in the latter case.

An author to explicitly distinguish between information about disabling conditions and the frequency of occurrence of p¬q cases in conditional reasoning is Verschueren (Verschueren et al., 2005, see Chapter 3 on reasoning). In her

dual process account of everyday conditional reasoning she differentiates two kinds of information: probabilistic information and counterexample information. She conceptualizes the former as frequency information about instances stored in long term memory that serves as input for a fast, heuristic probability estimation process allowing gradual responses. Counterexample information about disabling conditions is described as general knowledge about causal relations of events, for instance knowledge that broken bulbs don't emit light. This knowledge has to be retrieved analytically in a more effortful way and thus leads to slower judgments. Additionally, authors of the Mental Model theory (e.g. Klaczynski, 2001; Markovits & Quinn, 2002) claim that this analytical process also leads to categorical answers: if a disabling condition is retrieved and a model of a counterexample subsequently constructed, inferences are refuted, otherwise they are accepted.

In their study, Verschueren et al. (2005) obtained probability ratings and disabler ratings for everyday conditionals such as, for instance, "if you water a plant well, then it stays green". Instructions for the probability ratings were: "A plant is well watered, how likely is it that it will stay green?" People had to choose between the answer options: never, almost never, sometimes, most of the time, and always. Ratings of these frequency categories were taken as subjective probability estimates. Instructions for the disabling conditions were: Can a plant be well watered and not stay green? If yes, list one reason. Another group of participants evaluated inference tasks from these conditionals. As the authors could show, results confirm the dual process assumption: when people took longer to evaluate MP and MT, they mainly relied on disabling information, for fast answers they mainly relied on subjective probability ratings.

A weakness of Verschueren et al.'s studies is that the two types of information are usually highly correlated for everyday conditionals: The more different causes there are to prevent q despite the presence of p, the more frequent instances of p¬q will usually be. For any given inference evaluation supposedly based on frequency information, it can never be ruled out that people have come to their answer by thinking of disabling conditions.

Chapter 4 therefore follows multiple goals. First, two types of information that so far have mostly been used in separate studies, although in related fields, are explicitly contrasted while avoiding the potential confound that is present in the Verschueren et al. (2005) study. Second, and maybe even more important, the hitherto mainly unrelated findings on the believability of the conditional obtained with the probabilistic truth table task, and the findings on the suppression effect on acceptance of inferences are integrated. Therefore, we measured people's belief in a conditional *and* their willingness to derive MP and MT inferences from it.

The combination of the two different experimental methods and assessment of all the above mentioned dependent variables allows to test different theoretical predictions at the same time. According to a probabilistic approach people's belief in a conditional should depend on the relative frequency of situations in which antecedent and consequent are both true out of all situations in which the antecedent is true, that is, P(q|p). People's confidence in an inference drawn from this conditional should depend on their belief in the conditional (Evans & Over, 2004), and hence should be influenced by the same frequency information.

The mental model account, in contrast, predicts that the retrieval of a disabling conditions blocks inferences from a conditional, regardless of how many exceptions are caused by this condition. Against this assumption, de Neys et al. (2003b) have shown that people's acceptability ratings of MP and MT inferences gradually decline with the number of disabling conditions that were named for the conditional in a separate study. It cannot be ruled out, however, that when the major premise of an inference problem is a conditional with many disabling conditions, people are simply more likely to retrieve any one of them, leading them to reject the inference. De Neys et al. also consider an extension of the mental model theory in which each disabling condition is represented as a separate model. This modified mental model theory would therefore predict an effect of the number of disabling conditions on inference acceptance ratings. We see no role for frequency information in an inference procedure based on mental models, so the mental model account should not predict an effect of the frequency of exceptions.

Evidence that both types of information might exert an influence on believability of the conditional and the acceptance of inferences comes from Weidenfeld (2005) and Verschueren et al. (2005). According to their dual process assumptions we should expect additive effects of both variables.

A fourth possibility is that solving the two types of task – evaluating the believability of a conditional, and evaluating inferences from it – draw on different sources of information. From all research reported above it seems justified to hypothesise that the analytic nature of drawing inferences might lead people to draw more heavily on information about disabling conditions, whereas the estimation of the believability of a conditional downrightly invites people to give probabilistic estimates based on frequencies. If this is correct, the bulk of research efforts reviewed above was intuitively justified in only testing their side of things. This view predicts that people's judgment of the probability of a conditional should depend on the frequency of exceptions but not the number of disabling conditions; the reverse pattern should be observed for acceptance of MP and MT inferences as dependent variable. A task dependent selection and

processing of information has been shown for different reasoning tasks with conditionals before (Thompson, 2000) and thus does not seem too unlikely.

To distinguish between these four hypotheses, in Experiment 4.1 we combined a probabilistic truth table task introducing explicit frequencies of truth table cases (Oberauer & Wilhelm, 2003) with a suppression paradigm presenting additional disabling conditions. Experiment 4.2 replicates that design with a reduced array of frequency information. Experiment 4.3 tested whether the frequently shown suppression effect can be replicated with our disabling-condition information in the absence of frequency information. Experiment 4.4 extended the findings to conditionals with everyday content, for which the two critical information dimensions were measured by independent ratings. We asked participants to evaluate the believability of the conditional and all four basic inferences from conditional premises, Modus Ponens, Modus Tollens, Acceptance of the Consequent (AC), and Denial of the Antecedent (DA). Based on theoretical assumptions and previous studies, we expect the manipulations of frequency of exceptions and number of disablers to affect only MP and MT; endorsement of DA and AC depend on another kind of counterexamples called *alternative causes* which were not investigated here (for effects of alternative causes on AC and DA see Cummins, 1995; Thompson, 1995, 2000; Verschueren et al., 2005).

4.2 In Experiment 4.1: the probabilistic truth table task and disabling conditions

In Experiment 4.1 we manipulated information about number of exceptional cases and disabling conditions independently. Exceptional cases were introduced through explicit information about the frequencies of the conjunction of p¬q cases, as well as the three other cases of the conditional's truth table, the conjunctions of pq, ¬pq, and ¬p¬q. Different cover stories introduced a conditional statement concerning a population of 2000 instances, which were distributed over the four truth-table cases in varying frequencies. Disabling conditions were given in the form of a conditional sentence introducing a circumstance in which ¬q is possible although p is the case (DeNeys et al., 2003b) The frequencies of exceptions and numbers of disablers assigned in each condition are summarized in Table 10. We will refer to the four conditions as FF, FM, MF, and

Table 10: Experimental manipulation in Experiment 4.1.

	FF	**FM**	**MF**	**MM**
Exceptional cases/ p cases	**100** /1000	**100** /1000	**900** /1800	**900** /1800
Disabling conditions	**0**	**3**	**0**	**3**
Legend: FF= few single exceptions, few (zero) disabling conditions, FM= few single exceptions, many (three) disabling conditions, and so forth.				

MM, with the first letter referring to few or many exceptions, and the second letter referring to few or many disablers. We chose zero as a low number and three as high number of disabling conditions, because DeNeys et al. (2003b) showed that availability of at least three disabling conditions has a reasonably large effect on the acceptance of inference tasks.

4.2.1 Method

Participants

Participants were 27 first-year psychology students from the University of Potsdam (age range: 19-30 years). Order of tasks was varied between subjects, yielding a factor task order with a group of 14 participants that gave probabilistic judgement first and then solved reasoning tasks and 13 participants working on the tasks in reversed order.

Material and Procedure

The experiment was a computerized study realizing the design in Table 1 within subjects. For the probabilistic judgements, participants received eight tasks presented on a monitor, two for each condition in Table 1. Every site presented a fictional short story set on a foreign planet called "Noxus" concerning pseudo scientific circumstances of this planet and contained a conditional statement made by an "expert" of the matter. The cover stories were inspired by Cummins (1995) who used a similar condition in her second experiment to prevent people retrieving disabling conditions from their world knowledge. Besides the story she used in her study, we made up seven more cover stories with completely arbitrary content. Preceding the conditional statement we presented zero or three disabling conditions for the conditional statement and 2000 cases with different numbers of p¬q cases according to Table 10. Here is an example of a cover story (for the condition MM):

A team of biologist examines the different species on Noxus. They focus on genetic relationships between body characteristics such as number of legs and the shape of ears. They found out that, if an animal belongs to the family of grocks then it has 6 legs.

It is also known that:

- *If a grock has a genetic mutation, then it has less than 6 legs.*
- *If a grock is born to lop eared parents, then it has less then 6 legs.*
- *If a grock has run into a pincer trap for rat like animals, then it has less then 6 legs.*

Of 2000 animals that have been examined over the last 6 months, scientists have made the following records:

900 animals that belong to the family of grocks had 6 legs

900 animals that belong to the family of grocks did not have 6 legs

100 animals that do not belong to the family of grocks had 6 legs

100 animals that do not belong to the family of grocks did no have 6 legs."

An expert claims that:
"If the animal belongs to the family of grocks, then it has 6 legs".

Participants were asked to rate the probability that the expert was right on a scale from 0 ("absolutely impossible") to 100 ("absolutely certain").

Either following or preceding the probabilistic judgements of all the conditional statements, participants had to solve the four types of inference tasks for each cover story. Random order of cover stories and conditions was the same as in the probability-judgement part of the experiment. The cover stories and conditionals were presented again, this time followed by MP, MT, DA and AC on one specific animal drawn from the sample. Order of the inference tasks were randomised anew for each cover story. Here is an example for MP:

Expert's statement: "If the animal belongs to the family of grocks, then it has 6 legs".

Observation: The animal belongs to the family of grocks.

Conclusion: It has 6 legs.

Participants had to rate their confidence in the conclusion on a 6 point scale ranging from "certain that I *can* draw the conclusion" to "certain that I c*annot* draw the conclusion".

4.2.2 Results

Since the factor task order did not have a general effect nor interacted with any of the other factors, all data were collapsed over this factor and submitted to a 2 x 2 ANOVA with number of exceptional p¬q-cases (few-many) and number of disabling conditions (few-many) as factors. As in previous chapters, answers for the inference tasks were coded from +5 ("certain that I can draw this conclusion") to -5 ("certain that I cannot draw this conclusion") in steps of 2 to generate equal numerical distances between the six answer options.

Probability of the conditional and reasoning tasks

Results of Experiment 4.1 on the believability judgments of the conditional sentences are shown in Figure 13. The number of exceptional cases had an effect on the probability ratings of the conditional P(cond), $F(1,26) = 23.1$, $p<0.001$, $\eta_p^2 = 0.47$.

Number of exceptions also had an effect on MP, $F(1,26) = 34.0$, $p<0.001$, $\eta_p^2 = 0.50$ and a somewhat smaller effect on MT, $F(1,26) = 17.0$, $p<0.001$, $\eta_p^2 = 0.39$. For MP there was a trend of an interaction between the number of exceptions and disabling conditions $F(1,26) = 3.8$, $p=0.06$, $\eta_p^2 = 0.13$. None of the other effects reached significance, all $F<1$. For AC and DA neither of the two types of information had an effect, all $F<2$.

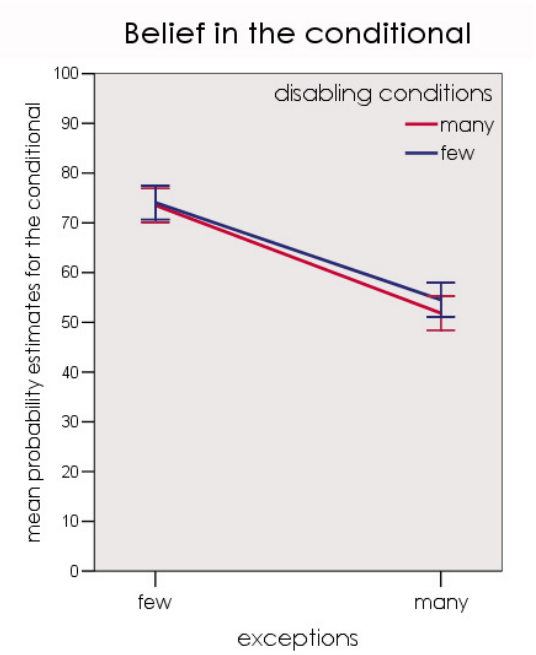

Figure 13: Believability of the conditional
Probability estimates (on a scale from 0 to 100) for the conditional. No of exceptions is grouped on the x-axis, no. of disabler is indicated by different patterns.

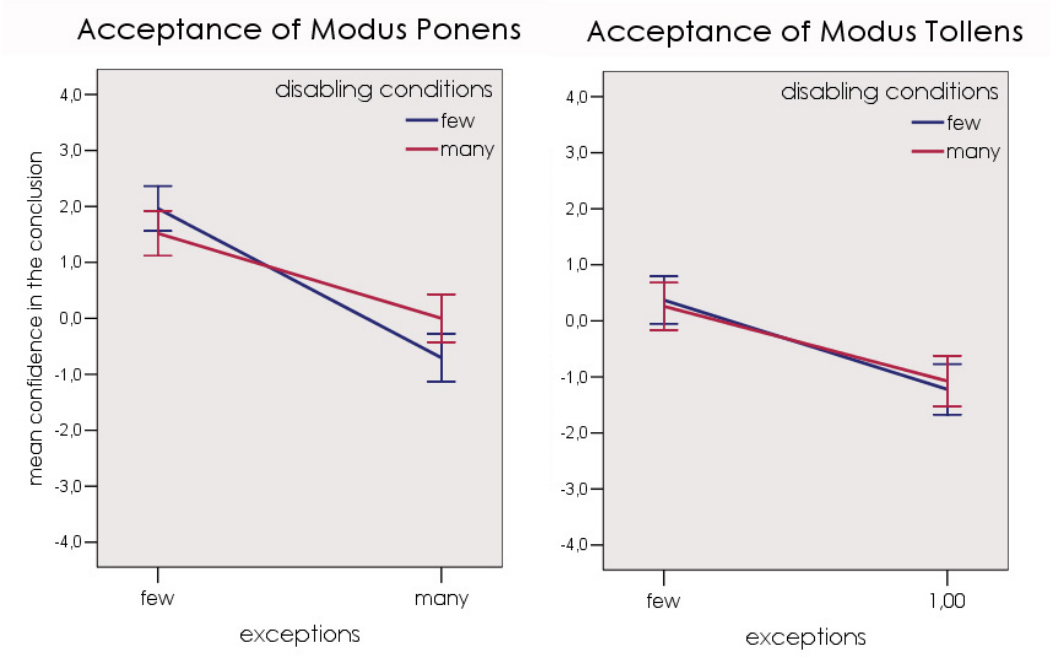

Figure 14: Acceptance of Modus Ponens and Modus Tollens
Confidence ratings in the conclusions (on a scale from -5 to +5).

4.2.3 Discussion

The number of disabling conditions had no effect on any of the dependent variables when additional frequency information was available. The suppression effect of disabling conditions on the inference tasks MP and MT is a very well established effect, (Byrne et al., 1999; Cummins, 1995; DeNeys et al., 2003b; Johnson-Laird, 2001; Quinn & Markovits, 1998) and the availability of disabling conditions has also been shown to reduce the perceived sufficiency of a conditional (Dieussaert et al., 2002) and as a consequence of this, its perceived probability (Weidenfeld et al., 2005). In this experiment these effects of the availability of disabling conditions seem to have been superseded by the processing of frequency information. The notion that the availability of frequency information renders the adherence to disabling information unnecessary has been raised by Weidenfeld (2004), who also found that when combined with frequency information, the availability of disabling conditions no longer affected MP and MT.

One reason for the lack of effects of the disabling conditions could be the way in which the information was presented. The material in Experiment 4.1 stated the disabling conditions as additional information, unrelated to the frequency of p¬q cases in question. To address the issue whether the effect of higher order disabling information would prevail if it was made superordinate to frequency information about single exceptions we conducted a second experiment introducing some minor changes.

