Causal Conditional Reasoning and Conditional Likelihood

Philip M. Fernbach (philip_fernbach@brown.edu)
Adam Darlow (adam_darlow@brown.edu)

Brown University, Department of Cognitive and Linguistic Sciences, Box 1978
Providence, RI 02912 USA

Abstract

We hypothesized that causal conditional reasoning reflects judgment of the conditional likelihood of causes and effects based on a probabilistic causal model of the scenario being judged. Although this proposal has much in common with Cummins' (1995) theory based on the number of disabling conditions and alternative causes, it takes more variables into account and therefore makes some differing predictions. To test this idea we collected judgments of the causal parameters of the conditionals and used them to derive predictions from a model with zero free parameters. We compared these predictions to Cummins' acceptability ratings and to analogous likelihood judgments that we also collected. The hypothesis was borne out for Affirming the Consequent and the analogous diagnostic likelihood judgments, where the model obtained close fits to both data sets. However, we found a surprising dissociation between Modus Ponens and judgments of predictive likelihood leading to a relatively poor fit to the Modus Ponens acceptability ratings. We propose an explanation for this in the discussion.

Key Words: Causal Conditional Reasoning, Conditional Probability, Reaction Time, Probabilistic Model, Modus Ponens, Affirming the Consequent

Causal Conditional Reasoning

When reasoning about deductive arguments people are biased to accept conclusions that are consistent with their beliefs and reject those that are inconsistent, regardless of argument validity (Evans, 2007). In a set of seminal papers, Cummins (1995; Cummins et al., 1991) showed that these belief biases follow systematic principles when people reason about conditional arguments with causal content. People judged the validity of four argument schemata: Modus Ponens (MP), Modus Tollens (MT), Denying the Antecedent (DA) and Affirming the Consequent (AC), though we focus on just MP and AC in this paper.

Despite MP being deductively valid and AC invalid regardless of content, Cummins predicted that for arguments where the antecedent is a cause of the consequent, acceptance rates for MP would be affected by the number of disabling conditions while AC would be affected by the number of alternative causes for the effect.

In the case of MP, thinking of a disabling condition provides a counterexample to the argument and hence may lead people to reject it. An example is given below. Cummins' predicted that (a) would be judged more acceptable than (b) because the conditional in (a) has fewer disablers; reasons why one could put fertilizer on plants and not have them grow quickly are more available than reasons why one could jump into a pool and not get wet.

(a) If Mary jumped into the swimming pool then she got wet.
Mary jumped into the swimming pool.
Therefore she got wet.

(b) If fertilizer was put on the plants then they grew quickly.
Fertilizer was put on the plants.
Therefore they grew quickly.

In the case of AC, alternative causes provide an alternative explanation for the effect and hence make the antecedent seem less necessary. For example Cummins predicted that (c) would be judged more acceptable than (d). It is hard to think of alternative causes for a gun firing besides the trigger being pulled but it is relatively easy to think of causes of wetness besides jumping into a swimming pool.

(c) If the trigger was pulled then the gun fired.
The gun fired.
Therefore the trigger was pulled.

(d) If Mary jumped into the swimming pool then she got wet.
Mary got wet
Therefore she had jumped into the swimming pool.

To test these ideas Cummins' asked one group of participants to spontaneously generate alternative causes and disabling conditions for a host of conditionals and then divided the conditionals into four groups of four conditionals each based on the number of alternatives and disablers (many alternatives, many disablers; many alternatives, few disablers; few alternatives, many disablers; few alternatives, few disablers). A different group was given the arguments based on the 16 conditionals and asked to judge the extent to which the conclusion could be drawn from the premise. Responses were on a 6 point scale from "very sure that the conclusion cannot be drawn" (-3) to "very sure that the conclusion can be drawn" (3). The results provided good support for both predictions.

A Causal Model Theory

Following Oaksford, Chater and Larkin (2000), if the conditional schemata are interpreted in terms of conditional probability, the acceptability of MP maps onto *P(Effect|Cause)* and AC to *P(Cause|Effect)*. Throughout the paper, we refer to *P(Effect|Cause)* as a *predictive* likelihood judgment and to *P(Cause|Effect)* as a *diagnostic* judgment.

