Memory & Cognition
2010, 38 (4), 485-492
doi:10.3758/MC.38.4.485

Conditional reasoning, frequency of counterexamples, and the effect of response modality

Henry Markovits, Hugues Lortie Forgues, and Marie-Laurence Brunet
Université du Québec à Montréal, Montréal, Québec, Canada

Geiger and Oberauer (2007) found that when asked to reason with conditionals, people are very sensitive to information about the relative frequency of exceptions to conditional rules and quite insensitive to the relative number of disabling conditions. They asked participants to rate their degree of certainty in a conclusion. In the following studies, we investigated the possibility that this kind of response encourages a more probabilistic mode of processing compared with the usual dichotomous response. In Study 1, participants were given a variant of the problems used by Geiger and Oberauer with either the same *scaled* response format or a dichotomous *categorical* response. The results with the scaled response were identical to those of Geiger and Oberauer. However, the results with the categorical response presented a very different profile. In Study 2, we presented similar problems using only frequency information, followed by a set of abstract conditional reasoning problems. The participants who performed better on the abstract problems showed a significantly different response profile than those who did worse on the abstract problems in the categorical response condition. No such difference was observed in the scaled response condition. These results show that response modality strongly affects the way in which information is processed in otherwise identical inferential problems and they are consistent with the idea that scaled responses promote a probabilistic mode of processing.

Understanding the nature and the processes involved in how people make inferences is a critical question in cognitive psychology. Unfortunately, there is currently no consensus as to what is involved in the inferential process. Complicating the situation is the fact that there is no uniform definition of what an inferential task is; current competing models both suggest and use very different task paradigms. Interpreting the consequent results is thus difficult, since despite the identical labeling of tasks, there is no guarantee that participants necessarily deploy the same processes when important parameters of inferential tasks are varied.

In the present article, we concentrate on conditional reasoning, which involves making inferences with a given major premise of the form "P implies Q" and one of four possible minor premises. *Modus ponens* (MP) is the logical principle that involves reasoning with the premises "P implies Q and P is true" and leads to the logically correct conclusion "Q is true." *Modus tollens* (MT) involves reasoning with the premises "P implies Q and Q is false" and leads to the logically correct conclusion "P is false." These two principles are valid logical forms, since they both lead to a single logically correct conclusion. *Affirmation of the consequent* (AC) involves reasoning with the premises "P implies Q and Q is true." *Denial of the antecedent* (DA) involves reasoning with the premises "P implies Q and P is false." In both cases, the implied conclusions—"P is true" for AC and "Q is false" for DA—are not logically correct. Neither of these forms leads to a single logically correct conclusion, and the correct response would be to deny the implied (biconditional) conclusion in both cases.

Currently, there are two major theories that attempt to explain how people make conditional inferences and what kinds of processes they use to do so. Mental model theory (Johnson-Laird & Byrne, 1991, 2002) supposes that reasoners will generate a representation of the premises using symbolic tokens. Tokens represent classes of possibilities, and a conclusion will be accepted if there is no counterexample available in the representation. This theory is constructed specifically to explain reasoning on standard deductive tasks, which generally require a dichotomous response (i.e., a conclusion must be judged as *certain* or as *uncertain*). Probabilistic theories suppose that a key factor in making a conditional inference is the subjective conditional probability of the conclusion given the premises (Oaksford, Chater, & Larkin, 2000). Importantly, people are assumed to hold variable degrees of belief in conditional statements, which clearly has an impact on the strength of the inferences that they are prepared to make. These theories are constructed specifically to explain reasoning on a probabilistic inference task, in which a natural response is one that is on a scale from *unbelievable* to *believable*.

Of course, the simplest way of reconciling these two theories would be to postulate the existence of two sep-

H. Markovits, henrymarkovits@gmail.com

arate forms of inference, each of which might deploy different cognitive processes. In fact, some recent studies suggest that this may well be the case (Markovits & Handley, 2005; Markovits & Thompson, 2008; Oberauer, 2006; Verschueren, Schaeken, & d'Ydewalle, 2005a). However, both mental model theorists and probability theorists claim that their underlying models provide a universal explanation for all kinds of inferences. Since the natural paradigms for the two kinds of theory differ, the interpretation of the consequent results is made difficult. One of the more potentially pernicious forms of variation concerns the nature of the response required for what is otherwise the same task, and this will be our focus in the present studies.