4.3 Experiment 4.2: the probabilistic truth table task and disabler: a reduced array version

In Experiment 4.2, we used a reduced version of the probabilistic truth table task. Instead of giving frequency information about all truth table cases, only the overall number of true-antecedent cases (1000) was stated and the number of exceptions (p¬q cases) were directly attributed to one or to three categories of disabling conditions. In this way, the role of disabling conditions was emphasized as causing (or enabling) a certain number of exceptions, whereas overall frequency information about the exceptions had to be indirectly inferred (by adding all p¬q cases attributed to the different disabling conditions and relating them to all p cases). Numbers of exceptions were slightly changed (150 vs. 450) to make them easily divisible by 3 and to keep P(q|p) reasonably close to 0.5 for the many-exceptions condition (Table 11). Number of disabling conditions in the "few"-conditions was changed from *zero* (in Experiment 4.1) to *one* since the existing exceptions always were attributed to some reason. In the condition with only one category of disabling conditions, this category accounted for all (150 vs.

Table 11: Experimental manipulation in Experiment 4.2.

	FF	**FM**	**MF**	**MM**
Exceptional cases/ p cases	**150** /1000	**150** /1000	**450** /1000	**450** /1000
Disabling conditions	**1**	**3**	**1**	**3**

For legend see Table 10.

450) exceptions; in the condition with three categories of disabling conditions, each of them accounted for a third (50 vs. 150) of the exceptions.

The different disabling conditions can be regarded as categories for the exceptions they cause, and therefore the four conditions can be interpreted as two packed conditions (one reason for all exceptions) vs. two unpacked conditions (three reasons for the same amount of exceptions). The notion of packed vs. unpacked descriptions for the same probability was first introduced by Tversky and collaborators in their support theory (Rottenstreich & Tersky, 1997; Tversky & Koehler, 1994). According to support theory, subjective probability judgements should be higher for a description of a situation that is broken down into more specific categories (that is "unpacked"). In our case, it is the probability of exceptions that is unpacked into three different categories. If this probability is judged higher in the conditions with unpacked causes, the ratings for the believability judgements for the conditionals and acceptance of the inferences should be lower in these conditions. Tversky and collaborators would therefore predict that the number of disabling conditions should have an effect even if frequency information is simultaneously available and held constant: the more different causes for the same number of exceptional cases are presented, the lower the believability ratings for the conditional and acceptance of inferences should be.

4.3.1 Method

Participants

Participants were 30 students from the University of Potsdam studying different subjects (age range: 20-31 years). Order of tasks was varied between subjects, yielding a factor task order with a group of 15 participants that gave probabilistic judgement first and then solved reasoning tasks and 15 participants working on the tasks in reversed order.

Material and Procedure

Material and procedure were kept exactly the same as in Experiment 4.1, besides the minor changes in the presented material mentioned in the introductory section. Here is an example of a cover story (for the condition MM):

A team of biologist examines the different species on Noxus. They focus on genetic relationships between body characteristics such as number of legs and the shape of ears. They found out that, if the animal belongs to the family of grocks, then it has 6 legs.

Of 1000 animals of the family of grocks that have been examined, it's also known that:

150 of these grocks do not have 6 legs because of a genetic mutation.

150 of these grocks do not have 6 legs because they have lop eared parents.

150 of these grocks do not have 6 legs because they run into a pincer trap for rat like animals.

The presentation of questions for the believability of the conditional and the 4 inference tasks were kept identical to Experiment 4.1 in form and order.

4.3.2 Results

Since the factor task order did not have a general effect nor interacted with any of the other factors, all data were collapsed over this factor and submitted to a 2 x 2 ANOVA with number of exceptional ($p\neg q$) cases (few-many) and number of disabling conditions (few-many) as factors.

Probability of the conditional and reasoning tasks

The number of exceptions had an effect on the believability ratings of the conditional P(cond), $F(1,29) = 26.8$, $p<0.001$, $\eta_p^2 = 0.48$. Number of disabling conditions did not have an effect on this variable ($F<1$).

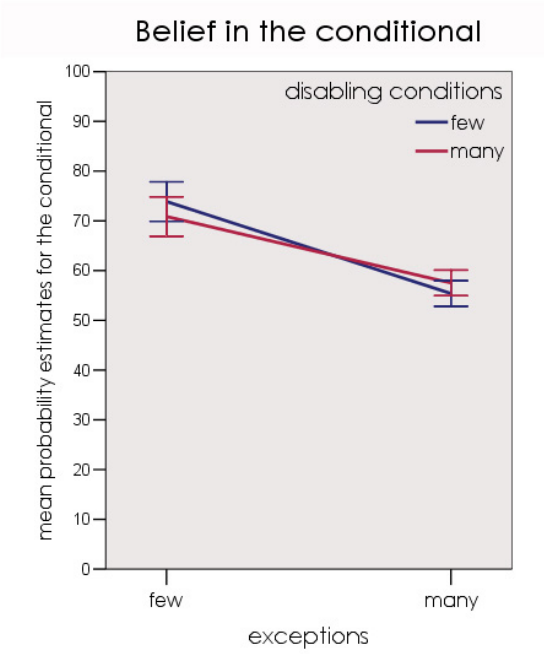

Figure 15: Believability of the conditional
Probability estimates (on a scale from 0 to 100) for the conditional. For legend see Figure 13: Believability of the conditional

The number of exceptions also had an effect on MP, $F(1,29) = 10.3$, $p<0.01$, $\eta_p^2 = 0.26$, but not on MT ($F=1.2$). The number of disabling conditions had a small effect on MT, $F(1,29) = 4.6$, $p<0.05$, $\eta_p^2 = 0.14$. Contrary to general findings on the suppression effect, a higher number of disabling conditions led to a higher acceptance of MT. None of the other effects including those on AC and DA reached significance, all $F< 2.1$.

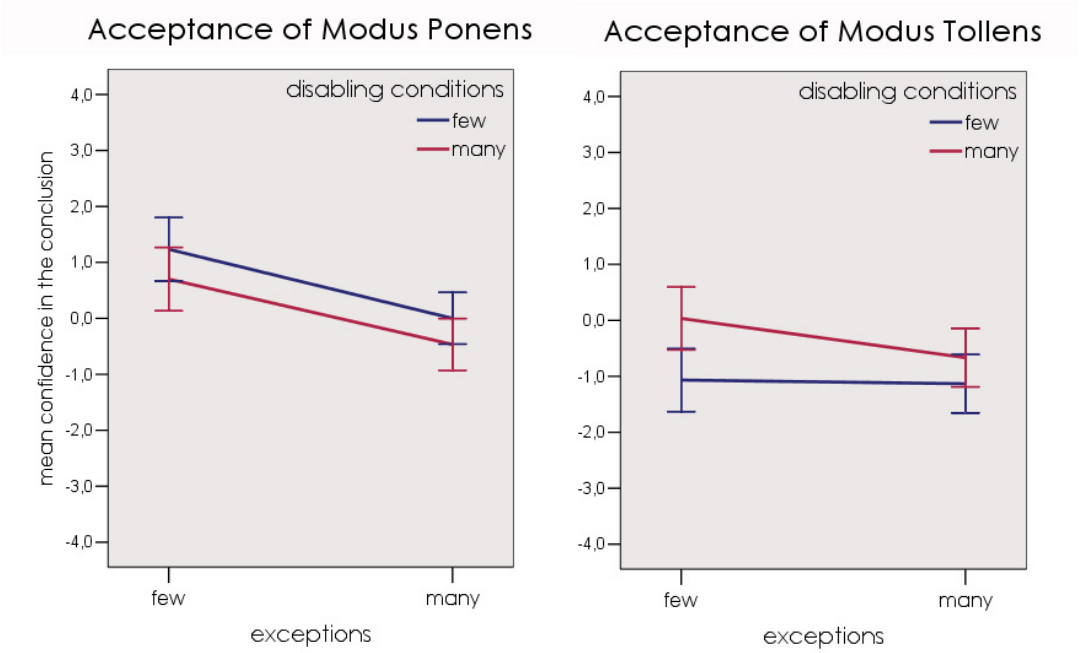

Figure 16: Acceptance of Modus Ponens and Modus Tollens
Confidence ratings in the conclusions (on a scale from -5 to +5).

4.3.3 Discussion

For the belief in the conditional and acceptance of MP we could replicate our findings from Experiment 4.1. Although the role of disabling conditions as causes of the exceptions was emphasized by assigning different numbers of exceptions directly to them, there was no effect of number of disabling conditions. Participants seem to have evaluated their belief in the conditional and the conclusion of MP solely on the frequency information.

Different from Experiment 4.1, the number of exceptions did not have an effect on MT in Experiment 4.2. An explanation for the discrepant results could be that, in Experiment 4.1, participants used a cognitive shortcut to evaluate MT, that is, they determined their confidence in the conclusion of MT by computing $P(\neg q|\neg p)$ from the given frequencies. This procedure to evaluate inferences follows the probabilistic approach of Oaksford & Chater 2001. Support for this shortcut strategy has been obtained in Experiment 3.1 (see Chapter 3) which also showed that the shortcut is used only in tasks where explicit frequencies are

available. In Experiment 4.1 the given frequency information enabled people to use the shortcut strategy, which would yield exactly the results that we obtained. In Experiment 4.2, in contrast, no information about the base rate of ¬q cases was given, so P(¬p|¬q) could not be computed, and participants might therefore have been reluctant to conclude ¬p.

However, the finding that a high number of disabling conditions led to higher MT acceptance (even though only to a small extent) contradicts all former findings in research on the suppression effect. It also directly contradicts predictions of the support theory (Tversky & Koehler, 1994). Unpacking the conditions that lead to a certain probability of exceptions did not lead to a decrease of the probability judgment of the conditional or reduce inference acceptance. Since this counterintuitive finding has not been replicated it will not be discussed here. Again, as in Experiment 4.1, there was no effect of any of the manipulations on AC and DA as would be expected from literature on necessity and sufficiency of conditionals.

4.4 Experiment 4.3: the disabling information only: does it indeed disable?

An important caveat to the results of Experiment 4.1 and 4.2 could consist in the type of disabling conditions used. It could be claimed that disabling conditions did not show the usual suppression effect because the material consisted of artificial reasons for unfamiliar circumstances and thus did not work in general. That is why Experiment 4.3 tests whether the information about disabling conditions used in Experiment 4.1 and 4.2 would yield the expected result of diminishing the belief in the conditional and suppressing the inferences of MP and MT when used without frequency information on exceptions. Therefore participants were exclusively presented with disabling conditions in a third experiment, omitting any frequency information about exceptional cases, and were asked again for their belief in the conditional and their confidence in the four inferences.

4.4.1 Method

Participants

Participants were 22 students from the University of Potsdam studying different subjects (age range: 20 - 27 years). Order of tasks was again varied between subjects, yielding a factor task order with a group of 11 participants that gave probabilistic judgement first and then solved reasoning tasks and 11 participants working on the tasks in reversed order.

Material and Procedure

The experiment was a computerized study realizing the design in Table 3 within subjects. Procedure was kept exactly the same as in Experiment 4.1 and 4.2. The material consisted of the same 8 cover stories, this time without providing participant with any frequency information about exceptional cases. Numbers of disabling conditions used in Experiments 4.1 and 4.2 were varied between 0, 1, 2 and 3. We tested all intermediate levels of disabling information to be able to compare our results to DeNeys et al. (2003b). The overall number of cases (i.e. items that were investigated) was the only frequency information we kept in the cover story to provide participants with some anchor to base their belief on the otherwise completely arbitrary conditionals. The selection of disabling conditions, when less than three were displayed, was randomised for each participant and conditional anew. A complete list of the disabling conditions used is displayed in the Appendix 4.1. Here is an example of a cover story (for condition 4):

A team of biologist examines the different species on Noxus. They focus on genetic relationships between body characteristics as number of legs and the shape of ears. They found out that, if the animal belongs to the family of grocks, then it has 6 legs. The biologists have examined 2000 animals so far.

It's also known that:

- *If a grock has a genetic mutation, then it has less than 6 legs.*
- *If a grock is born to lop eared parents, then it has less then 6 legs.*
- *If a grock has run into a pincer trap for rat like animals, then it has less then 6 legs.*

The presentation of questions for the believability of the conditional and the four inference tasks were kept identical to Experiment 4.1 and 4.2 in form and order.

4.4.2 Results

Since the factor task order did not have a general effect nor interacted with any of the other factors, all data was collapsed over this factor and submitted to a ANOVA with number of disabling conditions (1, 2, 3 vs. 4) as a factor.

Probability of the conditional and reasoning tasks

Figure 17, left panel, shows the effect of number of disabling conditions on participants´ evaluation of the believability of the conditional statement, $F(1,19) = 35.2$, $p<0.001$, $\eta_p^2 = 0.63$. Planned contrasts revealed that any additional disabler up to two significantly lowered the believability ratings (0 vs. 1 disabler: $F(1,21)=21.2$, $p<0.001$, 1 vs. 2 disabler: $F(1,21)=12.0$, $p<0.01$). Introducing the third did not have an additional effect (2 vs. 3 disabler: $F<1$).

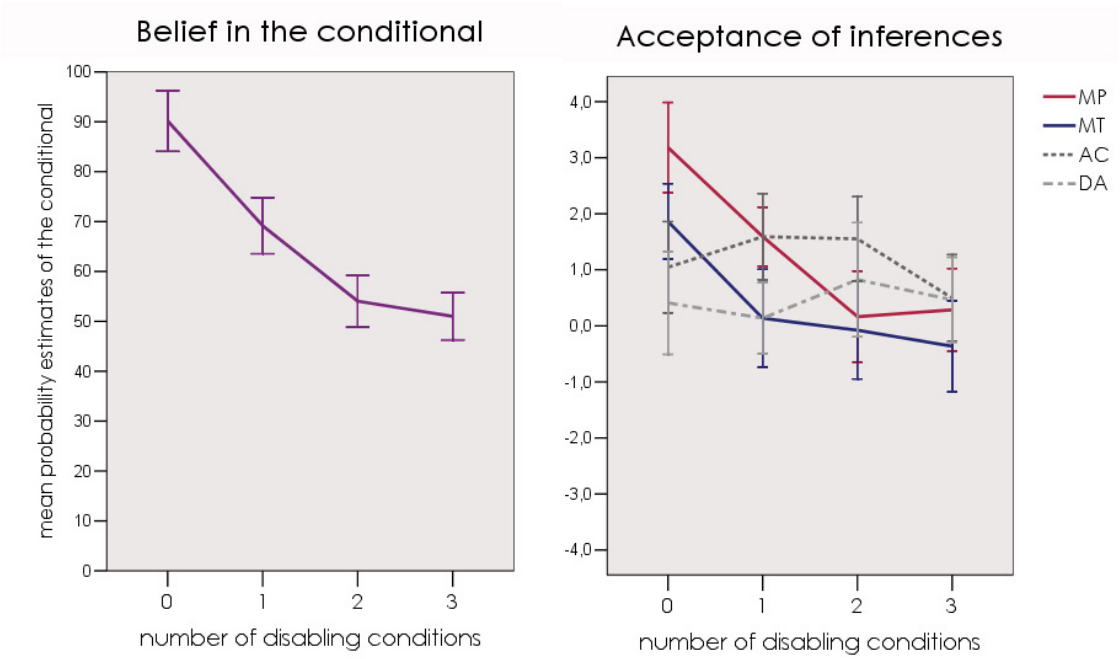

Figure 17: Believability of the conditional and acceptance of inferences
Left: Probability estimates (on a scale from 0 to 100), Right: confidence ratings in the conclusions (on a scale from -5 to +5).

Figure 17, right panel shows confidence ratings of all inferences. The number of disabling conditions had an overall effect on MP, $F(1,19) = 12.1$, $p<0.001$, $\eta_p^2 = 0.37$,. with additional effects for the first two disabler (0 vs. 1 disabler: $F(1,21)= 11.1$, $p<0.01$, 1 vs.2 disabler: $F(1,21)=8.2$, $p<0.01$) and no additional effect for the third disabler (2 vs. 3 disabler: $F <1$). There was also an overall effect of disabler on MT, $F(1,19) = 5.0$, $p<0.01$ $\eta_p^2 = 0.19$. For MT only the first disabler significantly lowered acceptance rates (0 vs. 1 disabler: $F(1,21)= 7.5$, $p<0.05$), the second and the third did not additionally lower inference acceptance compared to the closest condition, both $F<1$. Effects on AC and DA did not reach significance, all $F< 1.5$.