By assuming the conditional scenarios approximate a noisy-or common effect model (Cheng, 1997) the expressions in (1) and (2) can be derived for MP and AC respectively (Fernbach & Darlow, 2009; Waldmann et al., 2008). The noisy-or model assumes that there are multiple independent causes for a given effect, each of which may or may not be effective on a given trial.

$$MP \approx P(Effect \mid Cause) = W_c + W_a - W_c W_a \quad (1)$$

$$AC \approx P(Cause \mid Effect) = 1 - (1 - P_c)\frac{W_a}{P_c W_c + W_a - P_c W_c W_a} \quad (2)$$

W_c is the causal power of the cause, the probability that the cause successfully brings about the effect (e.g. the probability that pulling the trigger causes the gun to fire), W_a is the combined strength of all alternative causes, equivalent to the probability of the effect in the absence of the cause (e.g. the probability of the gun firing given the trigger wasn't pulled) and P_c is the prior probability of the cause (e.g. the probability of the trigger being pulled).

According to the full probabilistic model MP increases with both the causal power of the cause and the strength of alternatives (because alternative causes raise the probability of the effect). However, in previous work, we have found that people are not sensitive to the strength of alternative causes when judging predictive likelihood despite its relevance (Fernbach, Darlow & Sloman, 2010). Thus, like Cummins we predicted no effect of W_a and our model for MP is given in (3).

$$MP \approx P(Effect \mid Cause, \sim Alternatives) = W_c \quad (3)$$

AC is a function of all three parameters. It increases with P_c and W_c and decreases with W_a.

Relation Between Cummins' Analysis and Model

According to the causal model the determinants of causal inferences, and hence MP and AC acceptability, are causal power, strength of alternatives and prior probability of the cause. The number of disablers and number of alternatives are factors in the first two parameters, respectively. Causal power is inversely related to the number of disablers. All else being equal, as the number of disablers increases, the probability that the cause fails to bring about the effect increases, corresponding to a decrease in causal power. Thus the model is consistent with the decrease in MP as number of disablers increases, as predicted and found by Cummins. However, not all disablers are equally likely or equally effective in preventing the effect. A single strong disabler could lead to a lower causal power than several weaker disablers, making number of disablers an imperfect predictor of causal power.

Similarly, the number of alternatives is a factor in strength of alternatives. All else being equal, as the number of alternatives increases so does the probability that they will bring about the effect. Therefore, the model predicts that AC will decrease with number of alternatives. As with disablers though, number of alternatives is only a partial predictor of strength of alternatives.

Despite these similarities, the model suggests that Cummins' analysis is incomplete because it only takes a single parameter into account for each judgment. The implication for MP is that its acceptability should increase with the strength of alternative causes but as discussed above we predicted no effect of alternative causes on MP. Our prediction for MP only differs from Cummins in that we expected W_c to provide a better fit than number of disablers.

The model identifies three factors relevant to the acceptability of AC arguments. First, according to the model the prior probability of the cause plays an important role in diagnostic strength. For instance, a cause that is very improbable is unlikely to have occurred relative to other more likely causes and is therefore not as good an explanation for the effect. The second factor is the overall strength of alternatives. This differs from the number of alternatives because not all alternative causes are created equal. In the causal model the strength of alternatives reflects the probability of the effect in the absence of the cause and thus is a joint function of the prior probabilities and causal powers of alternatives. For instance, even a large number of highly improbable or weak alternatives should have less effect on the judgment then a single probable, strong cause. Finally, causal power -- and hence disablers -- should have some influence on AC. All else being equal, if the causal power of the cause is higher, the cause is more likely responsible for the effect. Table 1 summarizes how our predictions differ from Cummins' theory.

Table 1: Best Predictors for MP and AC judgments and Predictive and Diagnostic Likelihood Judgments According to Cummins (1995) and According to our Model

	MP	*AC*
Cummins' Theory	No. of Disablers	No. of Alternatives
Causal Model	Causal Power (W_c)	Full Diagnostic Model
	Predictive Likelihood	*Diagnostic Likelihood*
Cummins' Theory	No Prediction	No Prediction
Causal Model	Causal Power (W_c)	Full Diagnostic Model

Qualitative Support for Causal model

Some trends appear in Cummins' (1995) data that are not predicted by her theory. One is that acceptability ratings of AC for conditionals with many alternative and few disablers were lower than those with many alternatives and many disablers. Both groups had many alternatives and thus should have yielded similar AC judgments according to Cummins. The difference was replicated by De Neys, Schaeken and D'ydewalle (2002) who found lower AC ratings for all few disabler items compared to many disabler items.