Most studies of deductive inferences (that have not required an explicitly probabilistic judgment) require a dichotomous response, in which a conclusion is judged to be logically valid or not. However, in some studies, a scaled response has been used that not only was concentrated on the dichotomous judgment of certainty, but that required the reasoners to rate their degree of certainty that a putative conclusion could or could not be drawn from premises (De Neys, Schaeken, & d'Ydewalle, 2003b; Geiger & Oberauer, 2007). Interestingly, the results of these studies have tended to support a general view that is consistent with some form of scaled process, such as that required for probabilistic reasoning. One question that can of course be raised by these studies is whether the form of response might have induced a concomitant change in the reasoners' interpretation of the task. In fact, Geiger and Oberauer explicitly raised this possibility in their discussion, and our specific goal in the present studies was to explicitly examine the influence of the response format on the reasoning process.

We specifically focus on factors that influence acceptance of the MP inference. There is a great deal of evidence that people tend to accept the invited conclusion for the MP (and the MT) inference, which is the logically appropriate response, unless some form of counterexample to the conditional relation is available (Beller & Kuhnmünch, 2007; Byrne, 1989; Cummins, 1995; Cummins, Lubart, Alksnis, & Rist, 1991; De Neys, Schaeken, & d'Ydewalle, 2002; Markovits & Potvin, 2001). Geiger and Oberauer (2007) examined an important question concerning the way in which evidence contradicting the truth of a given major premise is processed and its effect on the MP and MT inferences. Specifically, they looked at the relative importance of two forms of counterexample evidence: the number of disabling conditions and the frequency of exceptions. *Disabling conditions* refers to causes and/or states that could invalidate a given "if P, then Q" relation; that is, could make it possible to have both P true and Q false, which we will refer to as a *p.nq case* (Cummins, 1995; Cummins et al., 1991). For example, consider the following conditional premise: "If a rock is thrown at a window, the window will break." In this case, examples of disabling conditions are "the rock is not thrown hard enough" or "the rock is very small" or "the window is made of reinforced glass," and so on. Previous research has clearly shown a relation between the number of disabling conditions and the tendency to reject the MP and MT inferences (e.g., Cummins, 1995; De Neys, Schaeken, & d'Ydewalle, 2003a). However, another way of organizing counterexample information is suggested by current probabilistic theories. These theories suppose that the key factor in determining the believability of a conditional relation is given by a reasoner's estimate of the relative frequency of confirming to disconfirming cases. One current suggestion for how this is done is that this estimate is made by a procedure based on the Ramsey test, in which reasoners perform a mental simulation in which P is assumed to be true and then estimate the number of times that Q will be true and the number of times that Q will be false (e.g., Evans & Over, 2004). Several studies have indeed shown that an important determinant of people's belief in a conditional statement is the relative frequency of p.nq cases to p.q cases (in which P and Q are both true)—that is, the frequency of exceptions (Evans, Handley, Neilens, & Over, 2007; Evans, Handley, & Over, 2003; Oberauer & Wilhelm, 2003).

These two forms of information are intrinsically correlated, since any evaluation of the frequency of exceptions must, at some level, rely on information about disabling conditions. This makes the interpretation of previous results difficult, since simple measures of the number of disabling conditions fail to account for concomitant effects of the frequency of exceptions and vice versa. However, they are, in principle, dissociable, since it is possible to have disabling conditions that occur with varying levels of frequency, thus producing variable effects on measures of the frequency of exceptions. Geiger and Oberauer (2007) employed an ingenious procedure in which both disabling conditions and frequency information about a given conditional were provided to participants who were then asked to evaluate inferences on the basis of this conditional. The participants were specifically asked to rate their confidence in putative conclusions on a scale (adapted from De Neys et al., 2003b) varying from *certain that I can draw the conclusion* to *certain that I cannot draw that conclusion*. Their results clearly show that when provided with both forms of information, the tendency to reject the MP and MT inferences is strongly related to the frequency of exceptions and is very weakly related to the number of disabling conditions.