4.4.3 Discussion

In Experiment 4.3 we could show that the disabling information used in Experiment 4.1 and 4.2 did indeed have the expected suppressing effect on belief in the conditional, MP, and MT when presented without frequency information concerning the number of exceptions. This replicates the findings DeNeys et al. (2003b), who employed a similar manipulation of disabling information on inference tasks. Number of disabler in their study were varied between 0 and 4 and concerned everyday content. These authors also report stronger effects for the first two disablers on MP and MT and no effects at all on AC and DA.

Crucially, with Experiment 4.3 we demonstrated that the disabling information used in Experiments 4.1 and 4.2 was effective when used without frequency information on exceptions. The lack of effect of disabling conditions in these experiments can therefore not be attributed to the fictional content of the material.

4.5 Experiment 4.4: every day conditionals

With the fourth study we tested whether our findings would extend to everyday conditionals. It could be that in the previous experiments people used frequency information exclusively because it was explicitly given and therefore very easily available, potentially creating a task demand to use them. If frequency information is used, disabler information is rendered redundant. With everyday conditionals we can test a more natural situation in which both kinds of information must be retrieved from memory. In this situation, frequency information might not be available as readily, and people might draw on causal knowledge about disablers instead, or in addition, to knowledge about frequencies. One could even speculate that frequency information is available in memory only through disabler information because the frequency of exceptions is represented as the frequency (or probability) of the disabling condition causing them.

To test this hypothesis, we conducted a fourth study using conditionals with everyday content. In a pre-test, 85 conditionals were rated for their average number of exceptional cases and number of disabling conditions. We selected 20 conditionals on the basis of these ratings to fill the four cells of the design presented in Table 12.

Table 12: Experimental manipulation in Experiment 4.4.

		Disabling conditions	
		Few	**Many**
Exceptional cases	**Few**	FF: If water is heated to 100°C, then it will boil.	FM If you open the fridge, then the light inside goes on.
	Many	MF: If a horse is white, then it is an albino.	MM If you drink coffee in the evening, then you won't be able to fall asleep.

Legend: FF= few single exceptions, few disabling conditions, few single exceptions, many (three) disabling conditions, MF= many single exceptions, few disabling conditions, MM= many single exceptions, many disabling conditions.

4.5.1 Pretests: the exceptions and the disabler dimension

A set of 85 conditionals with everyday content was constructed that would ensure a range of combinations concerning the two dimensions of interest. The conditionals in this set were rated for the average number of exceptions to the conditional rule and for possible disabling conditions by two different samples of participants.

Pretest A: Exceptions

To ensure additivity of P(pq) and P(p¬q), we asked for frequencies of pq cases and of p¬q-cases. 20 first-year university students answered the following questions about all 85 conditionals. Instructions for the two rating questions were as follows:

> *Consider the following statement: "If you open the fridge, then the light goes on inside."*
> *Please imagine 100 occasions where you open the fridge. How often, in your opinion, does it happen on average out of these 100 occasions, that*
>
> *a) the light goes on inside:*
>
> *b) the light does not go on inside:*
>
> *Please consider that of course, the two numbers have to add to 100.*
> *Estimate the frequency simply on grounds of your everyday experience*.

Examples of an answer scheme were given in the general instructions for this pretest as well as for the pretest B. As a control analysis we computed the sum of both answers and they always added up to 100.

Pretest B: Disabling Conditions

Since the free answer format question for disabling conditions is more time-consuming than the exceptions question, each participant received only 10 conditionals, and the survey was run through the internet to recruit a sufficient number of participants. Two hundred participants answered the questions for disabling conditions for the 85 conditionals. Data produced by people who did not complete the whole questionnaire was nevertheless analysed. The mean number of judgments for each conditional was 22.3 (sd =4.) The mean number of judgement each participant gave was 9.5 (sd =1.5). The instructions were as follows:

> *Consider the following statement: "If you open the fridge, then the light goes on inside."*
>
> *Can you imagine circumstances, in which the following situation is possible: you open the fridge and the light inside does not go on.*
>
> *o yes o no*
>
> *If yes, please give as many reasons for this situation in the table below, as you can think of (six empty text input slots were provided).*
>
> *Please give as many reasons as you think are plausible simply on grounds of your everyday experience. If you can think of less than 6 reasons, leave the rest of the fields empty."*

Two independent raters scored the disabling conditions given in a free-answer format, and eliminated all responses that were judged to be unrealistic items, or to be variations of one single idea.

On the basis of the two ratings five conditionals were chosen to fit in each cell of the 2 x 2 design of few vs. many number of exceptions and number of disabling conditions. Conditionals in each cell were chosen if their score was at above the 60 percentile or below the 40 percentile line. Cell means were calibrated to a roughly equal distance to the overall median on each dimension; for instance, the mean of the "few disabler" condition (1.2) differed from the median rating of 1.9 by the same absolute amount as the mean in the "many disabler" condition (2.6). Ratings for all 85 conditionals are shown in Appendix 4.2, the ratings for the 20 conditionals used in Experiment 4.4 are listed separately in Appendix 4.3.

4.5.2 Method main study

Participants

Participants were 30 last-year high school and first-year university students studying different subjects (age range: 17 - 23 years). Order of tasks was varied between subjects such that 15 participants gave probabilistic judgements first and then solved reasoning tasks and 15 participants worked on the tasks in reversed order.

Material and Procedure

The experiment was a computer based study realizing the design in Table 4 within subjects. The material participants saw consisted of the conditional sentence only. Each conditional and the inference tasks were presented on a separate page, respectively. The procedure used to assess the dependent measures was the same as in Experiment 4.1 - 4.3.

4.5.3 Results

Since everyday conditionals could vary in other untested dimensions, we refrain from analysing the answers of AC and DA. Variables with well known influences on these inferences (mainly: necessity and availability of alternative antecedents) have not been established for the conditionals used in Experiment 4.4. Data of the remaining three dependent variables (belief in the conditionals, MP and MT) were submitted to 2 x 2 x 2 ANOVAs with task order (probability judgements first – last), number of exceptions (few – many) and number of disabling conditions (few - many) as factors. Results for belief in the conditional and reasoning tasks are reported separately.

Probability of the Conditional

Figure 18 shows participants´ evaluation of the believability of the conditional statement. Number of exceptions had a large effect with $F(1,28) = 288.6$, $p<0.001$, $\eta_p^2 = 0.91$. Number of disabling conditions had a smaller effect on the believability of the conditional with $F(1,28) = 8.1$, $p<0.01$, $\eta_p^2 = 0.22$. Unexpectedly, conditionals with many disabling conditions were rated

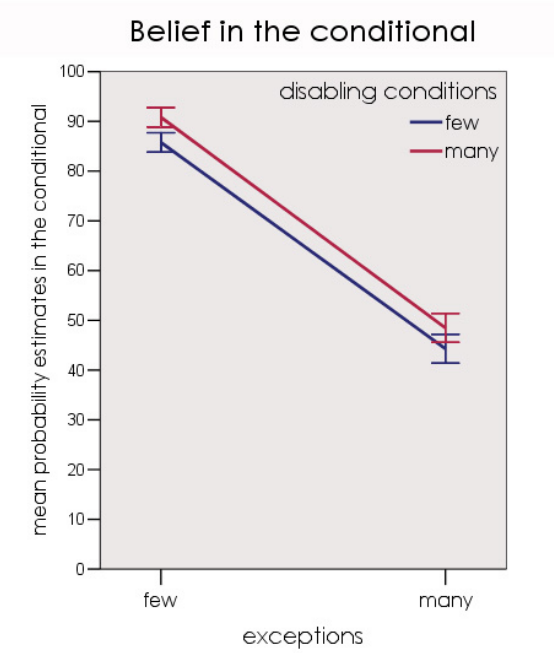

Figure 18: Believability of the conditional
Probability estimates (on a scale from 0 to 100) for the conditional. For legend see Figure 13.

higher in their believability than conditionals with few disabling conditions. Task order also had a main effect on believability of the conditional. When these judgements were given first, the believability was rated lower, $F(1,28) = 14.6$, $p=0.01$, $\eta_p^2 = 0.34$.

Different from previous experiments, task order interacted with both other factors. The influence of exceptions as well as of disabling conditions was higher when probability judgements were given first. There was a small interaction effect of task order and exceptions, $F(1,28) = 7.3$, $p<0.05$, $\eta_p^2 = 0.21$ such that the suppression effect of exceptions was higher when the conditional was judged first. The interaction of task order and disabling condition also reached significance, $F(1,28) = 6.4$, $p<0.05$, $\eta_p^2 = .0.19$, such that more disabling conditions led to higher believability ratings only when the conditional was judged first, $t(1,14) = -3.18$, $p<0.01$. The interaction of all three factors was significant as well, $F(1,28) = 6.3$, $p<0.05$, $\eta_p^2 = 0.18$.

Reasoning tasks

Figure 19, left panel, shows participants´ confidence in MP. Number of exceptions again had the largest effect with $F(1,28) = 42.2$, $p<0.001$, $\eta_p^2 = 0.60$. There was a non-significant trend for MP to be endorsed more when fewer disablers were available, $F(1,28) = 3.2$, $p=0.09$, $\eta_p^2 = 0.10$. For the interaction of both factors we also observed a trend that just missed the conventional criterion of significance, $F(1,28) = 3.2$, $p=0.06$, $\eta_p^2 = 0.12$. Number of disabler seemed to have a suppressing effect only when number of exceptions was high.

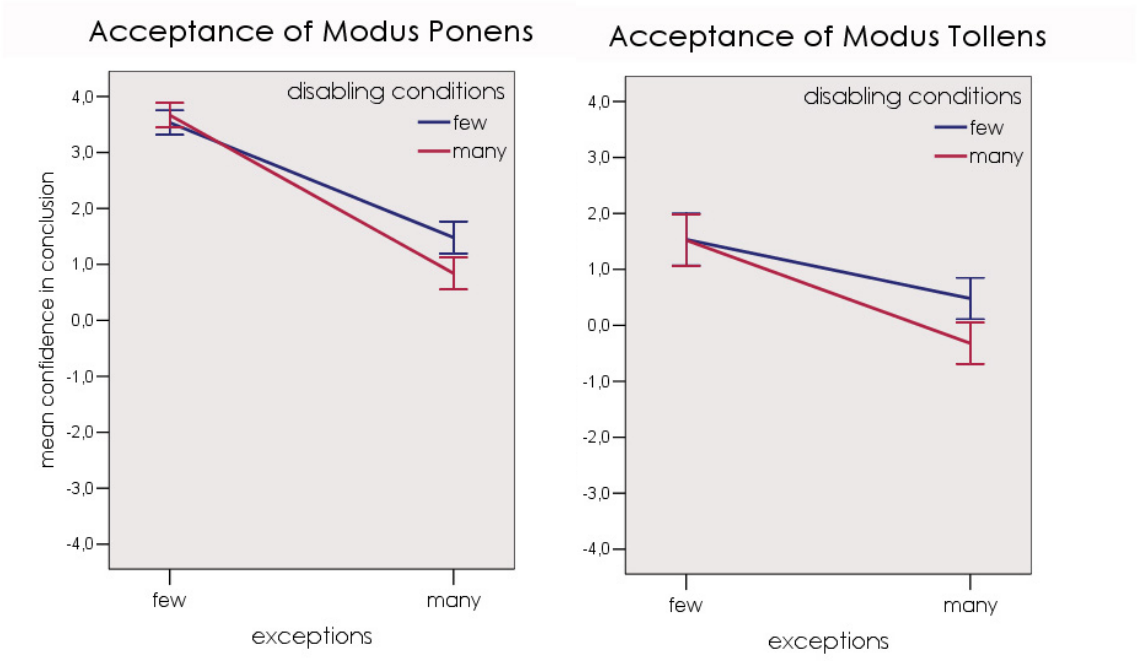

Figure 19: Acceptance of Modus Ponens and Modus Tollens
Confidence ratings in the conclusions (on a scale from -5 to +5).

Task order had a main effect on MP; when probability judgements were given first, MP was endorsed more hesitantly, $F(1,28) = 4.4$, $p<0.05$, $\eta_p^2 = 0.14$. Moreover, task order interacted with number of disabling conditions, $F(1,28) = 6.8$, $p<0.05$, $\eta_p^2 = 0.20$; disabling conditions only had a suppressing effect when inferences were evaluated first.

Figure 19, right panel, shows participants´ confidence in MT. Number of exceptions again had the largest effect on this inference with $F(1,28) = 31.0$, $p<0.001$, $\eta_p^2 = 0.53$. Number of disabling conditions did not have a general effect on MT, $F(1,28) = 2.8$, $p>0.1$ but interacted with task order, $F(1,28) = 8.2$, $p<0.01$, $\eta_p^2 = 0.23$. A suppressing effect of disabling conditions only showed in the group that evaluated inferences first, $t(1,14)= 2.76$, $p<0.05$. There was no effect in the group that judged the conditionals first, $t(1,14) = -1.0$, $p= 0.3$. None of the other factors reached significance, all $F< 1.6$.

4.5.4 Discussion

Experiment 4.4 orthogonally varied the number of disabling conditions and the relative frequency of exceptions, drawing on people's knowledge about conditionals with everyday contents. In the main experiment, both kinds of information had to be retrieved from memory, thereby avoiding any demand characteristic by presenting information in the task context. People could base their judgments about the believability of the conditional, and about the inferences from it That is why in Experiment 4.3 it is tested, whether the information about disabling conditions used in these experiments would yield the expected result of diminishing the belief in the conditional and suppressing the inferences of MP and MT when used without frequency information on exceptions. on one information or the other or a combination of both.

As in the previous experiments, participants strongly favoured the number of exceptional cases over the number of different disabling conditions to evaluate their belief in a conditional statement and their confidence in MP and MT. The large suppressing effect of frequency of exceptions on these three variables could be replicated. Similar to reasoning with arbitrary conditionals that have so far been used most often in probabilistic truth table tasks (Experiment 4.1 and 4.2), frequency information seems to have a dominating role in evaluation of and reasoning with everyday conditionals as well.

Number of disabling conditions only exerted a minor effect on belief in the conditional, in an unpredicted way of leading to a higher belief when there were more disabling conditions available. This paradoxical effect might be due to specific characteristics of the 20 conditional sentences used in Experiment 4.4 that had an untested impact on the believability of those sentences. For MP and MT, disabling conditions only showed a trend towards a suppression effect, in the

case of MT only when reasoning tasks were evaluated first. Taking these results together it seems justified to assume that for everyday conditionals as much as for arbitrary conditionals people mainly rely on frequency information for their evaluation of the believability of a conditional, and their evaluation of MP and MT inferences from that conditional.

This main effect was modulated by several small interaction effects of the experimental manipulations with task order. When the believability judgement for the conditional had to be given first, number of exceptions had a larger effect on this judgement compared to when it was given second. On the other hand, when reasoning tasks were solved first, number of disabling conditions had a larger effect (respectively some small effect at all) on acceptance of MP and MT.

The present results seem to contradict those from the study of Weidenfeld, Oberauer, and Hörnig (2005). They investigated correlations across conditionals between the number of disabling conditions people could think of, their degree of belief in a conditional, their estimates of P(q|p), and endorsement rates for MP and MT. For their "inductive" inference condition, which is most comparable to the instructions used here, they found that the number of disabling conditions had an equally strong direct effect on the rates of endorsement for MP and MT as the belief in the conditional and P(q|p).

In the present study, we found a different overall pattern of effects. Across the 20 conditionals, mean subjective P(q|p) values (calculated from the ratings of the frequency of q, given p) strongly correlated with mean endorsement of MP ($r = .90$, $p < .01$) and MT ($r = .59$, $p < .01$), whereas the number of disabling conditions was not correlated with neither MP nor MT ($p>.18$). Since in the Weidenfeld et al. study participants who rated the inferences did not give any other ratings at all, their results are best compared with our group that rated the inferences first. In that group, we found a small, but significant suppression effect of the number of disabling conditions for both MP and MT. The discrepancy between the present findings and those of Weidenfeld et al. (2005) therefore are more quantitative in nature than qualitative – whereas they found that disabling condition had an equally strong effect on inference endorsement as P(q|p), we found disabling conditions to have a smaller effect than P(q|p).