De Neys et al. (2002) proposed that when there are many disablers, they interfere with searching memory for alternatives, leading to the observed difference. A perusal of the individual conditionals suggests an alternative explanation based on the causal model. The two groups appear to vary not just in number of disablers but also in some of the factors that the probabilistic analysis says should affect diagnostic judgments. Specifically, the items that obtain low acceptability scores share the property that the cause is weak or improbable relative to the strength of alternatives (see Table 2). For instance, jumping into a swimming pool is improbable relative to other causes of wetness. Likewise, pouring water onto a fire is not the most

common cause of a campfire going out. On the contrary, the high ratings obtain for arguments in which the cause is strong and probable relative to alternatives. There may be many alternatives for a car slowing, but braking is likely the dominant cause. Likewise, studying hard is probably the strongest cause of doing well on a test. Thus, number of alternatives may be equated across groups, but diagnostic strength is not.

Table 2: Mean Acceptability of AC arguments for Two Groups of Conditionals from Cummins' (1995) Exp.1

Conditional	Acceptability (-3 to 3)
Many Alternatives, Many Disablers	
If fertilizer was put on the plants, then they grew quickly	1.00
If the brake was depressed, then the car slowed down	1.00
If John studied hard, then he did well on the test	1.50
If Jenny turned on the air conditioner, then she felt cool	1.08
Many Alternatives, Few Disablers	
If Alvin read without his glasses, then he got a headache	0.75
If Mary jumped into the swimming pool, then she got wet	0.25
If the apples were ripe, then they fell from the tree	1.00
If water was poured on the campfire, then the fire went out	-0.08

Another trend unexplained by her analysis is that few alternative conditionals obtained slightly higher MP judgments than many alternative conditionals despite being equated across number of disablers. Again, the probabilistic analysis suggests why this may be so. Several of the many alternative items have somewhat low causal powers (e.g. 'if the apples were ripe then they fell from the tree') while virtually all of the few alternative items have very high causal powers (e.g. 'if the gong was struck then it sounded.'). Thus, while number of disablers was equated across groups, causal power may have varied leading to differing MP judgments.

Experiment

To test whether the causal model accounts for the causal conditional acceptability ratings we collected judgments of the relevant parameters: the prior probability of the cause (P_c), the causal power of the cause (W_c) and the strength of alternatives (W_a) for Cummins' (1995) conditionals. Using these judgments we derived predictions with zero free parameters to which we compared Cummins' acceptability ratings.

Another implication of our argument is that judgments of the conditional probability of effects and causes should be similar to Cummins' acceptability ratings and should also be accounted for by the causal model. Thus, we collected predictive and diagnostic conditional probability judgments from a second group of participants. We also collected reaction times for these judgments. De Neys et al. (2002) showed that reaction times for causal conditionals basically supported Cummins' analysis. Collecting reaction times with materials phrased in conditional likelihood language allowed us to verify and extend these findings.

Method

Participants 133 Brown University students were approached on campus and participated voluntarily or participated through the psychology research pool in return for class credit.

Design, materials and procedure All experimental conditions used questions based on the 16 conditionals from Cummins' (1995) experiment 1. We therefore adopted Cummins' 2 (number of alternatives; few/many) X 2 (number of disablers; few/many) design with four conditionals in each condition. Judgments were on a 0 ('impossible') to 100 ('definite') scale.

17 Participants provided judgments of the prior probabilities (P_c) and strength of alternatives (W_a) for the 16 conditionals. The questions were split onto two pages with all of the P_c questions on the first page and all of the W_a questions on the second page. The order of questions was randomized on each page. For each question we first stated the conditional and then asked the relevant likelihood question. Examples of P_c and W_a questions are given in (e) and (f) respectively.

(e) If John studied hard then he did well on the test.
How likely is it that John studied hard?

(f) If John studied hard then he did well on the test.
John did not study hard. How likely is it he did well on the test?

A minority of participants interpreted the conditional statement in the P_c questions as indicating that the cause was present and therefore gave ratings of 100 for all of the P_c questions. We removed these responses from the dataset for all subsequent analyses.

