The Meaning(s) of "If": Conditional Probabilities and Mental Models

Klaus Oberauer (ko@rz.uni-potsdam.de)
Allgemeine Psychologie I, University of Potsdam
PO box 60 15 53, 14415 Potsdam, Germany

Oliver Wilhelm (oliver.wilhelm@rz.hu-berlin.de)
Department of Psychology, Humboldt-University Berlin
Hausvogteiplatz 5-7, 10117 Berlin, Germany

Abstract

Three experiments with a probabilistic truth-table evaluation task suggest that most people interpret conditionals as asserting a high conditional probability of the consequent, given the antecedent. A minority seems to endorse an interpretation in terms of a single explicit mental model (Johnson-Laird & Byrne, 1991). There was no evidence that a substantive number of people interpret conditionals as material implications. We propose a revision of the theory of mental models that can accommodate both prevalent interpretations as two levels of elaboration of model-based representations.

Introduction

How do people understand statements of the form "if *p* then *q*"? Conditionals seem to have a chameleon-like meaning that varies with content and context (c.f. Johnson-Laird & Byrne, 2002). Still, people can reason systematically from conditional statements even with abstract material presented out of context, as in typical experiments on deduction (for an overview see Evans, 1993). This suggests that there is a psychological core meaning associated with the connective "if ... then".

The experiments reported here put to test two theories of the psychological meaning of conditionals - the theory of mental models (Johnson-Laird & Byrne, 1991) and the hypothesis that conditionals are interpreted as conditional probabilities (Edgington, 1995; Evans, Handley, & Over, 2003; Oaksford & Chater, 2001; Oberauer & Wilhelm, in press). According to the theory of mental models, a conditional of the form "if *p* then *q*" is initially represented as one explicit mental model together with one implicit model (expressed by the three dots):

[p] q
...

The square brackets around *p* signify that there are no other possible cases with *p*, so all the implicit models must be cases of ¬*p*. The initial representation can be "fleshed out" if necessary, yielding three explicit models:

p q
¬p q
¬p ¬q

Thus, the full set of mental models represents those cases from a truth-table that make the conditional statement true. The initial representation corresponds to what has been called a "defective truth table", based on the observation that people often regard the cases with a negation of the antecedent (¬*p*) as irrelevant to the truth of the conditional (Johnson-Laird & Tagart, 1969).

According to the conditional probability view, this is not a defective judgment at all. Edgington (1995), building on earlier work in the philosophy of logic, proposed that the reasonable degree of belief in a conditional "if *p* then *q*" equals the subjective conditional probability $P(q|p)$. This depends on the relative frequency of *pq* cases and *p*¬*q* cases, whereas cases with negated antecedent are irrelevant.

Our experiments used a probabilistic truth-table evaluation task. Participants were given information about the frequencies of the four cases of the truth-table and asked to judge how likely a conditional statement was true (Experiments 1 & 2) or whether it was true or false (Experiment 3). Two factors were varied orthogonally: The relative frequency of *pq*, and the ratio of *pq* to *p*¬*q* cases. Table 1 shows the resulting design.

Table 1: Design of Experiments

Cases | Conditions →	HH	HL	LH	LL
p q	900	900	90	100
p ¬q	100	900	10	100
¬p q	500	100	950	900
¬p ¬q	500	100	950	900

Legend: HH = high frequency of *pq*, high $P(q|p)$, HL = high frequency of *pq*, low $P(q|p)$, LH = low frequency of *pq*, high $P(q|p)$, LL = low frequency of *pq*, low $P(q|p)$.

If people interpret "if *p* then *q*" as asserting that the conditional probability of *q*, given *p*, is high, then their judgment of how likely this statement is true should depend only on the ratio variable. Thus, they should

judge the likelihood of the conditional to be high in conditions HH and LH, and low in the other two conditions. Assuming further that people accept a conditional as "true" if $P(q|p)$ surpasses a threshold close to, but not equal to one, this account would predict that a larger number of people would be willing to accept the conditional statement as true in conditions HH and LH than in the other two.