This discrepancy might be due to methodological differences in the two studies. Whereas Weidenfeld et al. used pseudo-naturalistic conditionals with fictional contents that could only indirectly be related to world knowledge, we used conditionals that bore directly on people's knowledge. Moreover, Weidenfeld et al. made no effort to deconfound P(q|p) and the number of disabling conditions, so the stronger effect of disabling conditions in their study might be explained by confounds with the P(q|p) measure. A third difference regards the response format for the inference tasks. Weidenfeld et al. asked participants to make a categorical judgment of accepting or rejecting the

conclusion, whereas we asked them to rate their confidence in the conclusion on a continuous scale. The continuous scale could have motivated participants to frame the task as one of probability estimation, for which it seems rational to search for information about the relative frequencies of cases in which the conclusion holds. When, however, the task is to decide whether a conclusion can or cannot be drawn from a set of premises, it seems rational to search for reasons for refuting the conclusion, namely disabling conditions.

4.6 General discussion of Experiments 4.1-4.4

There is a wide range of evidence that the availability of counterexample information suppresses otherwise endorsed inferences (Byrne, 1989; Byrne et al. 1999, Cummins 1995, DeNeys et al. 2003a, 2003b; Quinn & Markovits, 1998) and lowers the perceived sufficiency of a conditional sentence (Dieussaert et al., 2002). This counterexample information is usually conceptualised as causal information that expresses a condition or factor that prevents the consequent of a conditional to occur although the antecedent is given.

Another, more recent approach to the understanding of conditionals advocates a probabilistic view, according to which conditionals are evaluated by comparing rule confirming instances of the pq conjunction with exceptional instances of the p¬q conjunction (Evans & Over, 2004, Evans et a. 2003, Oberauer et al., in press). In this line of research a probabilistic truth table task has been established that presents explicit frequency information about exceptional cases. One of the major findings with this paradigm is that the belief in a conditional is highly dependent on the ratio of pq-cases to p¬q-cases. A corresponding effect of explicit frequency information on the inference tasks MP and MT could be demonstrated in a recent study (Geiger & Oberauer, submitted).

One author to systematically investigate these two different types of information is Verschueren et al. (2005), who found that frequency information as well as counterexample information that can be retrieved from memory exert a suppressing effect on inferences, depending on the speed of judgments and contextual factors (e.g. association strength of counterexamples).

The studies reported here took a similar approach. Experiment 4.1 and 4.2 combined explicitly mentioned frequency information (as used in the probabilistic approach) with explicitly mentioned counterexample information (as used in studies on suppression effects). In both experiments the effect of frequency information about exceptions outplayed the influence of information about categories of disablers. As Experiment 4.3 established, this was not due to specificities of the materials used in Experiment 4.1 and 4.2. The last experiment largely replicated the findings of the first two experiments using a carefully

selected set of everyday conditionals across which the number of disablers and the frequency of exceptional cases was varied orthogonally. This independent manipulation is important since the two dimensions are usually highly correlated (i.e. the more disabling conditions there are to prevent q in the presence of p, the more exceptional p¬q cases there will usually be). Choosing sentences that varied independently on these two dimensions, we could isolate the effects of any one dimension. The results unambiguously show that people give priority to frequency information over causal information about the number of different disabling conditions.

Why do people prefer frequency information over disabling conditions although the latter could be more informative, for example, in giving us reasons for why exception occur or similar circumstances to expect exceptions? In a probabilistic framework relying exclusively on frequency information is rational. According to this view, the probability of the conditional is a function of P(q|p), which in turn depends on the relative frequency of exceptions regardless of the number of disabling conditions by which these exceptions were caused. MP and MT in turn depend on the believability of the major conditional premise and are therefore likewise affected by the relative frequencies of exceptions. In a probabilistic framework there is no reason to take the number of disabling conditions into account, if the believability of a conditional can be evaluated via frequency information. Previous findings showing that endorsement of inferences from conditionals is blocked by the availability of disabling conditions are probably mediated by people's beliefs about the relative frequencies of exceptions: In the absence of independent information about the frequency of exceptions, the number of different possible causes of such exceptions is a good estimate of their probability of occurrence. Moreover, in nonselected samples of everyday conditionals people's beliefs about the relative frequency of exceptions is highly correlated with the availability of disabling conditions.

The present results have implications for theories of how people reason from conditional premises. It has been well documented that content and context modulate people's willingness to accept even the logically valid inferences MP and MT. In the mental model framework, this so-called inference suppression effect has been explained by assuming that people can retrieve models of counterexamples to the conditional premise, that is, models of the p¬q conjunction. Retrieving a single such model should be sufficient to block endorsement of MP and MT, because if p¬q is represented as a possibility, it constitutes a counterexample to the conclusions of MP and MT, and hence these conclusions are not supported by the set of mental models of the situation described by the premises. Against this assumption, De Neys et al. (2003b) have shown that people's degree of endorsement of MP and MT declines linearly with

every additional counterexample that is presented to them or that they could retrieve for the conditional premise in a separate part of the experiment.

The results here add further evidence against the mental models account of the inference suppression effect. Whether a mental model of the p¬q conjunction is constructed should depend on whether people think of that conjunction as a possibility, and not on how frequently people believe this conjunction occurs. The theory of mental models allows for attaching numerical information about frequencies or probabilities to mental models (Johnson-Laird, Legrenzi, Girotto, Legrenzi, Caverni, 1999), but they play a role only in probability estimation tasks, not in conditional inference. The only way in which the model theory could explain why beliefs about the frequency of exceptions affects endorsement of MP and MT is by framing these inference tasks as probability estimation tasks. A probabilistic model theory of reasoning with conditionals could assume that people assign probabilities to each mental model and estimate the probability of the conclusion from these values. For instance, in an MP argument the minor premise "p" serves to reduce the set of models to those involving p, that is, the pq model and the p¬q model. People might assign a probability of .8 to the pq model and probability of .2 to the p¬q model, and from that infer a degree of belief in the conclusion q of .8. A probabilistic adaptation of the model theory is not implausible in light of the fact that we asked participants to evaluate conclusions on a continuous scale rather than to make categorical judgments of acceptance or rejection. Of course, such an adaptation would bring the model theory very close to probabilistic theories of reasoning with conditionals.

The findings here are consistent with probabilistic theories of the interpretation of and reasoning from conditionals. These theories agree that people's degree of belief in a conditional is determined by their subjective conditional probability of the consequent, given the antecedent. The effect of the relative frequency of exceptions on belief in the conditional confirms this assumption.

For inferences people draw from conditionals, probabilistic theories differ in their assumptions (see Chapter 3). The present data don't distinguish between the different views – they all predict that, in the present design, the relative frequency of exceptions should influence how readily people endorse MP and MT, and the results for these tasks are comparable with those of Chapter 3.

According to Verschueren et al.'s (2005) dual-process account, inferences from conditional premises are evaluated through two processes. A fast, heuristic process draws on knowledge about the frequency or probability of exceptions to assess the believability of the conditional, and evaluates the conclusion accordingly. A slower, analytical process draws on causal knowledge about disabling conditions and operates according to the assumptions of the mental

model theory (Markovits & Barouillet, 2002; Schroyens, Schaeken & Handley, 2003). The dual-process theory predicts that fast judgments on conditional inferences should be affected mostly by frequency information and slow judgments should be affected mostly by information about disabling conditions. In all Experiments here participants made their judgments without time pressure, so our results should reflect a mixture of fast and slow judgments. The theory of Verschueren et al. (2005) should therefore predict that both the frequency of exceptions and the number of disabling conditions should have effects on people's endorsement of MP and MT. This was not what we found. The dual-process account of Verschueren et al. (2005) could explain our data only with the additional assumption that participants in our experiments relied nearly exclusively on the fast, heuristic process. With this assumption, however, the theory reduces to a purely probabilistic theory.

Conclusions

In four experiments we unambiguously showed that people establish their belief in a conditional and their confidence in conclusions drawn from it depending on the relative frequencies of confirming and exceptional cases, regardless of number of disabling cases that cause these exceptions. The findings conform with probabilistic theories of conditional reasoning and introduce an alternative explanation for the well proven suppression effect on inferences. Instead of blocking inferences directly by introducing a counterexample, the suppression effect might be well caused by the number of exceptional cases to a conditional rule that lower its believability. The current formulation of the Mental Model Theory without a probabilistic annotation can not explain the gradual effect of exceptions on the inference tasks. We did not find evidence for an additional search process for counterexamples above and beyond probabilistic estimations, which casts doubt on the necessity of the analytic process assumed by dual process theories on reasoning.

Chapter 5: A comprehensive probabilistic approach on conditionals?

This dissertation explored the potential of a probabilistic approach of explaining the way people interpret, assign believability to, and reason from conditional statements. In Chapter 2, further refinements of the suppositional account of interpreting conditionals were investigated. Chapter 3 explored the role of probabilities in drawing conclusions from conditionals, and Chapter 4 investigated different types of information feeding into these processes. A large body of this research was conducted with "basic conditionals" (Johnson-Laird & Byrne, 2002), that is conditionals with either an arbitrary ("If the card is spades, then it is a nine") or fictional content ("If the flying object has invisible wings, then it has two jet propulsions") that cannot be related to any former knowledge. Very importantly though, the findings on conditional reasoning (Chapter 3) and the findings on different types of information (Chapter 4) were further justified by one experiment each, using conditionals taken from an everyday conversational context. These conditionals (Experiment 3.4 and 4.4.) were rated in different dimensions and selected according to specific features of interest. Although these conditionals might have many disadvantages (potential confounds with other, unknown features, artificial-seeming or at least odd, see Appendix 3.5 and 4.5), they nevertheless provide an important link between laboratory findings on quite abstract tasks to effects that might be found in outside-the-laboratory, real-world thinking processes. In all the different experiments presented here using different material, results converge on a comprehensive probabilistic view on conditionals: a probabilistic way of interpreting conditionals and assigning believability to them and a probabilistic way of reasoning from them.

5.1 Probabilities – what can they explain?

As found in a large body of research reviewed in Chapter 1 and 2, all experiments reported in this dissertation confirm the major role of P(q|p) on the believability of the conditional. They all made use of frequency information on truth table cases from which this probability could be and was derived.

As Figure 20 illustrates with bold arrows, frequency information is the main information source that not only allows the evaluation of the conditional probability of q given p, for evaluation of the believability of the conditional, it is also the main source for deriving judgments about the probability of a conclusion, given a minor premise for the evaluation of inferences from a conditional. For MP, this conditional probability happens to be the same, but for MT it is different. As established with three experiments in Chapter 3, in conditional reasoning processes people first evaluate the believability of the

conditional (the major premise) via P(q|p) and if this is high, they evaluate the conditional probability P(conclusion | minor premise). Acceptance of inferences is a product of both these processes, unless explicit frequency information (as used in the standard version of the probabilistic truth table task) invites people to omit the first step and accept inferences as a result of the second step straight away.

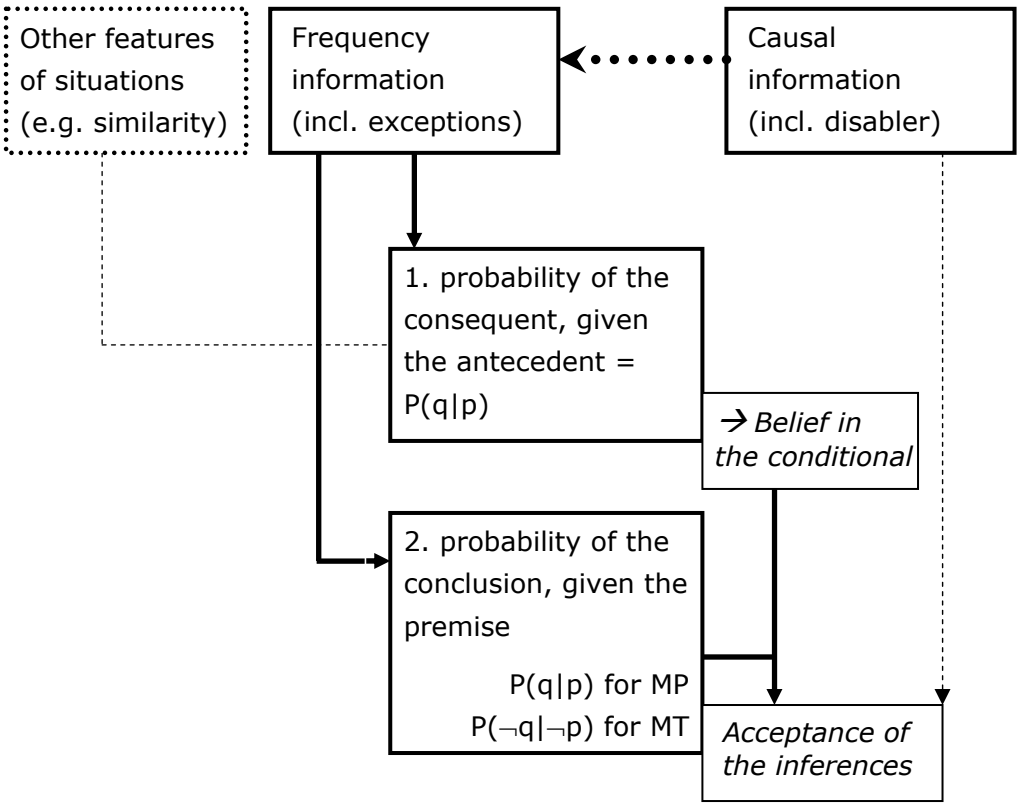

Figure 20: An extended theoretical model of interpretation and reasoning from conditionals:
Frequency information about situations in the world is used to compute conditional probabilities (here: P(q|p)), that in turn are used to derive the degree of belief people are willing to assign to a conditional. If people are asked to draw inferences from the conditional, they additionally compute the probability of the conclusion, given the minor premise in a second step. Both probabilities combined lead to the acceptance or rejection of inferences. Information on causal circumstances are only conceptually included (as causing situations in the world), they are not involved in the processes leading to believability judgments and only marginally affect inference acceptance. The opposite goes for feature information: it has a small effect on believability judgments, but no direct effect on the inferences.

Causative information on disabling conditions did not have a direct influence on the evaluation of the conditional and only a very marginal influence on inference acceptance, as shown in Chapter 4 of this thesis. In so far that disabling conditions are the cause of exceptions (e.g., broken light bulbs for dark fridges) these two measures are usually highly correlated for everyday conditionals and have been deliberately disassociated in the Experiments of Chapter 4. With this artificial disassociation an effect of each of the variables should be detected, if present. Unlike in preceding research (e.g. Weidenfeld et al., 2005), not much evidence for a direct effect of disabling conditions on drawing the inferences MP and MT was found here. Only a small trend of taking causative information into account was found, when inferences were evaluated with no former other tasks. From experiments in Chapter 4, where the two types of information were disentangled, it can be inferred that previous research that found direct effects of disabling conditions did so as a result of a residual confound between frequency information and causative information.

Besides frequency of situations and their causal circumstances, this dissertation showed how another characteristic, namely similarity of situation, only exert a small effect on the believability of conditionals.

Reviewing all the results in sum, it seems that conditional sentences are treated as expressing a certain conditional probability and that inferences from this conditional are accepted according to this probability as well. Both of these probability estimates seem to be almost exclusively based on frequency information with little attention given to further information on specific situations. None of the results on inferences tasks require an analytic, dichotomous validation process as proposed by dual process authors relying on assumptions of the MMT (e.g. Verschueren et al. 2005; Schroyens, Schaeken & Handley, 2003). Results of Chapter 3 can in fact not be explained by assumptions of the MMT. From these consistent findings over 10 Experiments it seems justified to conclude that judging the believability of a conditional *and* drawing inferences from them can be explained in probabilistic terms without the need of an analytical process, whether it be based on mental models or not.