An additional 21 participants judged causal power (W_c). Methods were identical except that there was just one page of questions. An example of a W_c question is given in (g).

(g) How likely is it that John studying hard for the test causes him to do well?

95 participants provided predictive and diagnostic likelihood judgments, fully within-participant. Each of these participants therefore answered 32 questions, one predictive and one diagnostic for each conditional. In order to avoid any reaction time differences due to reading time, the wordings of the questions were modified such that each had between 13 and 15 words and between 65 and 75 characters and such that the mean number of words and characters was equated across the four groups of conditionals. Examples of predictive and diagnostic questions are given in (h) and (i):

(h) John studied hard. How likely is it that he did well on the test?

(i) John did well on the test. How likely is it that he studied hard?

This part of the experiment was administered on a computer in the lab. For each question, participants input their answer using the number keys and hit 'return' to move to the next question. Reaction times were measured from the moment the question appeared on the screen to when the participant

hit 'return'. Order of questions was randomly determined for each participant.

Parameter Judgments and Modeling Results

For the following tests we collapsed over conditionals and compared participant means, using Bonferroni correction to control family-wise error rate. As expected, W_a was judged higher for many alternative items compared to few alternative items ($t(16)=13.4$, $p<0.001$) and didn't vary across few and many disablers ($t(16)=1.4$, *ns*).

W_c also varied across the number of alternatives manipulation; W_c was judged higher for few alternative items (M=83.4) compared to many alternative items (M=73.9), $t(20)=4.8$, $p<0.001$). This was not intended by Cummins, but confirmed our intuitions about the unexplained trend in MP; weak alternative items seemed to have lower causal powers despite being equated across number of disablers. Surprisingly, W_c did not vary across the many/few disablers manipulation ($t(20)=1.2$, *ns*) suggesting that number of disablers and causal power were not as closely linked as we expected. The low correlation between number of disablers and W_c ($r=-0.11$, *ns)* also supported this conclusion. P_c did not vary across either manipulation.

Applying the Model Simply computing Equations 2 and 3 using item means would have been inappropriate because the parameter judgments were collected between participants. We therefore used a sampling procedure to generate model predictions. For each conditional we took 10,000 samples each of W_a, P_c and W_c uniformly and randomly from participant responses, and calculated Equations 2 and 3 for each set of samples. We therefore generated 10,000 samples of each probability for each conditional and then took the mean over samples for each conditional as the output of the model. Reruns of the model yielded only negligible differences.

Fits to AC and Diagnostic Judgments Figure 1a depicts Cummins' acceptability ratings for AC on the X-axis plotted against model fits (Equation 2) on the Y-Axis for each of the 16 conditionals, along with the least squares regression line. Figure 1b shows diagnostic judgments plotted against model fits. The model predictions were highly correlated with both Cummins' acceptability ratings (AC) ($r=0.87$, $p<0.001$) and the diagnostic judgments (D) ($r=0.93$, $p<0.001$). To test whether the model is a better predictor of AC and D than the number of alternatives, we performed hierarchical multiple regression analyses of AC and D responses using the model predictions and the number of alternatives as predictors. The model accounted for a significant amount of unique variance beyond what number of alternatives accounted for, both for AC ($F(1,14)=10.7$, $p<0.01$) and for D ($F(1,14)=38.4$, $p<0.001$). Number of alternatives did not account for any unique variance for AC ($F(1,14)=0.24$, *ns*) or for D ($F(1,14)=0.46$, *ns*).

Fits to MP and Predictive Judgments Figure 1c depicts Cummins' acceptability ratings for MP plotted against model fits (equal to W_c according to Equation 3). Figure 1d shows predictive judgments plotted against model fits. Surprisingly, MP ratings and predictive judgments were not highly correlated ($r=0.30$, *ns*), and each was correlated with a different independent variable. MP ratings were significantly correlated with number of disablers ($r=0.53$, $p=0.035$)but not with the model ($r=0.39$, *ns*). Conversely, predictive judgments were highly correlated with the model ($r=0.81$, $p<0.001$) but not with number of disablers ($r=0.04$, *ns*). As predicted, alternative strength did not add any explanatory power; the full model was poorer than W_c at accounting for both MP and predictive judgments.