The theory of mental models predicts that people estimate the probability of a statement by the relative frequency associated with the mental models of the statement, set in relation to the number of all cases (Johnson-Laird, Legrenzi, Girotto, Legrenzi, & Caverni, 1999). Thus, people working with the initial model representation should judge the likelihood of a conditional as a function of the relative frequency of *pq* cases, independent of the ratio factor. People who employ the full set of three explicit models should estimate the probability of a conditional as the sum of the frequencies of the cases *pq*, $\neg pq$, and $\neg p\neg q$, divided by the sample size (i.e., 2000), which equals 1-$P(p\neg q)$. Their judgments should thus depend on the relative frequency of the $p\neg q$ cases only. Within our design, this can be decomposed into a main effect of ratio and a main effect of frequency of *pq*. Note that the predicted effect of the frequency of *pq* goes in the opposite direction of that expected from the initial model representation: Within each category of ratio, high frequency of *pq* goes with high frequency of $p\neg q$, which should yield low estimates of the likelihood of the conditional.

Experiments 1 and 2

Method

Participants. Experiment 1 was a paper-pencil study realizing the design in Table 1 within subjects. Participants were 61 high school students (age range 17-21). Experiment 2 realized the same design between subjects as an internet survey, to which 2255 people contributed data. We accepted only respondents who provided an email address for feedback that was not entered before in order to reduce the likelihood of multiple participations of the same person.

Materials and Procedure. Participants of Experiment 1 received a five-page booklet, with one page for a brief introduction into the task and one for each condition. Each condition introduced an imaginary set of 2000 cards, each card having either an A or a B printed on it in either red or blue. Next the frequency information was given about each combination of letter and color (e.g., "There are 900 cards with a red A"). Participants were then asked four sets of questions for each condition. The first set of four questions asked about the probability that a card selected at random had a particular feature combination (e.g., that it had a blue A). The second pair of questions asked about the probability that a single card, which happens to have an A printed on it, was red, and the probability that it was blue. These two questions targeted directly the conditional probability corresponding to the conditional statement introduced in the third question. As an introduction to the third question, participants were informed that a random set of 10 cards was drawn from the pack. They were then asked to estimate how likely it is that the following statement is true for the 10 cards: "If a card has an A, then it is red". All estimates of probabilities were to be given on a scale from 0 ("absolutely impossible") to 100 ("absolutely certain"). The final question asked participants to imagine a bet on the truth of the conditional. If the conditional turned out to be true for the sample of 10 cards, they would win 100 DM, otherwise they would lose their bet. They were required to indicate the maximum amount they would be willing to bet on the truth of the conditional.

Results

Figure 1 plots participants' estimates of the conditional probability $P(q|p)$ and their estimates of the probability of the conditional, P("if *p* then *q*") in Experiment 1. The third line represents subjective probabilities of the conditional calculated from participants' bets by the formula $P = bet/(bet+100)$, multiplied by 100 for compatibility with the estimates. Figure 2 shows the corresponding data from Experiment 2.

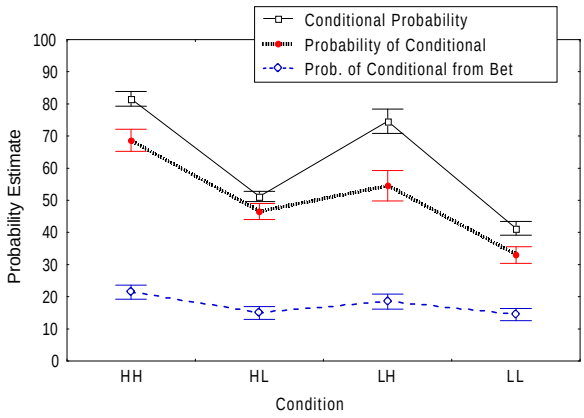

Figure 1: Probability estimates (on a scale from 0 to 100) for the conditional probability $P(q|p)$, the probability of the conditional statement, and the probability of the conditional calculated from participants' bet; Experiment 1. Conditions H = high, L = low, first letter refers to the frequency of *pq*, second to the ratio of *pq* to $p\neg q$. Error bars represent one standard error.

The estimated probabilities of the conditional statements were submitted to an ANOVA with frequency of *pq* (2) and ratio (2) as factors. Both main effects were significant in both experiments; for the effect of ratio, F (1, 60) = 58.67 and F (1, 1998) = 576.4, for Experiment 1 and 2, respectively; for the effect of frequency of *pq*, F (1, 60) = 21.7 and F (1, 1998) = 33.5. The interaction was not significant (F = .02 and 3.7, respectively).