For judging the believability of the example conditional "If it is a sunny day on the weekend, I´ll go to the beach" people obviously do this by thinking of sunny days and compare the frequency of those where they ended up on the beach with those where they did other things. For this conditional they will probably end up at a fairly high believability. To arrive at this judgment, they firstly do *not* take more features of the days into account, once they have made up their mind what counts as a sunny weekend day (i.e. is part of the relevant set) and secondly, do *not* base this judgment on further disabling causal information (such as pressing work, sickness or beach-shy visitors). The same way, when evaluating an inference from this conditional and given a minor

premise "I am not at the beach" (for MT), they evaluate the conditional probability of "then it is not a sunny weekend day", given this minor premise, by simply comparing frequencies of days spent somewhere else, according to whether those were sunny weekend days or not. Realizing that most of these days were busy weekdays (and not sunny weekend days) in conjunction with a high believability of the conditional, they will accept the conclusion of the MT without trying to retrieve potential disabling causes (such as pressing work, see above).

A caveat of this quite general conclusion consists of the fact that in all our experiments we phrased the answer option for the judgment of the inferences in a gradual way ("how sure are you, that you can draw this conclusion"). This might have invited participants to treat the inference task much more as a probability judgment than they would have done with an instruction stressing logical validity (cf. Discussion 4.4). More evidence that inference tasks are treated less as a gradual estimation than believability judgments of the conditional comes from the persistent finding of higher variances of the inference tasks compared to the believability judgments (stemming from more answers completely rejecting the inference, compared to zero-probability answers for the conditional, see larger error bars for the inference tasks in all experiments).

A purely probabilistic view on conditionals, especially drawing inferences from them, might be confined to situations where people are not pressed into drawing logically sound inferences and when it is left for them to decide what exactly they think licenses a conclusion. In all experiments reported here, as mentioned above, the instructions were rather vague regarding the logical soundness ("being able to draw a conclusion") and made no use of words like "necessary" or "beyond doubt". As other research has shown, people seem to have deductive ways of dealing with inferential tasks and will employ this way if required by the instructions they are given (see Markovits & Handley, 2005; Rips, 2001). Even when using gradual response options and liberal instruction like the ones used here, DeNeys (2003) obtained stepwise answers with a clear cut between acceptance and rejection for a minority of people, hinting on a logical evaluation strategy. As Oberauer (2006b) concludes, trying to explain a range of conditional reasoning tasks (basic vs. pseudonatural with varying causal content, yes/no forced choice responses) requires either a mechanism based on mental models or assumptions based on dual process theories on reasoning. Neither of the two mechanisms was found to be necessary to explain the results presented here, which might be due to the liberal, non-strictly deductive nature of reasoning tasks used.

Other potential task demands besides logical necessity might also lead to the consideration of causative information, as for example dealing with contradictory information or generalisation to new situations, where relying on pure frequency information might fall short of giving the most desirable answer. So in a sense, the experimental settings used in this dissertation consist of the specific situation, in which rather casual judgments were allowed, being liberal did not cause any disadvantages, and no further requirements besides what is sensible to believe and conclude were present.

5.2 Open research questions

The two-stage probabilistic reasoning process was proposed for results on MP and MT, respectively. Although for the specific case of MP, the two steps fall into one, since they both include calculating the conditional probability P(q|p), they are quite distinct for all three other inference patterns. It was beyond the scope of this dissertation to extend the probabilistic two-stage idea of conditional reasoning to the inference patterns of AC and DA. Since Oaksford & Chater (2001), Verschueren et al. (2005) and the outline given here in Chapter 3 allow deriving precise and differing predictions, there are straightforward tests that could be implemented through using a probabilistic truth table task. According to the two-stage proposal, there are two ways to block inferences: everything that lowers the belief in a conditional should lower the acceptance of all four inferences, at least to some extent. Nevertheless each inference can be blocked by lowering the probability of the conclusion, given the minor premise as well. For AC and DA the well proven negative effect of alternative causes could be attributed to the second stage. Suppressing effects of disabling conditions (or p¬q cases) on acceptance of AC and DA on the other hand could be read as evidence of affecting the first stage of conditional reasoning. These effects on AC and DA were not found in any of the experiments here (nor in the studies of e.g. Byrne, 1989; Cummins, 1995 or Thompson 1995, for contrasting results see George, 1997) which speaks against the generalisation of the two-stage reasoning idea on all four of the inference forms.

Relating to this question of whether a two stage inference process is applicable to AC and DA as well, it has not been established whether the relationship of frequency information to causative information generalizes to alternative causes and the frequency of ¬pq cases as well. There is no theoretical evidence as to why there should be differences for the dimension that describes the necessity rather then the sufficiency of conditionals. Analogous to the experiments in Chapter 4, it would be desirable to investigate effects of alternative causes, and whether the inferences of AC and DA can be exclusively explained with probabilistic thought processes as well. Applied to the example above, given that "I am at the beach" people might infer a high probability of "it

is a sunny weekend day" because they can think of a lot of days with both of these properties compared to beach days that weren't sunny weekend days. They would then accept the AC conclusion without reviewing plausible alternative causes such as the need for shells, which might force us to the beach on a windy, rainy day.

5.3 In closing

Regarding all the evidence gathered in this dissertation it seems justified to draw the picture of a comprehensive probabilistic view on conditionals quite optimistically. Probability estimates not only explain the believability people assign to a conditional sentence, they also explain results on drawing inferences from them. Looking at the input information people use to derive any of their judgments, it seems they almost exclusively rely on frequency information, and that is the case for conditionals with explicit frequencies as well as conditionals taken from everyday contexts that convey this information indirectly. If more converging evidence on probabilistic reasoning processes can be accumulated and coherently formulated within the suppositional account of interpreting conditionals, we might see a paradigmatic change in explaining conditionals and their meanings in the not too distant future expressing a quite human perspective on conditionals:

"If there's an exception to the rule, then it is still a rule".

List of references

Adams, E. W. (1981). Truth, proof, and conditionals. *Pacific Philosophical Quarterly, 62*, 323-339.

Anderson, J. R. (1995). *Cognitive psychology and its implications*. New York: W. H. Freeman.

Bennett, J. (2003). *A philosophical guide to conditionals*. Oxford: Oxford University Press.

Betsch, T., Biel, G.-M., Eddelbüttel, C. & Mock, A. (1998). Natural Sampling and base-rate neglect. *European Journal of Social Psychology*, 28, 269-73.

Byrne, R. M. (1989). Suppressing valid inferences with conditionals. *Cognition, 31*, 61-83.

Byrne, R. M., Espino. O. & Santamaria, C. (1999). Counterexamples and the suppression of interferences. *Journal of Memory and Language*, 40, 347-373.

Cummins, D. (1995). Naive theories and causal deduction. *Memory and Cognition*, 23, 646-658.

Cummins, D., Lubart, T., Alksnis, O. & Rist, R. (1991). Conditional reasoning and causation. *Memory and Cognition, 19*, 274-282.

DeNeys, W., Schaeken, W. & D'Ydewalle, G. (2002). Causal conditional reasoning and semantic memory retrieval: A test of the semantic memory framework. *Memory & Cognition*, 30, 908-920.

DeNeys, W., Schaeken, W. & D'Ydewalle, G. (2003a). Causal conditional reasoning and strength of association: The disabling condition case. *European Journal of Cognitive Psychology*, 15(2), 162-167.

DeNeys, W., Schaeken, W. & D'Ydewalle, G. (2003b). Inference suppression and semantic memory retrieval: every counterexample counts. *Memory & Cognition, 31*(4), 581-595.

Dieussaert, K., Schaeken, W. & d'Ydewalle, G. (2002). The creative contribution of content and context factors on the Interpretation of conditionals. *Experimental psychology*, 49(3), 81-195.

Edgington, D. (1991). Do conditionals have truth conditions? In F. Jackson (Ed.), *Conditionals* (pp. 176-201). Oxford: Oxford University Press.

Edgington, D. (1995). On Conditionals. *Mind, 104*, 235-329.

Evans, J. S. B. T. (2003). In two minds: dual process accounts of reasoning. *Trends in Cognitive Sciences, 7*(10), 454-459.

Evans, J. S. B. T. (in press). The heuristic-analytic theory of reasoning. extension and evaluation. *Psychonomic Bulletin & Review*.

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. St. B. T., Newstead, S. E., & Byrne, R. M. J. (1993). *Human Reasoning: The Psychology of Deduction*. Hove & London: Erlbaum.

Evans, J. S. B. T., Over, D. E. & Handley, S. J. (2005). Suppositionals, extensionality, and conditionals: a critique on the mental model theory of Johnson-Laird and Byrne (2002). *Psychological Review,* 112(4), 1040-1052.

Evans, J. S. B. T. & Over, D. E. (2004). *If*: Oxford University Press.

George, C. (1997). Reasoning from uncertain premises. *Thinking & Reasoning,* 3(3), 161-189.

Gigerenzer, G. & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: frequency formats. *Psychological Review*, 102, 684-704.

Gigerenzer, G., Hoffrage, U. & Kleinbölting, H. (1991). Probabilistic mental models: A Brunswikian theory of confidence. *Psychological Review, 98*, 506-528.

Girotto, V. & Johnson-Laird, P. N. (2004). The probability of conditionals. *Psychologia, 47*, 207-225.

Heit, E. (1997). Features of similarity and category-based induction. *Proceedings of the Interdisciplinary Workshop on Categorization and Similarity, University of Edinburgh,*(115-121.).

Johnson Laird, P. N., & Tagart, J. (1969). How implication is understood. *American Journal of Psychology, 2*, 367-373.

Johnson-Laird, P. N. (1983). *Mental Models*. Cambridge: Cambrige Universtiy Press.

Johnson-Laird, P. N. (2001). Mental models and deduction. *Trends in Cognitive Sciences, 5*, 434-442.

Johnson-Laird, P. N. & Byrne, R. M. (1991). *Deduction*. Hove & London: Earlbaum.

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

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.

Juslin, P. Olsson, H. & Olsson, A.-C. (2003). Exemplar effects in categorization and multiple-cue judgment. *Journal of Experimental Psychology: General, 132*, 563-607.

Juslin, P. & Persson, M. (2002). PRObabilities from EXemplars: a "lazy" algorithm for probabilistic inference from generic knowledge. *Cognitive science, 26*, 133-156

Kahneman, D. & Frederick, S. (2002). Representativeness revisited: Attribute substitution in intuitve judgment. In T. Gilovich, D. W. Giffrin & D. Kahneman (Eds.), *Heuristics and Biases: the psychology of intuitive judgment* (pp. 49-81). New York: Cambridge University Press.

Kahneman, D. & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology, 3*, 430-454.

Klaczynski, P. A. (2001). Analytic and heuristic processing influences on adolescent reasoning and decision making. *Child Development*, 72, 844-871.

Klauer, K. C., Musch, J., & Naumer, B. (2000). On belief bias in syllogistic reasoning. *Psychological Review*, 107, 852-884.

Kuhn, T. S. (1962). *The structure of scientific revolutions*: University of Chicago Press.

Lewis, D. (1991). Counterfactual dependence and time's arrow. In F. Jackson (Ed.), *Conditionals* (pp. 46-75). Oxford: Oxford University Press.

Liu, I.-M., Lo, K.-C., & Wu, J.-T. (1996). A probabilistic interpretation of "If-Then". *Quarterly Journal of Experimental Psychology Section A: Human Experimental Psychology, 49A*, 828-844.

Liu, I.-M. (2003). Conditional reasoning and conditionalization. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 29, 694-709.

Lopez, A., Gelman, S. A., Gutheil, G., & Smith, E. E. (1992). The development of category based induction. *Child development, 63*(5), 1070-1090.

Mannhaupt, H.-R. (1994). Produktionsnormen für verbale Kategorien zu 40 geläufigen Kategorien. In M. Hasselhorn (Ed.), *Handbuch deutschsprachiger Normen* (pp. 86-92). Göttingen: Hogrefe.

Markovits, H., & Handley, S. (2005). Is inferential reasoning just probabilistic reasoning in disguise? *Memory & Cognition*, 33,(7) 1315-1323.

Markovits, H., & Barrouillet, P. (2002). The development of conditional reasoning: A mental model account. *Developmental Review*, 22, 5-36.

Markovits, H., & Potvin, F. (2001). Suppression of valid inferences and knowledge structures: the curious effect of producing alternative antecedents on reasoning with causal conditionals. *Memory & Cognition*, 29(5), 736-744.

Markovits, H., & Quinn, S. (2002). Efficiency of retrieval correlates with "logical" reasoning from causal conditional premises. *Memory & Cognition*, 30(5), 696-706.

Nilsson, H., Olsson, H., & Juslin, P. (2005). The cognitive substrate of subjective probability. *Journal of Experimental Psychology: Learning, Memory and Cognition, 31*(4), 600-620.

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. (2001). The probabilistic approach to human reasoning. *Trends in Cognitive Sciences, 5*, 349-357.

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

Oberauer, K. (2006a). Conditionals: their meaning and their use in reasoning. *manuscript submitted for publication*.

Oberauer, K. (2006b). Reasoning with conditionals: A test of formal models of four theories. *Cognitive Psychology*, *53*, 238-283.

Oberauer, K., Geiger, S. M., Fischer, K., & Weidenfeld, A. (in press). Two meanings of "if"? Individual differences in the interpretation of conditionals. *Quarterly Journal of Experimental Psychology*.

Oberauer, K., Weidenfeld, A. & Hörnig, R. (2004). Logical reasoning and probabilities- a comprehensive test of Oaksford and Chater (2001). *Psychonomic Bulletin & Review, 11(3), 521-527.*

Oberauer, K., & Wilhelm, O. (2003). The meaning(s) of conditionals: Conditional probabilities, mental models, and personal utilities. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, 29, 680-693.

Osherson, D., Smith, E. E., Wilkie, O., Lopez, A., & Shafir, E. (1990). Category-based induction. *Psychological Review, 97*(185-200).

Over, D. E. & Evans, J. St. B. T. (2003). The probability of conditionals: The psychological evidence. *Mind & Language*, 18, 340-358.

Over, D. E., Hadjichristidis, C., Evans, J. S. B. T., Handley, S., & Sloman, S. A. (in press). The probability of causal conditionals. *Cognitive Psychology*.

Quinn, S., & Markovits, H. (1998). Conditional reasoning, causality, and the structure of semantic memory: Strength of association as a predictive factor for content effects. *Cognition*, 68, B93-B101.

Ramsey, F. P. (1931). *The foundations of mathematics and other logical essays*. London: Routledge & Kegan Paul.

Rips, L. J. (2001). Two kinds of reasoning. *Psychological Science, 12*, 129-134.

Rips, L. J. (1975). Inductive judgements about natural categories. *Journal of Verbal Learning and verbal Behaviour, 14*, 665-681.

Rottenstreich, Y., & Tversky, A. (1997). Unpacking, repacking, and anchoring: Advances in support theory. *Psychological Review*, 104, 406-415.

Schroyens, W., Schaeken, W., & Handley, S. (2003). In search of counter-examples: Deductive rationality in human reasoning. *The Quaterly Journal of Experimental Psychology*, 56A(7), 1129-1145.

Schroyens, W., Schaeken, W., & d'Ydewalle, G. (2002). The processing of negations in conditional reasoning: A meta analytic case-study in mental model and/or mental logic theory. *Thinking and Reasoning, 7*(7), 121-172.

Sloman, S. A. (1998). Categorical inference is not a tree: The myth of inheritance categories. *Cognitive Psychology, 35*, 1-33.

Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin, 119*, 3-22.

Stalnaker, R. (1970). Probability and Conditionals. *Philosophy of Science* 37: 64-80.

Stanovich, K. E., & West, R. F. (2000). Individual Differences in Reasoning: Implications for the Rationality Debate? *Behavioral and Brain Sciences, 23*, 645-726.

Stevenson, R. J. & Over, D. E. (2001). Reasoning from uncertain premises: Effects of expertise and conversational context. *Thinking and Reasoning*, 7, 367-390.

Thompson, V. (1995). Conditional reasoning: the necessary and sufficient conditions. *Canadian Journal of Experimental Psychology*, 40(1), 1-60.

Thompson, V. (2000). Task specific nature of domain general reasoning. *Cognition*, 76, 209-268.

Tversky, A., & Koehler, D.J. (1994). Support theory: A nonextensional representation of subjective probability. *Psychological Review*, 101, 547-567.

Verschueren, N. (2004). *Contextualising conditional inferences*. Doctoral thesis, Leuven.