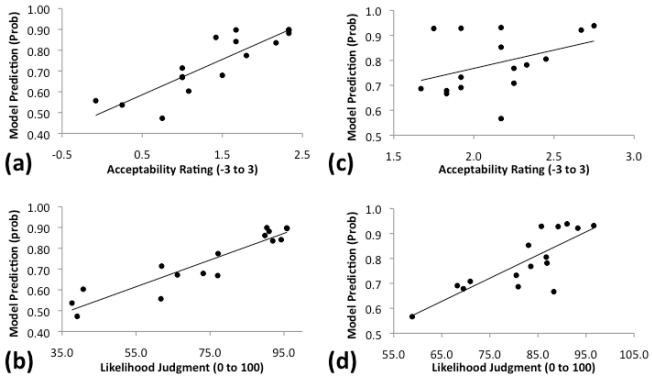

Figure 1: (a) Model fits against Cummins' AC acceptability ratings. (b) Model fits against diagnostic likelihood judgments. (c) Model fits against Cummins' MP acceptability ratings. (d) Model fits against predictive likelihood judgments.

Reaction Time Results

For the sake of concision, the analyses of the predictive and diagnostic judgments are described in the appendix and only the reaction times results are presented in this section. All statistical tests on reaction times used a log transform to normalize the data. Outliers were removed by eliminating all trials that fell more than four standard deviations above or below the participant's mean reaction time. Additionally any trial faster than 1 second was removed.

The reaction time results are depicted in Figure 2. The cleaned data were subjected to a 2 (direction of inference) X 2 (number of alternatives) X 2 (number of disablers) repeated measures ANOVA. There was a main effect of direction of inference; prediction (M=5.88 s) was faster than diagnosis (M=6.21 s), $F(1,95)=25.1$, $p<0.001$. There was also a significant interaction between number of alternatives and direction of inference, $F(1,95)=4.0$, $p<0.05$. No other main effects or interactions were significant.

The interaction between strength of alternatives and direction of inference was driven by diagnostic judgments being faster for items with few alternatives (M = 6.09 s) than for items with many alternatives (M=6.32 s), $t(94)=1.95$, $p=0.05$. Predictive judgments showed no difference in reaction time across the number of alternatives manipulation, $t(94)=0.61$, *ns*.

Number of disablers had no effect on reaction times for predictive judgments (t(94)=1.2, *ns)*. Since W_c accounted for the predictive judgments better than number of alternatives, we suspected it might also yield reaction time differences. To test this we split the conditionals at the median based on W_c and compared reaction times. Confirming the prediction, predictive judgments were faster for items with high W_c (M=5.71 s) than for items with low W_c (M=6.05 s), t(94)=4.19, p<0.0001.

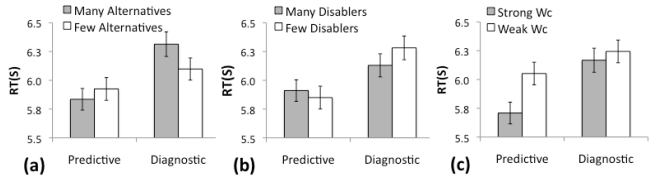

Figure 2: Reactions Times for Predictive and Diagnostic Judgments by (a) number of alternatives, (b) number of disablers and (c) strength of W_c

General Discussion

Summary and Interpretation of Results

Model Fits The diagnostic model achieved very good fits to both Cummins' AC data and our diagnostic likelihood judgments with zero free parameters. It also explained more variance than the single parameter number of alternatives. This confirmed the qualitative analysis indicating that AC judgments were sensitive not just to number of alternatives, but also to the other factors in the causal model in approximately the right way. The model also accounted for the previously unexplained trend in Cummins' AC data for higher AC ratings with more disablers. Altogether, it seems that when judging AC for causal conditionals, people are actually judging the likelihood of the cause (premise) given the effect (conclusion).

The model also matched the predictive judgments closely and differences in W_c explained the previously unexplained trend in Cummins' MP judgments for higher MP judgments with fewer alternatives, a pattern that also showed up in the predictive likelihood judgments (see appendix). But the model didn't match the MP data that well and in fact was slightly worse than the number of disablers at accounting for the variance. Additionally, number of disablers was a remarkably poor predictor of W_c judgments. This was surprising because we expected causal power to vary inversely with number of disablers.