Equivalent ANOVAs on the conditional probabilities of *q*, given *p*, yielded comparable results. There was a main effect of ratio, F (1, 60) = 233.4 and F (1, 1998) = 1199.6, for Experiment 1 and 2, respectively. The main effect of the frequency of *pq* also was significant, F (1, 60) = 8.44 and F (1, 1998) = 104.5. The interaction was not significant.

Finally, the same analysis was conducted with the probabilities calculated from bets, and again there was a main effect of ratio, F (1, 60) = 17.9 and F (1, 1998) = 57.4. The effect of frequency was not significant in Experiment 1 (F = 2.1), but it was in Experiment 2, F (1, 1998) = 4.9.

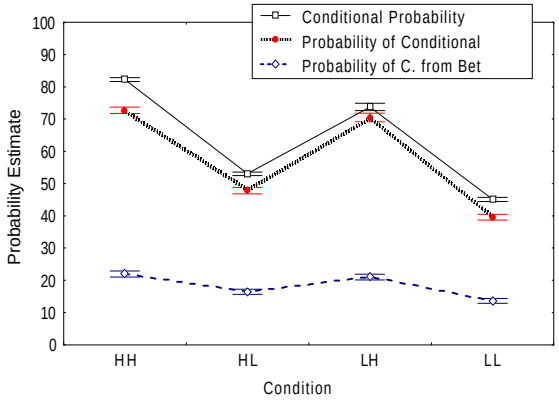

Figure 2: Results from Experiment 2; see Figure 1 for legend.

Figures 1 and 2 show that in all cases the ratio variable was the dominant determinant of people's judgments: High ratios of *pq* to *p*¬*q* were associated with high estimated conditional probabilities, with high estimated probabilities of the conditional statement, and with higher bets on its truth. In addition, higher frequencies of *pq* slightly increased the estimates of the probability of the conditional and the bets on its truth, and, surprisingly, also the estimates of $P(q|p)$.

In Experiment 2, 52% of participants gave an estimate of the probability of the conditional that exactly matched their estimate of $P(q|p)$. Ten percent gave estimates that matched exactly their estimates of the unconditional probability of the *pq* case.

Discussion

The data provide compelling support for the view that a majority of people interpret statements of the form "if *p* then *q*" as expressing a high conditional probability of *q*, given *p*. In addition, the smaller effect of frequency of *pq* suggests that a minority of people judge the probability of a conditional statement on the basis of a mental model representation that contains only the case *pq* as an explicit model. This minority could be responsible for the difference in overall level between the estimated probabilities of the conditional and the $P(q|p)$ estimates: For this group, the probability that the conditional is true would be considerably smaller than $P(q|p)$ – except in condition HL, where the two estimates were in fact very close in both experiments. The low level of probabilities derived from bets are probably due to conservative betting (hardly anyone bet more than 100 DM), which can be explained by loss aversion (Kahneman & Tversky, 1984).

Experiment 3

The results of the foregoing experiments could have been biased toward a probabilistic reading of the conditional statement because the task was embedded in a context of probability estimations. Therefore, Experiment 3 was designed as a further test of the two theories in a context that avoided mention of probabilities altogether.

Method

The third experiment was again a web-based survey, using the same design as Experiment 2. The only difference was that each participant answered a single question: "Someone claims that the following general rule about the playing cards holds: 'If there is an A on a card, then it is red'. Do you think this is true or false?" Responses were obtained from 2198 people.

Results and Discussion

The percentage of participants regarding the conditional statement as true in the four conditions was 21 for HH (i.e., high frequency of *pq*, high ratio), 10 for HL, 20 for LH, and 12 for LL. A log-linear analysis performed on the frequencies of responses in each condition revealed a significant effect of ratio, Chi^2 (2) = 43.0, but no significant effect of frequency of *pq*, Chi^2 (2) = 1.90.

The results are in good agreement with the probabilistic account of conditionals. People can be assumed to accept a conditional statement as true when the conditional probability of the consequent, given the antecedent, surpasses a threshold close to one. The precise location of the threshold can vary between individuals, such that some, but not all of them are willing to accept a conditional when $P(q|p)$ is only .9,

and considerably less, but still a few are apparently willing to accept it even when $P(q|p)$ is only .5.