Verschueren, N. & Schaeken, N. (2006). Denial inferences: Oaksford, Chater & Larkin (2000) on shaky ground. *Proceedings of the Twenty-Seventh Annual Meeting of the Cognitive Science Society*, Stresa, Italy. Mahwah: Erlbaum Ass.

Verschueren, N., Schaeken, W., & d`Ydewalle, G. (2005). A dual process theory on everyday conditional reasoning. *Thinking and Reasoning, 11*(3), 239-278.

Walsh, N.D. (1997). *Conversations with God. An uncommon dialog: Book 2*. Hampton Roads Publishing Company.

Wason, P. (1966). Reasoning. In B. M. Foss (Ed.), *New Horizons in Psychology I*. Harmandsworth: Penguin.

Wason, P. C., & Johnson-Laird, P. N. (1972). *The Psychology of Deduction: Structure & Content*. Cambridge, MA: Harvard University Press. London: Batsford.

Weidenfeld, A. (2004). *Interpretation of and Reasoning with conditionals - Probabilities, Mental Models, and Causality*. Doctoral thesis. University of Potsdam.

Weidenfeld, A., Oberauer, K. & Hörnig, R. (2005). Causal and noncausal conditionals - an integrated model of interpretation and reasoning. *Quarterly Journal of Experimental Psychology*, 58, pp. 1479-1513(35).

Willemsen, M., & Johnson, E. (2004). MouseLabWebDesigner. *http://www.mouselabweb.org/*, online resource, last access 10.10.2006.

Appendix

Appendix 2.1: Material used in Experiment 2.1

The list shows categories used, overall number of cases, proportion of exception and representatives of the categories used in Experiment 2.1. Ratings show the generation frequency of an exemplar in a category generation task. Numbers give the percentage of subjects who spontaneously generate a given exemplar when presented a given category.

	Category	**N**	**High no. of p¬qs (45%)**	**Low no. of p¬qs (5%)**	**Typical members**	**Rating**	**Atypical members**	**Rating**
1	Profession	1250	562	62	Teacher	48.0	Journalist	1.0
					Doctor	37.0	Taxi driver	1.0
2	Flowers	700	315	35	Rose	86.0	Nettles	1.5
					Carnation	69.0	Heather	1.0
3	Trees	600	270	30	Oak	75.0	Palm	9.0
					Fir	66.5	Juniper	1.0
4	Fish	900	405	45	Shark	58.0	Ray	6.0
					Trout	50.5	Octopus	10.0
5	Fruit	1100	495	55	Apple	86.5	Blue berry	7.0
					Orange	74.0	Pumpkin	3.0
6	Vegetable	250	112	12	Carrot	64.0	Mushroom	1.5
					Beans	43.5	Artichoke	2.5
7	Illness	850	382	42	Flu	44.0	Heartburn	1.0
					Cancer	38.0	Stroke	1.0
8	Furniture	650	292	32	Table	86.0	Hallstand	1.0
					Wardrobe	84.0	Footstool	1.0
9	Musical instrument	350	157	17	Violin	93.5	Triangle	13.5
					Piano	75.0	Jew's Harp	2.5
10	Sport	400	180	20	Swimming	65.5	Yoga	1.0
					Football	60.5	Billiard	1.0
11	Criminal offence	1300	585	65	Murder	79.5	Failure for assistance	1.0
					Theft	74.5		
							Pollution	1.0
12	Birds	750	337	37	Black bird	67.0	Penguin	3.0
					Eagle	43.5	Partridge	2.0

Appendix 2.2: Materials tested for Experiment 2.2

Results of typicality ratings for birds tested for Experiment 2.3. Ratings were given on a scale from 0 ("not similar at all to a typical bird") to 6 ("very similar to a typical bird").

Bird names tested:	**English bird names:**	**Ratings**
Taube	Pigeon	4.91
Rabe	Raven	4.87
Amsel	Black bird	5.30
Fink	Finch	4.96
Meise	Tit	5.15
Adler	Cuckoo	4.47
Schwalbe	Swallow	4.98
Rotkehlchen	Robin	5.21
Wellensittich	Budgerigar	4.49
Elster	Magpie	4.74
Specht	Woodpecker	4.89
Bussard	Buzzard	4.31
Habicht	Hawk	4.04
Falke	Falcon	4.38
Möwe	Seagull	4.79
Papagei	Parrot	3.38
Kakadu	Cockatoo	3.09
Zaunkönig	Wren	4.63
Reiher	Egret	3.43
Pfau	Peacock	2.60
Geier	Vulture	3.23
Kolibri	Humming bird	3.78
Eule	Owl	3.45
Uhu	Eagle owl	3.37
Kuckuck	Cuckoo	4.28
Storch	Stork	3.38
Schwan	Swan	3.30
Ente	Duck	2.96
Albatros	Albatross	3.25
Gans	Goose	2.89
Ibis	Ibis	2.68
Huhn	Chicken	2.77
Wachtel	Quail	3.55
Rebhuhn	Partridge	2.52
Blesshuhn	Coot	2.55
Pelikan	Pelican	2.74
Flamingo	Flamingo	2.34
Pinguin	Penguin	1.53
Emu	Emu	1.44
Strauss	Ostrich	2.09

Appendix 2.3: Materials used in Experiment 2.2

Results of typicality ratings for birds used in Experiment 2.3. Ratings were given on a scale from 0 ("not similar at all to a typical bird") to 6 ("very similar to a typical bird").

Similarity category	**Bird names**	**Ratings**
1	Black bird	5.30
1	Robin	5.21
1	Swallow	4.98
1	Pigeon	4.91
	mean	**5.10**
2	Falcon	4.38
2	Cuckoo	4.28
2	Hawk	4.04
2	Humming bird	3.78
	mean	**4.12**
3	Eagle owl	3.37
3	Vulture	3.23
3	Cockatoo	3.08
3	Duck	2.95
	mean	**3.15**
4	Partridge	2.52
4	Flamingo	2.34
4	Ostrich	2.09
4	Penguin	1.53
	mean	**2.12**

Appendix 2.4: Bird features in Experiment 2.2

List of conditionals stating a specific, fictional feature a bird either had or didn't have in Experiment 2.2.

1. "Wenn es sich um einen Vogel handelt, dann hat er pulnare Arterien."
2. "Wenn es sich um einen Vogel handelt, dann hat das Tier Olinesterase im Blut."
3. Wenn es sich um einen Vogel handelt, dann enthält das Blut des Tieres Kryriozyten."
4. "Wenn es sich um einen Vogel handelt, dann zersetzt das Tier Zellulose aus der Nahrung zu Ziobiose."
5. "Wenn es sich um einen Vogel handelt, dann ist das Tier immun gegen Obrax-Viren."
6. "Wenn es sich um einen Vogel handelt, dann enthält der Speichel des Tieres Loxi-Viren."
7. "Wenn es sich um einen Vogel handelt, dann kann das Tier Cyanwasserstoff verdauen."
8. „Wenn es sich um einen Vogel handelt, dann nimmt das Tier Karbozium mit der Nahrung auf."
9. „Wenn es sich um einen Vogel handelt, dann verläuft eine Ansteckung mit dem Terwa-Virus bei diesem Tier tödlich."
10. "Wenn es sich um einen Vogel handelt, dann ist der Ansteckungsweg für LX3-Viren bei diesem Tier die Nahrungsaufnahme"
11. „Wenn es sich um einen Vogel handelt, dann findet sich Dykomon in der Niere"
12. "Wenn es sich um einen Vogel handelt, dann findet sich Endorpropen in der Speicheldrüse des Tieres"
13. "Wenn es sich um einen Vogel handelt, dann hat das Tier Gotagan im Magen"
14. "Wenn es sich um einen Vogel handelt, dann hat das Tier einen schnellen Puls."
15. "Wenn es sich um einen Vogel handelt, dann ist das Tier Wirt für Eckermilben."
16. "Wenn es sich um einen Vogel handelt, dann kann er Frequenzbereiche von 4H-25kHz hören"

Appendix 2.4: Individual regression weights in Experiment 2.2

Results of typicality ratings for birds used in Experiment 2.3. Ratings were given on a scale from 0 ("not similar at all to a typical bird") to 6 ("very similar to a typical bird").

	B weight Ramsey plain	**p**	**B weight Ramsey Probex**	**p**	**B weight Conj**	**p**
Participant 1	**.59**	**.02**	.40	.14	**.57**	**.03**
Participant 2	-.26	.36	-.18	.51	-.20	.48
Participant 3	-.89	.39	-.13	.69	-.05	.88
Participant 4	-.24	.53	-.25	.52	.19	.62
Participant 5	-.35	.26	-.42	.18	-.26	.42
Participant 6	-.05	.86	-.10	.72	.00	.99
Participant 7	-.84	.77	-.02	.94	.00	.99
Participant 8	-.01	.96	.01	.97	.05	.85
Participant 9	.32	.22	.32	.23	.29	.28
Participant 10	.15	.59	.16	.56	.07	.79
Participant 11	.09	.73	.11	.68	.10	.72
Participant 12	-.32	.23	-.28	.30	-.32	.23
Participant 13	-.29	.27	-.35	.19	-.18	.51
Participant 14	-.96	.17	-.94	.23	-.99	.08
Participant 15	.23	.40	.30	.28	.25	.36
Participant 16	-.21	.44	-.18	.50	-.21	.43
Participant 17	**.89**	**.00**	**.91**	**.00**	**.92**	**.00**
Participant 18	**.86**	**.00**	**.87**	**.00**	**.88**	**.00**
Participant 19	.72	.11	.42	.40	.67	.14
Participant 20	**.99**	**.00**	**.94**	**.00**	**.92**	**.00**
Participant 21	**.91**	**.00**	**.83**	**.00**	**.65**	**.02**
Participant 22	**.91**	**.00**	**.87**	**.00**	**.89**	**.00**
Participant 23	**.87**	**.00**	**.83**	**.00**	**.85**	**.00**
Participant 24	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
Participant 25	**.99**	**.00**	**.94**	**.00**	**.99**	**.00**
Participant 26	**.88**	**.00**	**.83**	**.00**	**.88**	**.00**
Participant 27	**.89**	**.00**	**.91**	**.00**	**.78**	**.00**
Participant 28	**.69**	**.01**	**.56**	**.05**	**.86**	**.00**
Participant 29	**.83**	**.00**	**.82**	**.00**	**.74**	**.00**
Participant 30	**.99**	**.00**	**.95**	**.00**	**.99**	**.00**
Participant 31	**.89**	**.00**	**.87**	**.00**	**.85**	**.00**
Participant 32	**.92**	**.00**	**.88**	**.00**	**.87**	**.00**

Appendix 3.1: Materials used in Experiment 3.1

List of all arbitrary coverstories used in Experiment 3.1.

1. Allergic desease

The research areas of a big medical animal laboratory in Australia concern physiological fundamental relations (e.g. the composition of blood in different species) and allergic diseases. Only recently has the laboratory discovered a new allergic disease in dogs called Midosis. Another department has discovered a hitherto unknown substance in the blood of cats and called it Xathylen. In the last weeks the researches have examined whether Xathylen also exists in dogs. For all dog that were examined it was recorded that:

900 dogs suffered from Midosis and had Xathylen in their blood.
900 dogs suffered from Midosis and didn't have Xathylen in their blood.
100 dogs didn't suffer from Midosis and had Xathylen in their blood.
100 dogs didn't suffer from Midosis and didn't have Xathylen in their blood.

Beate is a veterinarian and claims that: "If a dog suffers from Midosis, then it has Xathylen it its blood."

2. Alarm equipment

Paul works as a security guard for an insurance company. His special attention is on the central strongroom on the ground level. In this room there is a very sensitive alarm system. If somebody enters the vestibule without an appointment it can happen that the floodlight is switched on. Sometimes also the alarm went off. Because of that it was recorded, which security measure went off, every time when somebody entered the vestibule to the strongroom over the last months. The recordings show that in:

900 cases the floodlight went on and the alarm went off.
900 cases the floodlight went on and the alarm did not go off.
100 cases the floodlight did not go on and the alarm went off.
100 cases the floodlight did not go on and the alarm did not go off.

Paul claims that: "If the flood light is on, then the alarm goes off.

3. Computer virus

A couple f days ago, security experts of the agency "fun media" discovered, that a hacker has planted a computer virus in the intranet of the agency. The virus dispreads through the email program. The experts examine all computers. For documentation purposes they secure the IP number of the PC, the system's time and date, the homepage of the internet browser, the email program in use and of course whether the computer was infected with the virus. The examination yields following results.

900 computers had a system date of 24.10.01 and the homepage was www.joke.com
900 computers had a system date of 24.10.01 and the homepage was not www.joke.com
900 computers did not have a system date of 24.10.01 and the homepage was

www.joke.com
900 computers did not have a system date of 24.10.01 and the homepage was not www.joke.com
Steve works for "Fun media". He claims that: "If the computer's system date is 24.10.2001, then the homepage is www.joke.com".

4. DNA Mutation

Western biologists of the chemistry company "science United" investigate native mammals in a remote Chinese province. A sub group of the biologist's team is especially interested in e new DNA mutation in rabbits. Another group of biochemists investigates about a recently discovered substance called Natrolsan that is found in some rodents. For all the rabbits, that were examined, the scientists have recorded the following. In:

900 rabbits Natrolsan could be detected and the DNA mutation was found.
900 rabbits Natrolsan could be detected and the DNA mutation was not found.
100 rabbits Natrolsan could not be detected and the DNA mutation was found.
100 rabbits Natrolsan could not be detected and the DNA mutation was not found.

Angela, who works for the chemistry company claims, that: "If Natrolsan is detected in a rabbit, than the DNA mutation is found."

5. Mechanical art object

Susanne is an artist; she builds unusual sculptural objects on request. In the last weeks, she has created a special sculpture for Daniel, who is an engineer. It is a steel object of the approximate size and shape of a shoe carton. Sometimes a song is played and sometimes a light goes on inside. Of all occasions, where Daniel and his thrilled friends have operated the mechanism of the box, his wife has made the following notes:

900 times the light was on and a song was played.
900 times the light was on and no song was played.
100 times the light was not on and a song was played.
100 times the light was not on and no song was played

Daniel claims that: "If the light is on, then the song is played."

6. Outer space physics

In the year 4000, astrophysicists discovered a new inhabited planet in a foreign galaxy. Scientists are engaged in resolving the biophysical characteristics. In a lot of places the planet's atmosphere contains philoben gas unknown to terrestrial atmosphere. Of 2000 probes of this planet's atmospheric particles it is known that:

900 probes were rich of philoben gas and warmer than 22° centigrade.
900 probes were rich of philoben gas and not warmer than 22° centigrade.
100 probes were not rich of philoben gas and warmer than 22° centigrade.
100 probes were not rich of philoben gas and not warmer than 22° centigrade.

An expert claims that: "If the probe is rich in philoben gas, then it is warmer than 22° centigrade".

7. Tribal behavior

Maria is developmental worker in South America. She is a doctor in the area of the Yamarati-Indians. She worries about the many men of the tribe, who smoke Zenobia Herb that has similar health damaging effects than tobacco. A colleague from Germany who is interested in the frequency of hair loss in different cultures has asked her to keep track of the patients who additionally suffer from hair loss The records about all mean of the tribe reveal that:

900 men smoked Zenobia herbs and suffered from hair loss.
900 men smoked Zenobia herbs and didn't suffer from hair loss.
900 men didn't smoke Zenobia herbs and suffered from hair loss.
900 men didn't smoke Zenobia herbs and didn't suffer from hair loss.

Maria claims that: "If a man smokes Zenobia Herbs, then he suffers from hair loss."

8. Tropical plant

Biologists of an American University have recently discovered a new tropical plant and have called it Pherotelia. It blooms twice a year and grows close to the equator in meagre soil. There is a rare kind of beetle, the blue dot beetle that lives in the area where the Pherotelia grows. The scientist have found out that of all Pherotelia that hey examined:

900 Pherotelia bloomed and had blue dot beetles on their leaves.
900 Pherotelia bloomed and had no blue dot beetles on their leaves.
900 Pherotelia didn't bloom and had blue dot beetles on their leaves.
900 Pherotelia didn't bloom and had no blue dot beetles on their leaves.