Reaction Times The reaction time data yielded three noteworthy findings: First, predictive judgments were faster than diagnostic ones. This corroborates De Neys et al. (2002) who found that MP was faster than AC and it supports the claim that reasoning from cause to effect is easier in general than reasoning from effect to cause (Tversky & Kahneman, 1982). This difference likely reflects the time it takes to consider alternative causes and prior probability in diagnostic judgment.

Second, diagnostic judgments were faster with few alternatives. This also corroborates De Neys et al. (2002). It implies that searching for alternative causes takes time. It could also reflect the fact that when alternative causes are very weak the judgment is very high and may not require as much thought to calculate. Predictive judgments showed no reaction time differences across number of alternatives. This is more evidence that people don't think of alternatives when making predictions (Fernbach, Darlow & Sloman, 2010).

Finally, we found no reaction time differences for many versus few disablers. This failed to corroborate De Neys et al. (2002) who found that MP was faster for few versus many disablers. We did however find an effect of W_c on reaction times. Prediction was faster for high versus low W_c.

Explaining MP

Both the model fitting and reaction times imply dissociation between how people judged MP and how they judged predictive likelihood. Predictive likelihood judgments and reaction times were explained by differences in W_c but were uncorrelated with number of disablers. Conversely, number of disablers was slightly better at accounting for Cummins' (1995) MP acceptability ratings than W_c and also yielded reaction time differences for MP in De Neys et al.'s (2002) study. This leaves three open questions: First, why is number of disablers such a poor predictor of W_c? Second, why is W_c better at accounting for predictive likelihood judgments and reaction times? Third, why is it worse at accounting for MP?

A speculative answer to the first two questions comes from the possibility that when making predictive likelihood judgments people represent causal systems in terms of their normal, common or prototypical components. If asked to list disablers they may be able to come up with a relatively large number, some of them being very uncommon or atypical. But when asked to judge causal power or make a prediction they think only of the most important disablers. The 'depressed brake' provides a good example. It is not too hard to come up with disablers for why brakes would fail to slow a car (e.g. cut brake lines) but none of them is common. Thus, while number of disablers is relatively high, many of those disablers make a small impact on actual causal power and may have no effect on people's estimates of causal power. On this account, low causal power might still correlate with slower reaction time on the assumption that examples with a greater number of typical or high probability disablers yield lower W_c judgments, lower predictive judgments, and take longer to reason about.

This leaves the question of why W_c fails to account for MP judgments and reaction times, while number of disablers is somewhat better. We don't have a conclusive answer to this question, but we suspect it may be due to people using a mixture of strategies when judging MP. In a deductive context, people reason about MP more naturally than other conditional schemata (Johnson-Laird & Byrne, 2002). This suggests that some participants may be engaging in a different kind of thinking when judging MP in comparison to the other schemata. Perhaps more abstract

thinking leads to rejection of MP based on the ability to think of specific counterexamples without regard to their probability, in which case the number of disablers may be more important than W_c. This is consistent with work by Verschueren, Schaeken and d'Ydewalle (2005) showing two processes in causal conditional reasoning: A relatively quicker intuitive process that arrives at judgments that are highly correlated with conditional probability and a relatively slower, analytic process that correlates with number of alternatives or disablers. Of course, it's important not to jump to firm conclusions on the basis of so few examples (the poor fit to MP was primarily driven by 4 data points). Future work should aim to corroborate the differences in ratings and reaction times for MP versus predictive likelihood with a larger number of well-controlled items.

Conclusions

Our work provides some evidence in favor of the conditional probability approach to conditional reasoning (Oaksford & Chater, 2001, 2003; Over et al., 2007). One caveat to this is that the causal model we propose is incorrect in some important senses. People tend to neglect the strength of alternatives when making predictions, and while aggregate data are fit really well by the diagnostic model, individual data are less consistent. This suggests that people are not actually computing probabilities. It is more natural to think of the model as a computational solution that people only approximate. The literature on probabilistic causal reasoning tends to focus primarily on computational models like this to the detriment of process level implementations. The focus on semantic memory models in the causal conditional reasoning literature is admirable, but the downside of these models is that, as our work shows, people are sophisticated causal reasoners. Simple memory models based on the number of alternatives or disablers won't suffice. A complete model requires mechanisms for judging prior probability, for integrating over the strengths and probabilities of alternative causes, for judging causal power and for combining these various pieces of information in a reasonable way. These processes undoubtedly rely on retrieval from semantic memory – our reaction time data is strong evidence of that – but no current memory model can accommodate the balance of empirical evidence. Exploring how people construct their causal models from remembered alternatives, disablers and other parameters thus offers a promising avenue for future research.