General Discussion

The experiments summarized here (for a more detailed report see Oberauer & Wilhelm, in press) provide strong support for the contention that people interpret conditional statements as asserting a high conditional probability of the consequent, given the antecedent (Edgington, 1995; Oaksford & Chater, 2001). In addition, they provide evidence for a minority of people who base their judgments on a representation akin to the initial mental model representation (Johnson-Laird & Byrne, 1991). Hardly anyone, it seems, endorsed a representation corresponding to the full set of three mental models making the conditional true. This elaborate set of mental models corresponds to a reading of the conditional as a material implication. Thus, our results also demonstrate that people untrained in formal logic don't interpret conditionals as material implications. In a recent series of experiments similar to ours, Evans et al. (2003) obtained a pattern of results strikingly matching those presented here.

These findings pose a serious problem for a truth-functional account of the psychological meaning of conditionals. The most prominent such account is the theory of mental models (Johnson-Laird & Byrne, 1991). In its present form, the mental model theory of the conditional can account well for the pattern of responses demonstrated by a minority of participants, who based their degree of belief in the conditional on the frequency of the *pq* case. This would be expected if they represent the conditional by a single explicit model of *pq*. The mental model theory cannot account, however, for the observation that most people's degree of belief in "if *p* then *q*" is determined by their subjective conditional probability $P(q|p)$.

In order to accommodate the two most prevalent interpretations observed in our experiments, as well as those from Evans et al. (2003), we propose a modified version of the mental model account. Every statement in ordinary discourse implicitly refers to a domain of discourse, which defines what is relevant for the truth or falsity of the statement. For example, "Peter is in Paris or he is in Moscow" refers to a somewhat extended present on the time dimension and to a particular person named Peter whom both speaker and hearer happen to know. A statement such as "there is a triangle or there is a square" makes sense only when we assume a spatially as well as temporally restricted domain of discourse (e.g., what is drawn on a certain blackboard right now). Mental models for such statements can be interpreted only if the cognitive system frames them into a domain of discourse, which defines the scope of application of the model – otherwise, the cognitive system would have no idea what to do with a model such as:

Conditionals can be interpreted as stating the truth of the consequent in a domain of discourse in which the antecedent is true. The conditional is true to the extent that the consequent is true in the domain of discourse defined by the antecedent. It follows that a reasonable degree of belief in the conditional should depend on the proportion of *pq* cases among all *p* cases, that is, among all cases in the domain of discourse. This is exactly the conditional probability of *q*, given *p*. Cases of $\neg p$ are irrelevant to the truth of the conditional, because they are outside the domain of discourse. This is exactly what most participants express in truth-table evaluation or production tasks (e.g., Johnson-Laird & Tagart, 1969).

In order to represent the meaning of conditionals by mental models, one has to make explicit the domain of discourse specified by the antecedent in the representation. To this end we introduce the concept of a *reference frame* into the ontology of mental models. A reference frame defines explicitly a region in a mental space of possibilities relative to which a mental model should be interpreted. Reference frames are most obviously necessary in mental models of containment relations in space, and the computational implementation of the theory developed by Bara, Bucciarelli, and Lombardo (2001) uses them to represent, for example, statements like "There is no ashtray in Holmes' house" (p. 855).

Bara et al. (2001), however, did not introduce reference frames as a special kind of entity into their "ontology" of mental model representations. We propose that this should be done. A conditional can then be interpreted as an instruction to construct a reference frame defined by the antecedent and to construct a model of *q* (or of the conjunction *pq*) within it. Thus, reference frames take over the function of the square brackets in the original model theory of conditionals. Different from the square brackets, they are not symbolic annotations linked to individual models, but regions in mental space in which models are constructed.

The explicit representation of the domain of discourse by a reference frame can be done more or less completely, and this variation, we suggest, accounts in part for the individual differences in understanding of conditional statements. Different variants of representing "if *p* then *q*" by mental models are sketched in Figure 3.

We believe that the initial mental model constructed by logically untutored adults corresponds to one of the

first two variants (Figure 3a or 3b). With the model in Figure 3a the person endorses a "conjunctive" interpretation of the conditional. The domain of discourse is left implicit, such that the model of *pq* is interpreted relative to an unrestricted space of possibilities. Thus, the degree of belief in the conditional depends on the frequency of *pq* cases relative to all other cases in the population.