Stephanie is a German Biologist. She claims that: "If a Pherotelia blooms, it has blue dot beetles on its leaves."

Appendix 3.2: Conditionals tested for Experiment 3.4

List of all 100 conditionals that were tested for P(q|p) and P(¬p|¬q) for Experiment 3.4. The conditionals have been rated in German and are here displayed in their original wording.

No.	Conditionals:	P(q|p)	P(¬p|¬q)
1.	Wenn man viel Alkohol trinkt, dann wird man betrunken.	90.1	87.6
2.	Wenn man große Zahnschmerzen hat, dann hat man ein Loch im Zahn.	84.4	79.6
3.	Wenn man einen Regenbogen sehen kann, dann hat gleichzeitig die Sonne geschienen und es regnete	91.7	74.0
4.	Wenn ein Film aus den 20er Jahren ist, dann ist es ein Stummfilm	86.3	84.7
5.	Wenn viele Menschen in einer katholischen Kirche schwarz tragen, dann ist es ein Trauergottesdienst.	73.3	66.0
6.	Wenn man reich ist, dann hat man viel Geld	88.0	78.3
7.	Wenn die Wassertemperatur unter 0°C sinkt, dann gefriert es.	90.7	80.4
8.	Wenn in einem Garten Maulwurfshügel zu sehen sind, dann gibt es dort einen Maulwurf	88.9	74.9
9.	Wenn eine Frau im neunten Monat schwanger ist, dann steht sie kurz vor der Geburt ihres Kindes.	91.2	76.7
10.	Wenn jemand stirbt, dann wird er bestattet	90.6	86.0
11.	Wenn eine Gans gut gemästet wird, dann wird sie fett.	87.5	73.8
12.	Wenn Wasser auf 100°C erhitzt wird, dann kocht es	98.5	75.6
13.	Wenn man auf ebener Straße das Bremspedal betätigt, dann verlangsamt das Auto	92.0	82.1
14.	Wenn mein Telefon klingelt, dann ruft mich jemand an.	95.9	92.1
15.	Wenn ein Raubtier hungrig ist, dann geht es auf die Jagd.	97.5	80.6
16.	Wenn man ein Y-Chromosom besitzt, dann ist man ein Mann.	87.4	89.9
17.	Wenn man das Gaspedal betätigt, dann beschleunigt das Auto.	90.2	89.3
18.	Wenn man die Toilettenspülung betätigt, dann wird gespült.	93.0	83.6
19.	Wenn die Zugvögel wegfliegen, dann ist Herbst.	94.4	79.6
20.	Wenn draußen Schnee liegt, dann kann man eine Schneeballschlacht machen.	83.1	78.3
21.	Wenn ein Druckbleistift vorne eine Mine hat, dann schreibt er.	88.2	80.2
22.	Wenn eine Pfütze gefroren ist, dann ist es draußen kälter als 0°C.	93.3	79.0
23.	Wenn die Schranke geschlossen ist, dann kommt ein Zug.	95.9	81.4
24.	Wenn eine Briefmarke gestempelt ist, dann wurde der Brief befördert.	92.6	70.6
25.	Wenn es regnet, dann wird die Straße nass.	97.9	86.9
26.	Wenn ein Schwein weiblich ist, dann hat es einen Ringelschwanz.	90.2	56.5

27.	Wenn man in der ersten Jahreshälfte geboren wurde, dann hat man zwei Augen	96.8	57.7
28.	Wenn ein Haus eine gerade Hausnummer hat, dann hat es ein Dach.	91.4	57.4
29.	Wenn jemand beim Schach mit weiß spielt, dann versucht er, den Gegner matt zu setzen.	89.4	54.0
30.	Wenn ein Fußballteam auswärts spielt, dann pfeift ein Schiedsrichter das Spiel an.	95.3	47.7
31.	Wenn ein Tennisspieler gewinnt, dann duscht er nach dem Spiel.	87.8	30.3
32.	Wenn beim Roulette die Kugel auf rot fällt, dann wurde "Rien ne va plus" gesagt.	81.9	57.4
33.	Wenn man Spiegeleier lieber mag als Omelette, dann brät man sie in einer Pfanne.	88.0	43.8
34.	Wenn man Müsli lieber mag, als Cornflakes, dann isst man es aus einem Schälchen.	82.6	55.2
35.	Wenn ein Schüler in Geographie besser ist als in Geschichte, dann bekommt er zwei Mal im Jahr Zeugnisse.	84.9	52.6
36.	Wenn ein Haus eine ungerade Hausnummer hat, dann hat es weniger als 10 Stockwerke.	71.2	55.6
37.	Wenn jemand häufig in den Winterurlaub fährt, kauft er mehr als einmal monatlich Lebensmittel ein.	84.0	65.0
38.	Wenn man eine Frau ist, dann stirbt man, bevor man 100 Jahre alt ist.	88.0	45.8
39.	Wenn der Vater an einem geraden Tag Geburtstag hat, dann geht das 10jährige Kind zur Schule.	89.5	31.7
40.	Wenn Ostern im März ist, dann kommen im Frühling die Zugvögel wieder.	90.9	54.5
41.	Wenn ein Kind Frühling lieber mag als Herbst, dann spielt es am Wochenende.	90.0	54.2
42.	Wenn ein Paar Kinder hat, dann findet es am Strand Muscheln.	79.3	67.2
43.	Wenn in einem Restaurant mehr als 15 Tische sind, dann bringt der Kellner das Essen.	86.5	44.2
44.	Wenn nachts ein Halbmond am Himmel steht, dann schläft das Neugeborene.	80.4	60.0
45.	Wenn der Bauer in der ersten Jahreshälfte Geburtstag hat, dann wird die Kuh täglich gemolken.	89.2	63.3
46.	Wenn an der Nordsee Ebbe ist, dann schlafen die Einwohner nachts.	82.2	40.8
47.	Wenn der Bäcker Ulmen lieber mag als Erlen, dann steht er früh auf.	70.9	56.4
48.	Wenn jemand Fußball lieber mag als Tennis, dann trägt er beim Sport Schuhe.	84.5	55.6
49.	Wenn ein Pferd ein Rappe ist, dann hat es einen Schweif.	889	72.2
50.	Wenn ein Elefant aus Afrika kommt, dann hat er einen Rüssel.	99.9	51.4
51.	Wenn jemand Bauarbeiter ist, dann ist er in der ersten Hälfte des Jahres geboren.	71.6	57.8
52.	Wenn jemand Vegetarier ist, dann ist er unter 1,75 m groß.	69.6	58.6
53.	Wenn jemand Millionär ist, dann trinkt er lieber Kaffee als Tee.	67.7	71.8
54.	Wenn ein Deutscher Yoga macht, dann kommt er aus Süddeutschland.	68.7	70.1

55.	Wenn eine Frau eine Milcheiweißallergie hat, dann bekommt sie als erstes Kind einen Jungen.	75.1	55.4
56.	Wenn jemand gelb als Lieblingsfarbe hat, dann sieht er sich gerne Fußball an.	60.0	48.6
57.	Wenn jemand Linkshänder ist, dann mag er Nudeln lieber als Kartoffeln.	65.6	34.1
58.	Wenn ein Japaner blaue Augen hat, dann ist er verheiratet.	71.3	70.8
59.	Wenn ein Erwachsener unter 1,50 m ist, dann mag er Actionfilme lieber als Komödien.	65.5	64.7
60.	Wenn jemand eine Schuhgröße von 48 hat, dann mag er Schokoladeneis lieber als Vanilleeis.	65.3	22.2
61.	Wenn jemand rote Schuhe trägt, dann isst er vor sieben zu Abend.	58.3	68.5
62.	Wenn jemand über 2 m ist, dann wohnt er in einem Haus mit gerader Hausnummer.	66.8	72.8
63.	Wenn ein Paar Zwillinge bekommt, dann werden sie in der Nacht geboren.	64.9	71.3
64.	Wenn man ein Auto fährt, was vor 1930 gebaut wurde, dann mag man Gummibärchen lieber als Lakritz.	55.3	57.3
65.	Wenn in Potsdam 35°C sind, dann spielt der Schachspieler mit schwarz.	60.0	28.9
66.	Wenn der Spieler über 25 Jahre alt ist, dann fällt beim Roulette eine gerade Zahl.	65.0	51.4
67.	Wenn jemand ein vierblättriges Kleeblatt findet, dann ist der Tag wärmer als 15°C.	66.8	48.6
68.	Wenn jemand türkise Haare hat, dann mag er Weintrauben lieber als Mandarinen.	70.1	19.9
69.	Wenn ein Straßencafé im Dezember öffnet, dann ist der Besitzer älter als 40 Jahre.	73.3	65.7
70.	Wenn jemand eine Schlange als Haustier hat, dann geht er lieber joggen als Rad fahren.	70.4	64.7
71.	Wenn jemand einen 3 Meter hohen Weihnachtsbaum besitzt, dann hat das Haus eine Gegensprechanlage.	72.7	69.3
72.	Wenn ein Mann älter als 100 Jahre ist, dann wirft er beim Münzwurf Kopf.	65.9	67.8
73.	Wenn in Europa ein Vulkan ausbricht, dann ist es Nacht.	72.8	31.0
74.	Wenn ein Zebra auf der Straße steht, dann ist der Zirkus groß.	77.6	74.3
75.	Wenn ein Haus höher als 20 m ist, dann mag der Besitzer Hunde lieber als Katzen.	74.2	68.3
76.	Wenn in einem Zimmer eine Halogenlampe hängt, dann ist die Decke höher als 3m.	74.1	63.2
77.	Wenn der Knopf eines Gerätes auf an steht, dann ist das Gerät neu.	60.1	59.7
78.	Wenn man Schinken lieber mag als Salami, dann mag man Tannen lieber als Buchen	68.7	59.1
79.	Wenn man Orangen lieber mag als Äpfel, dann mag man Goethe lieber als Schiller.	63.5	54.8
80.	Wenn man Apfelkuchen lieber mag als Käsekuchen, dann mag man Erbsen lieber als Bohnen.	70.3	58.3
81.	Wenn jemand einen Internetanschluss hat, dann mag er Katzen lieber als Hunde.	66.6	50.7
82.	Wenn jemand weiße Rosen lieber mag als gelbe, dann hört er Musik lieber leise als laut.	70.8	60.7
83.	Wenn jemand lieber Bohneneintopf isst als Kartoffelsuppe, dann fährt er lieber nach Rom als nach	65.9	51.3

	Florenz.		
84.	Wenn jemand dunkle Brötchen lieber isst als helle, dann mag er Monet lieber als van Gogh.	71.8	53.8
85.	Wenn jemand Sean Connery lieber mag als Roger Moore, dann mag er lieber Hamster als Meerschweinchen.	64.2	50.5
86.	Wenn jemand Ken Follet lieber liest als Henning Mankell, dann trinkt er Apfelsaft lieber als Orangensaft.	75.8	62.8
87.	Wenn jemand Französisch lieber mag als Spanisch, dann trinkt er lieber Tee als Kaffee.	70.2	57.2
88.	Wenn jemand Science Fiction Filme lieber sieht als Krimis, dann isst er Huhn lieber als Pute.	67.7	59.4
89.	Wenn jemand Bier lieber mag als Wein, dann fährt er lieber Bus als Straßenbahn.	63.4	56.4
90.	Wenn jemand Frühaufsteher ist, dann mag er Schwimmen lieber als Reiten.	62.2	52.7
91.	Wenn sich jemand gerne sonnt, dann mag er Hip Hop lieber als Rap.	63.8	44.6
92.	Wenn jemand in einem Lebensmittelmarkt mit Bioprodukten einkauft, dann bezahlt er mit der Karte.	62.4	56.3
93.	Wenn jemand morgens Brötchen isst, dann trinkt er Kaffee.	78.7	55.5
94.	Wenn jemand Gartenmöbel aus Holz hat, dann hat er Rosen im Garten.	73.5	57.1
95.	Wenn eine Haustür einen Spion hat, dann hat die Tür einen Knauf.	77.5	65.2
96.	Wenn jemand eine Digitalkamera hat, dann hat er ein Handy mit einem Vertrag.	777	63.3
97.	Wenn jemand gerne Müsli isst, dann isst er vor 8 Uhr Frühstück.	72.3	60.9
98.	Wenn jemand ein Instrument spielen kann, dann hat er ein Haustier.	60.3	54.5
99.	Wenn eine Familie mehr als einen Fernseher hat, dann wohnt sie in einem Haus mit gerader Hausnummer.	57.6	47.6
100.	Wenn ein Oberteil langärmlig ist, dann ist es einfarbig	73.6	65.9

Appendix 3.3: Conditionals used Experiment 3.4

List of the 20 conditionals that were used in Experiment 3.4 in their original German wording and the *English translation*.

No.	Conditionals:	P(q|p)	P(¬p|¬q)
	Category 1: high P(q|p), high P(¬p|¬q)		
3.	Wenn man einen Regenbogen sehen kann, dann hat gleichzeitig die Sonne geschienen und es regnete. *If you can see a rainbow then the sun was shining and it was raining at the same time.*	91.7	74.0
8.	Wenn in einem Garten Maulswurfshügel zu sehen sind, dann gibt es dort Maulwürfe. *If you see molehills in a yard then there are moles.*	88.9	74.9
11.	Wenn eine Gans gut gemästet wird, dann wird sie fett. *If a goose is battened well then it grows fat.*	87.5	73.8
12.	Wenn Wasser auf 100 Grad erhitzt wird, dann kocht es. *If you heat water to 100°C then it boils.*	98.5	75.6
24.	Wenn eine Briefmarke gestempelt ist, dann wurde der Brief befördert. *If a stamp is postmarked, then the letter got sent.*	92.6	70.5
	mean:	**91.8**	**73.8**
	Category 2: high P(q|p), low P(¬p|¬q)		
29.	Wenn jemand im Schach weiß spielt, dann versucht er, den Gegner matt zu setzen. *If somebody plays with the white chessmen, he tries to checkmate his opponent.*	89.4	54.0
30.	Wenn ein Fußballteam auswärts spielt, pfeift ein Schiedsrichter das Spiel an. *If a soccer team plays out of town then a referee opens the game.*	95.3	47.7
33.	Wenn man Spiegeleier lieber mag als Omelette, dann brät man sie in der Pfanne. *If you prefer fried egg to scrambled egg then you fry them in a pan.*	88.0	43.8
38.	Wenn man eine Frau ist, dann stirbt man, bevor man 100 Jahre alt ist. *If you are a woman then you die before turning 100.*	88.0	45.8
48.	Wenn jemand Fußball lieber mag als Tennis, dann trägt er beim Sport Schuhe. *If somebody prefers soccer over tennis then you wear shoes when doing sports.*	84.5	55.6
	mean:	**89.0**	**49.4**
	Category 3: low P(q|p), high P(¬p|¬q)		
53.	Wenn jemand Millionär ist, dann trinkt er lieber Kaffe als Tee. *If somebody is a millionaire then he prefers coffee over tea.*	67.6	71.8
54.	Wenn ein Deutscher Yoga macht, dann kommt er aus Süddeutschland. *If a German practices yoga then he is from the south of the country.*	68.7	70.1
62.	Wenn jemand über 2 Meter groß ist, dann wohnt er in einem Haus mit gerader Hausnummer. *If somebody is taller then 2 meters then he is living in a house with an even house number.*	66.8	72.8

63.	Wenn ein Paar Zwillinge bekommt, dann werden sie in der Nacht geboren. *If a couple has twins then they are born at night.*	64.9	71.3
72.	Wenn ein Mann älter als 100 Jahre ist, dann wirft er beim Münzwurf Kopf. *If a men is older then a 100 years then he gets heads when flipping coins.*	65.9	67.8
	mean:	**66.8**	**70.8**
	Category 4: low P(q|p), low P(¬p|¬q)		
66.	Wenn ein Spieler über 25 Jahre alt ist, dann fällt beim Roulette eine gerade Zahl. *If a gambler is older then 25 years then an even number comes up in roulette.*	65.0	51.4
67.	Wenn jemand ein 4 blättriges Kleeblatt findet, dann ist der Tag wärmer als 15 Grad. *If somebody finds a four leaves clover then the day is warmer then 15 degrees.*	66.8	48.6
81.	Wenn jemand einen Internetanschluss hat, dann mag er Katzen lieber als Hunde. *If somebody has internet access then he prefers cats over dogs.*	66.6	50.7
91.	Wenn sich jemand gerne sonnt, dann hört er HipHop lieber als Rap. *If somebody enjoys sun bathing then he rather listens to Hip-Hop than Rap.*	63.8	44.6
98.	Wenn jemand ein Instrument spielen kann, dann hat er ein Haustier. *If somebody knows how to play an instrument then he has a pet.*	60.2	54.5
	mean:	**64.5**	**50.0**

Appendix 4.1: Conditionals and disabler used in Experiment 3.1-3

1. If a flopper has Xathylen in its blood, then it suffers from Midosis.

But,

If a flopper has an additional substance (Xalsosan) in its blood, then it doesn't suffer from Midosis.