Acknowledgments

This work was supported by a Galner Dissertation Fellowship and an APA Dissertation Research Award to the first author. We thank Steve Sloman, David Over and Dinos Hadjichristidis for helpful discussion and are especially grateful to Denise Cummins for digging up her data from 1995.

References

Cheng, P. W. (1997). From covariation to causation: A causal power theory. *Psychological Review,* 104, 367–405.

Cummins, D. D., Lubart, T., Alksnis, O. and Rist, R. (1991) Conditional reasoning and causation. *Memory and Cognition,* 19 (3), 274-282.

Cummins, D. D (1995) Naïve theories and causal deduction. *Memory and Cognition,* 23 (5), 646-658.

De Neys, W., Schaeken, W. & D'ydewalle, G. (2002). Causal conditional reasoning and semantic memory retrieval: A test of the semantic memory framework. *Memory and Cognition*, 30 (6), 908-920.

Evans, J. ST. B. T. (2007). *Hypothetical thinking: Dual processes in reasoning and judgement*. London: Taylor & Francis.

Fernbach, P. M., Darlow, A. & Sloman, S. A. (2010). Neglect of alternative causes in predictive but not diagnostic reasoning, *Psychological Science*, In Press.

Fernbach, P. M. & Darlow, A. (2009). Causal asymmetry in inductive judgments. *Proceedings of the Thirty-First Annual Conference of the Cognitive Science Society.*

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

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

Oaksford, M., & Chater, N. (2003). Conditional probability and the cognitive science of conditional reasoning. *Mind and Language*, 18 (4), 359 – 379.

Oaksford, M., Chater, N. & Larkin, J. (2000). Probabilities and polarity biases in conditional inference. *Journal of Experimental Psychology: Learning Memory and Cognition,* 26, 883-889.

Over, D., Hadjichristidis, C., Evans, J. St BT. Handley, S. J., & Sloman, S. A. (2007). The probability of causal conditionals. *Cognitive Psychology*, 54, 62-97.

Tversky, A. & Kahneman, D. (1982). Causal schemas in judgements under uncertainty. In D. Kahneman, P. Slovic & A. Tversky (eds.), *Judgement under uncertainty: Heuristics and biases (117-128).* Cambridge: Cambridge University Press.

Verschueren, N., Schaeken, W. & d'Ydewalle, G. (2005). A dual process specification of causal conditional reasoning. *Thinking & Reasoning*, 11 (3), 239-278.

Waldmann, M. R., Cheng, P. W., Hagmayer, Y., & Blaisdell, A. P. (2008). Causal learning in rats and humans: a minimal rational model. In N. Chater, & M. Oaksford (Eds.), *The probabilistic mind. Prospects for Bayesian Cognitive Science* (pp. 453-484). Oxford: University Press.

Appendix

The predictive and diagnostic judgments were subjected to a 2 (direction of inference) X 2 (number of alternatives) X 2 (number of disablers) repeated measures ANOVA. All of the main effects and two-way interactions were significant ($p<0.01$).

Further post hoc tests were performed on predictive and diagnostic judgments separately. Diagnostic judgments were sensitive to number of alternatives with higher judgments for the items with few alternatives ($M=90.7$) than for the items with many alternatives ($M=57.3$), $t(94)=27.9$, $p<0.001$. Diagnostic judgments also varied across number of disablers, with higher judgments for many disablers ($M=78.1$) than few disablers ($M=70.1$), $t(94)=8.9$, $p<0.001$.

As suggested by the differing W_c judgments, predictive judgments also varied across the number of alternatives; Few alternative items ($M=87.8$) yielded higher diagnostic judgments than those with many alternatives ($M=76.3$), $t(94)=6.0$, $p<0.001$. Predictive judgments did not vary with the number of disablers ($t<1$, *ns*). We also tested whether predictive judgments varied with the strength of W_c by dividing the 16 conditionals into two equal groups based on W_c and comparing predictive judgments. As expected, conditionals with high W_c obtained higher predictive judgments ($M=89.1$) than those with low W_c ($M=75.2$), $t(94)=7.0$, $p<0.001$.