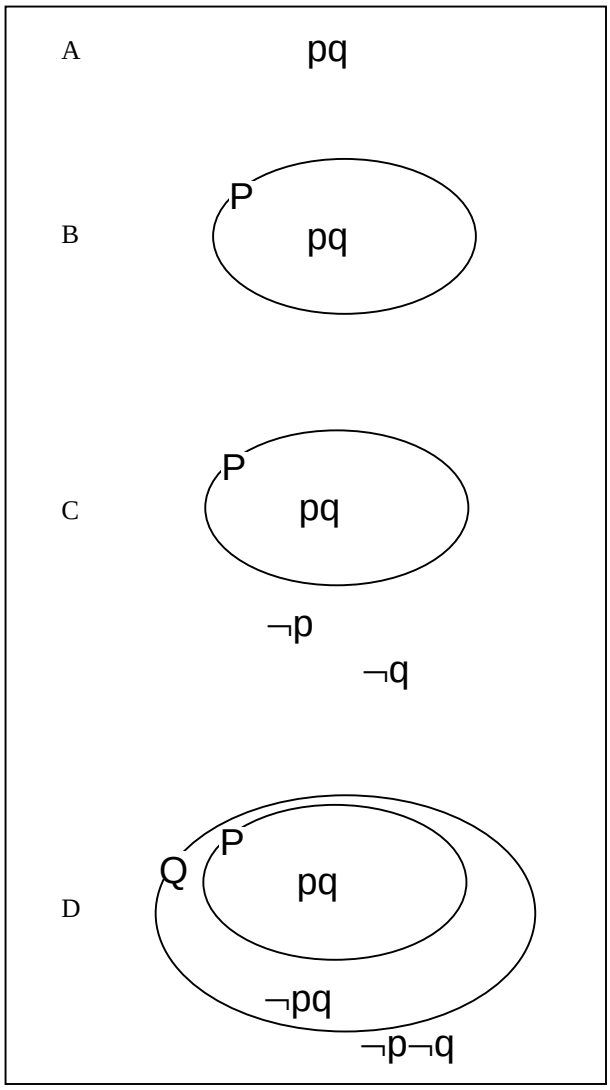

Figure 3: Four levels of elaboration of mental models of a conditional "if p then q". Ellipses designate explicitly represented domains of discourse in the space of possible situations. Capital letters define these domains by marking what is true in situations belonging to them. Small letters refer to mental models of situations.

Figure 3b illustrates a model where the domain of discourse is made explicit as a reference frame. Within this frame, again only *pq* is represented by a model. This follows from the "principle of truth", which is one of the basic assumptions in mental model theory: Models represent only true possibilities, not false ones (Johnson-Laird, 2001; Johnson-Laird, Legrenzi, Girotto, & Legrenzi, 2000). Therefore, $p\neg q$ is not represented. Since the domain of discourse defined by *p* is represented explicitly, it is possible to assign frequencies or probabilities to *p* as well as to *pq* in the way proposed by Johnson-Laird et al. (1999). Thus, with this representation the degree of belief in the conditional should equal the proportion of *pq* cases within the set of *p* cases, and hence the conditional probability $P(q|p)$. The frequency of *pq* relative to the total population (including $\neg p$ cases) should not be regarded as relevant.

Figures 3c and 3d depict more elaborate representations, which reasoners can form to meet special requirements. When the negation of *p* is brought to the person's attention – for example through explicit mention in discourse or by perception – it will be represented explicitly outside the domain of discourse defined by the conditional. This will be the case, for instance, in a *denial of the antecedent* inference task where the minor premise is "*p* is not the case". More generally, this is the kind of representation employed for counterfactual conditionals: The true situation is $\neg p$, and the conditional invites one to consider a counterfactual situation *p*. A negation of *q* will also be represented explicitly outside the reference frame defined by *p*, because the conditional disallows $\neg q$ cases within the reference frame. A natural consequence of this arrangement is to infer $\neg p$ from $\neg q$ (*modus tollens*) and the other way round (*denial of the antecedent*). This pattern of inferences corresponds to a "biconditional" interpretation of conditional statements.