If a flopper has genetic mutation, that makes immune, then it doesn't suffer from Midosis.

If a flopper develops antigens, then it suffers from Midosis.

2. If the probe is warmer than 22° Celsius, then it is rich of philoben gas.

But,

If the probe was sealed under enormous pressure, then it is not rich of philoben gas.

If the container has a leak, then the probe is not rich of philoben gas.

If the probe is stored for a long time before it was examined, then it is not rich of philoben gas.

3. If it thardons, then the streets get sticky. (cf. Cummins, 1995)

But,

If the atmosphere holds an additional substance (K-gas), then the streets do not get sticky.

If one of the rare cleaning vehicles is in use, then the streets do not get sticky.

If a powdery substance (Kalgoren) was strewed in advance, then the streets do not get sticky.

4. If the box is sealed, then it glows in the dark.

But,

If the box was forced open, then it does not glow in the dark.

If the box has a permeable spot, then it does not glow in the dark.

If the box is smaller than a match box, then it does not glow in the dark.

5. If an animal belongs to the family of grocks, then it has 6 legs.

But,

If the grocks has a genetic defect, then it does not have 6 legs.

If the grocks has lop-eared parents, then it does not have 6 legs.

If the grocks has run into a pincer trap for rat like animals, then it does not have 6 legs.

6. If the tree like plant has a square trunk, then it has purple leaves.

But,

If the wind is very heavy, then the tree like plant does not have purple leaves.

If the tree like plant is older than 300 yrs, then it does not have purple leaves.

If the treelike plant is located in soil that is rich of Krenalon, then it does not have purple leaves.

7. If the Karun roots have a striped pattern, then they contain valuable nutrients.

But,

If the Karun roots are harvested in midnight light, then they do not contain valuable nutrients.

If the Karun roots are stored together with another vegetable like plant called Ertonnel, then they do not contain valuable nutrients.

If the Karun roots are stored too long, then they do not contain valuable nutrients.

8. If the flight objects have invisible wings, then they have more than two jet propulsions.

But,

If the flight objects have a turbo engine, then they do not have more than two jet propulsions.

If the flight objects were intended for regional transport, then they do not have more than two jet propulsions.

If the flight objects are from the first generation, then they do not have more than two jet propulsions.

Appendix 4.2: Conditionals tested for Experiment 4.4

List of all 86 conditionals that were tested for exceptions and disabler for Experiment 4.4. The conditionals have been rated in German and are here displayed in their original wording. Number of single exceptions were given in relation to 100 p-cases and thus could range from 0 – 100. Number of disabling conditions were an absolute number that could range from 0 – 6.

No.:	**Conditional**	**Exceptions** (0-100)	**Disabler** (0-6)
1.	Wenn jemand Wasser auf 100 Grad erhitzt, dann kocht es.	6.1	0.91
2.	Wenn Sie jemanden anrufen, klingelt dessen Telefon.	13.2	2.60
3.	Wenn ein Raubtier hungrig ist, dann macht es Jagd auf Beute.	7.7	2.15
4.	Wenn man ein Y-Chromosom besitzt, dann ist man ein Mann.	2.0	0.56
5.	Wenn man bremst, dann wird das Auto langsamer.	5.1	2.21
6.	Wenn ein Hund Flöhe hat, dann kratzt er sich.	3.5	0.58
7.	Wenn man viel Salz isst, dann ist man durstig.	18.0	0.86
8.	Wenn man den Gong anschlägt, dann ertönt er.	2.7	1.23
9.	Wenn man sich in den Finger schneidet, dann fängt er an zu bluten.	6.0	1.15
10.	Wenn man den Abzug der Pistole betätigt, dann feuert sie.	31.8	2.18
11.	Wenn der Stecker des Computer gezogen wird, dann fährt er herunter.	37.0	1.10
12.	Wenn Äpfel reif sind, dann fallen sie vom Baum.	20.9	1.06
13.	Wenn eine EC Karte auf einen Magneten gelegt wird, dann ist sie kaputt.	28.5	1.05
14.	Wenn man Milch an der Luft stehen lässt, dann wird sie sauer.	19.6	0.79
15.	Wenn Butter erwärmt wird, dann schmilzt sie.	1.1	0.57
16.	Wenn man im Parkverbot parkt, dann bekommt man einen Strafzettel..	46.4	2.17
17.	Wenn ein Computer ein Virus hat, dann funktioniert er nicht mehr.	43.6	2.05
18.	Wenn ein Flugzeug abstürzt, dann sterben alle Insassen.	28.5	2.15

19.	Wenn ein Teddy sprechen kann, dann funktioniert er mit Batterien.	9.3	0.94
20.	Wenn man einen König auf der Straße trifft, dann ist Karneval.	18.3	1.00
21.	Wenn eine Pflanze ein Kaktus ist, dann hat sie Stacheln.	12.0	1.44
22.	Wenn ein Mann einen Rock trägt, dann ist er Schotte.	35.8	2.74
23.	Wenn ein Vulkan ausbricht, dann tritt Asche aus.	17.7	0.17
24.	Wenn jemand ein Kaninchen aus einem Hut zieht , dann ist er ein Zauberer.	12.9	1.20
25.	Wenn ein Tiger durch einen Feuer-Reifen springt, dann ist er dressiert.	2.4	1.54
26.	Wenn ein Kind an Blutkrebs erkrankt, dann wird es mit Chemotherapie behandelt.	26.8	2.15
27.	Wenn eine Birne so groß ist wie eine Pampelmuse, dann ist sie genmanipuliert.	34.3	1.40
28.	Wenn ein Vogel spricht, dann ist er ein Papagei.	29.3	0.93
29.	Wenn ein Wollstoff bei 60° gewaschen wird, dann verfilzt er.	19.1	0.75
30.	Wenn auf der Autobahn ein schwerer Unfall passiert, kommen Menschen zu Schaden.	16.2	1.61
31.	Wenn jemand Politiker ist, dann hat er einen vollen Terminkalender.	14.2	1.62
32.	Wenn die Kinder in der Schulzeit schulfrei bekommen, dann ist es draußen sehr heiß.	38.8	2.25
33.	Wenn ein Erwachsener weint, dann ist er traurig.	36.8	2.44
34.	Wenn jemand nur eine Zeitung kauft, dann bezahlt er sie bar.	6.8	1.41
35.	Wenn es draußen friert, dann sind die Straßen glatt.	27.8	1.61
36.	Wenn jemand lange in der Kälte ist, dann wird er krank.	51.8	2.95
37.	Wenn ein Mensch keine Sorgen hat, dann ist er glücklich.	35.0	1.84
38.	Wenn ein Ehemann mit einer anderen Frau flirtet, dann wird seine Frau eifersüchtig.	24.0	2.19
39.	Wenn ein Handy ins Wasser fällt, dann ist es kaputt.	20.9	1.95
40.	Wenn man ein lautes Geräusch hört, dann erschrickt man sich.	35.1	2.17
41.	Wenn ein Pferd weiß ist, dann ist es ein Albino.	72.2	1.29

42.	Wenn Valentinstag auf einen Sonntag fällt, dann steigt der Umsatz in den Blumenläden.	25.1	1.19
43.	Wenn ein Weihnachtsbaum vier Meter hoch ist, dann steht er in einer Kirche.	54.9	2.00
44.	Wenn der Anrufbeantworter nicht angeht, dann ist er ausgeschaltet.	31.1	1.91
45.	Wenn ein Baum entwurzelt wird, dann ist draußen starker Sturm.	36.6	1.69
46.	Wenn ein Haustier Männchen macht, dann ist es ein Hund.	33.8	1.19
47.	Wenn ein Pullover aus Kaschmir-Wolle ist, dann muss man ihn zum Waschen in die Reinigung bringen.	44.5	1.22
48.	Wenn ein Haus brennt, dann kommt die Feuerwehr.	13.9	1.83
49.	Wenn es in Europa ein schweres Erdbeben gibt, dann stürzen Häuser ein.	21.7	1.15
50.	Wenn ein Brief ohne Anschrift ankommt, dann hat ihn eine Taube gebracht.	83.4	1.28
51.	Wenn ein Dirigent ein Orchester dirigiert, dann hat er Noten bei sich.	8.8	1.78
52.	Wenn eine Frau Sex hat, dann wird sie schwanger.	80.5	2.95
53.	Wenn ein Produkt beworben wird, dann gehen seine Verkaufszahlen hoch.	30.3	2.29
54.	Wenn man viel Cola trinkt, dann wird man dick.	49.4	2.55
55.	Wenn jemand kurzsichtig ist, dann trägt er eine Brille.	31.3	2.92
56.	Wenn jemand eine Diät macht, dann verliert er an Gewicht.	32.0	2.72
57.	Wenn man ein Streichholz an der Reibefläche entlang zieht, dann brennt es.	20.9	3.65
58.	Wenn der Computer eingesteckt ist, dann kann man damit arbeiten.	39.6	3.04
59.	Wenn man am Abend Kaffee trinkt, dann kann man nicht einschlafen	46.0	2.23
60.	Wenn ein Mädchen hübsch ist, dann verlieben sich alle Jungs in sie.	51.2	2.09
61.	Wenn ein Mensch sportlich ist, dann geht er 3 Mal in der Woche joggen.	74.2	2.14
62.	Wenn ein Mensch im All ist, dann ist er jünger als 50 Jahre.	12.4	1.45
63.	Wenn ein Ehepaar sich streitet, dann lässt es sich	75.9	2.95

	scheiden.		
64.	Wenn ein Geschäft in einem Einkaufszentrum ist, dann findet man dort leicht einen Parkplatz.	37.3	2.71
65.	Wenn ein Arzt den Blutdruck misst, dann prüft er auch die Blutfettwerte.	66.8	1.82
66.	Wenn eine Frau einen Ring trägt, dann ist sie verheiratet.	64.3	2.35
67.	Wenn jemand schläft, dann liegt er im Bett.	19.2	2.94
68.	Wenn jemand den Kühlschrank öffnet, dann leuchtet innen das Licht.	9.3	2.75
69.	Wenn ein Auto einen Motor hat, dann fährt es.	27.1	2.96
70.	Wenn eine Blume blüht, dann duftet sie.	26.2	1.67
71.	Wenn eine Blume eine Rose ist, dann ist sie rot.	58.8	1.75
72.	Wenn Studierende viel lernen, dann bestehen sie die Prüfung.	27.6	3.11
73.	Wenn man die Klimaanlage anmacht, dann ist einem kühl.	47.9	1.88
74.	Wenn Benzin im Tank ist, dann fährt das Auto.	40.2	2.15
75.	Wenn man eine Pflanze gut gießt, dann bleibt sie grün.	30.4	2.32
76.	Wenn man den Lichtschalter betätigt, dann geht das Licht an.	9.4	2.86
77.	Wenn die Straße glatt ist, dann gibt es viele Unfälle.	22.4	2.28
78.	Wenn man reich ist, dann hat man viel Geld.	27.5	1.35
79.	Wenn jemand ein Handy anmacht, dann sieht er etwas auf dem Display.	5.4	2.42
80.	Wenn man im Café ist, dann trinkt man etwas.	15.9	2.89
81.	Wenn ein Wildschwein Junge hat, dann ist es gefährlich.	26.5	1.22
82.	Wenn ein Chirurg einen entzündeten Blinddarm entfernt, dann geht es dem Patienten hinterher besser.	14.8	2.38
83.	Wenn ein Pop-Star ein Konzert gibt, dann gibt er eine Zugabe.	23.9	3.00
84.	Wenn jemand krank wird, dann geht er zum Arzt.	46.5	4.59
85.	Wenn jemand duscht, dann werden seine Haare nass.	39.1	2.13

Appendix 4.3: Conditionals used in Experiment 4.4

List of the 20 conditionals that were used in Experiment 4.4 in their original German wording and the *English translation*. Number of single exceptions were given in relation to 100 p-cases and thus could range from 0 – 100. Number of disabling conditions were an absolute number that could range from 0 – 6.

No.	Conditionals:	Exceptions (0-100)	Disabler (0-6)
	Category 1: many exceptions, many disabler		
16.	Wenn man im Parkverbot parkt, dann bekommt man einen Strafzettel. *If you park your car on an illegal spot then you will get a parking ticket.*	46.4	2.17
36.	Wenn jemand lange in der Kälte ist, dann wird er krank. *If someone stays in the cold for a long time then he will get sick.*	51.8	2.95
52.	Wenn eine Frau Sex hat, dann wird sie schwanger. *If a woman has sexual intercourse then she will get pregnant.*	80.5	2.95
54.	Wenn man viel Cola trinkt, dann wird man dick. *If you drink a lot of coke then you will get thick.*	49.4	2.55
59.	Wenn man am Abend Cola trinkt, dann kann man nicht einschlafen. *If you drink coffee in the evening then you won't be able to fall asleep.*	46.0	2.23
	mean:	**54.8**	**2.57**
	Category 2: many exceptions, few disabler		
11.	Wenn der Stecker des Computers gezogen wird, dann fährt er runter. *If you unplug the computer then it will shut down.*	37.0	1.10
41.	Wenn ein Pferd weiß ist, dann ist es ein Albino. *If a horse is white then it is an albino.*	72.2	1.29
46.	Wenn ein Haustier Männchen macht, dann ist es ein Hund. *If a pet performs stunts then it is a dog.*	33.8	1.19
47.	Wenn ein Pullover aus Kaschmir ist, dann muss man ihn zur Reinigung bringen. *If a pullover is made of cashmere then it has to be brought to a dry cleaner.*	44.5	1.22
50.	Wenn ein Brief ohne Anschrift ankommt, dann hat ihn eine Taube gebracht. *If a letter arrives without an address then a carrier pigeon has brought it.*	83.4	1.28
	mean:	**54.1**	**1.22**
	Category 3: few exceptions, many disabler		
2.	Wenn Sie jemanden anrufen, dann klingelt dessen Telefon. *If you phone someone then his telephone will ring.*	13.2	2.60
5.	Wenn man bremst, dann wird das Auto langsamer. *If you depress the brakes then the car will slow down.*	5.1	2.21
68.	Wenn jemand den Kühlschrank öffnet, dann leuchtet das Licht. *If you open the fridge then the light goes on inside.*	9.3	2.75

76.	Wenn man den Lichtschalter betätigt, dann geht das Licht an. *If the light switch is turned then the light will go on.*	9.4	2.86
79.	Wenn jemand ein Handy anmacht, dann sieht er etwas auf dem Display. *If you switch on your mobile then you see something on the display.*	5.4	2.42
		8.5	**2.57**
Category 4: few disabler, few exceptions			
1.	Wenn jemand Wasser auf 100° erhitzt, dann kocht es. *If water is heated to 100°C then it will boil.*	6.1	0.91
6.	Wenn ein Hund Flöhe hat, dann kratzt er sich. *If a dog has fleas then it will scratch itself.*	3.5	0.58
9.	Wenn man sich in den Finger schneidet, dann fängt er an zu bluten. *If you cut your finger then it will bleed.*	6.0	1.15
31.	Wenn jemand Politiker ist, dann hat er einen vollen Terminkalender. *If somebody is a politician then his appointment calendar is full.*	14.2	1.63
34.	Wenn jemand nur eine Zeitung kauft, dann bezahlte er sich bar. *If somebody only buys a newspaper then he pays cash.*	6.8	1.41
		7.3	**1.13**