Figure 3d shows a more sophisticated way of elaborating the conditional when presented with negated cases. Here the set of *q* cases is represented as a second reference frame in which the *p* frame is embedded. This representation rests on the flexible engagement of *p* and of *q* as potential domains of discourse and thereby effectively coordinates the conditionals "if *p* then *q*" and "if *q* then either *p* or $\neg p$). This is a representation that supports conditional reasoning in accordance with propositional calculus: acceptance of *modus ponens* and *modus tollens*, and rejection of *denial of the antecedent* and *acceptance of the consequent*.

The four versions of mental models to represent conditionals can be mapped onto the three successive stages in development of mental models proposed by Barrouillet and Lecas (1999, Barrouillet, Grosset, & Lecas, 2000). In the first stage children endorse a conjunctive interpretation of the conditional, based on a single model of *pq*. They accept inferences with positive but not negative minor premises (i.e., *modus ponens* and *acceptance of the consequent*). This corresponds to the representation in Figure 3a. In the

second stage children construct an additional model $\neg p \neg q$. This leads them to interpret conditionals as biconditionals. They accept all inferences involving one of the four possible minor premises (p, $\neg p$, q, $\neg q$) as valid. This corresponds to an initial model like Figure 3b, which is elaborated as in 3c in the face of negative minor premises. The third stage of Barrouillet and Lecas (1999) is reached when people build all three models required to represent the conditional as material implication.

The evidence presented in this article suggests that people hardly ever represent the conditional as material implication. Instead, we believe, people on the second and third stage identified by Barrouillet and Lecas (1999) still represent the core meaning of a conditional as in 3b, but they elaborate it as in 3c or 3d when necessary, that is, when a negation of the antecedent or of the consequent is given. An elaboration into a representation as in 3c supports all four standard inferences, equivalent to a biconditional reading of the conditional premise, whereas the more complex representation depicted in 3d supports only *modus ponens* and *modus tollens*, corresponding to a use of the conditional as material implication. We regard these elaborations not as part of the core meaning of "if", but as context-dependent modulations of it, of which there are probably many more than the two discussed here (c.f. Johnson-Laird & Byrne, 2002).

The modified mental model theory of conditionals specifies four levels of elaboration for a representation of conditionals. Which level is actually used depends on the working memory capacity of the person (higher levels requiring more capacity) and the requirements of the task (assessing the truth or probability of a conditional requires only a representation of level B, whereas reasoning with negated premises requires higher levels). Through the introduction of reference frames that explicitly designate a relevant domain of discourse, the theory can explain why most adults interpret conditionals in terms of conditional probabilities. At the same time, it preserves the explanatory power of the model theory for conditional reasoning, because the model (or models) built within the reference frame (or frames) are the same as in the original model theory.

Acknowledgments

This research was supported by Deutsche Forschungsgemeinschaft (DFG, grant FOR 375 1-1). We thank Karina Schimanke for help with collecting the data, as well as Robin Hörnig and Andrea Weidenfeld for valuable comments on an earlier version of the text.

References

Bara, B. G., Bucciarelli, M., & Lombardo, V. (2001). Model theory of deduction: a unified computational approach. *Cognitive Science, 25*, 839-901.

Barrouillet, P., Grosset, N., & Lecas, J. F. (2000). Conditional reasoning by mental models: chronometric and developmental evidence. *Cognition, 75*, 237-266.

Barrouillet, P., & Lecas, J. F. (1999). Mental models in conditional reasoning and working memory. *Thinking and Reasoning, 5*, 289-302.

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

Evans, J. S. B. T. (1993). The mental model theory of conditional reasoning: critical appraisal and revision. *Cognition, 48*, 1-20.

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

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

Johnson-Laird, P. N., & Byrne, R. M. J. (1991). *Deduction*. Hillsdale: Erlbaum.

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. (2000). Illusions in reasoning about consistency. *Science, 288*, 531-532.

Johnson-Laird, P. N., Legrenzi, P., Girotto, V., Legrenzi, M. S., & Caverni, J. P. (1999). Naive probability: A mental model theory of extensional reasoning. *Psychological Review, 106*, 62-88.

Johnson-Laird, P. N., & Tagart, J. (1969). How implication is understood. *American Journal of Psychology, 82*, 367-373.

Kahneman, D., & Tversky, A. (1984). Choices, values, and frames. *American Psychologist, 39*, 341-350.

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

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