STUDY 1

The results of these studies are interpreted as strongly supporting a probabilistic theory of conditionals. However, Geiger and Oberauer (2007) raised the possibility that the response format, which requires a scaled response, could have induced a probabilistic mode of reasoning more than would the dichotomous response used in most deductive tasks. The aim of Study 1 was to examine the effects of response mode using the same basic paradigm. In order to do this, we adapted the same basic procedure that was used by Geiger and Oberauer, but we used two versions of the questions: one that required a scaled response of

confidence about a conclusion and a second version that required a dichotomous decision.

Method

Participants. A total of 138 college-level students (76 female, 62 male; average age = 18 years, 6 months) were randomly assigned to one of the two conditions. All of the participants were French-speaking students from the same college in Montreal, Canada, and were volunteers.

Material. Four versions of a basic paper-and-pencil booklet were constructed. On the first page, the participants were asked for basic demographic information. They were also given the following introductory paragraph (translated from the original French):

> Recently, scientists have discovered a new inhabited planet called Planet Kronus. A team of scientists was then sent to this planet. These scientists discovered some things that do not exist on Planet Earth. In the following pages, you will see a description of their discoveries. Read these carefully, because they give important information about these discoveries. Then, you will be asked to evaluate whether some conclusions that are described can be drawn logically from this information.

Following this paragraph, the participants received a series of five situations. Each situation described a causal conditional relation involving a nonsense term. Directly after the situations, the participants were presented with both disabling conditions and frequency information concerning the relative numbers of p.nq and p.q cases. Following this information, the participants were given two inferences corresponding to an MP inference (P implies Q; P is true. Conclusion: Q is true) and an MT inference (P implies Q; Q is false. Conclusion: P is false). In the categorical condition, the participants were asked whether the conclusion could be logically drawn from the given information, and a dichotomous response (*yes*, *no*) was required. In the scaled condition, the participants were asked to indicate their level of certainty that the conclusion could be logically drawn from the given information. This required using a 10-point scale ranging from −5 (*absolutely certain that the conclusion cannot be drawn*) to +5 (*absolutely certain that the conclusion can be drawn*). Each situation described a different combination of the number of disabling conditions and the frequency of exceptions. These combinations were one disabling, 10% exceptions; three disabling, 10% exceptions; one disabling, 50% exceptions; three disabling, 50% exceptions; and no disabling, 0% exceptions. The order of the MP and MT inferences was varied among the situations.

We describe the first situation as an example of the format used:

> A team of meteorologists observed the climate of Kronus and remarked on an interesting phenomenon. They affirm that on Kronus, if it thardons, the ground becomes soft.
>
> They also know that if it thardons, but if there is any Gas K in the air, the ground does not become soft.
>
> Of the 1,000 last times that it has thardoned, the meteorologists observed that 900 times, it has thardoned, and the ground became soft; 100 times, it has thardoned, and the ground did not become soft.
>
> From this information, Jean reasoned in the following manner: The meteorologists have affirmed that if it thardons, the ground becomes soft.
>
> Observation: It thardons.
>
> Conclusion: The ground becomes soft.

A first booklet was prepared with the five situations in the order described, using a categorical response. A second booklet was identical to the first, except that the disabling conditions and the frequency of exceptions associated with the situations were modified to be, in order, one disabling, 50% exceptions; three disabling, 50% exceptions; one disabling, 10% exceptions; three disabling, 10% exceptions; and no disabling, 0% exceptions. Two further booklets were then prepared that were identical to the first two, except that a scaled response was used.

Procedure. The booklets were distributed to all of the participants, who were told to read the instructions carefully and to take as much time as required to respond.

Results

We calculated the mean number of times that the conclusion was judged to have been drawn logically from the premises in the categorical condition and the mean rating of certainty in the scaled condition. The initial analyses indicated that both the categorical judgments and the scaled ratings were globally lower for the MT than for the MP inferences but that the pattern of variation across the different combinations was the same for the two measures. We then combined the results across the MP and MT inferences (see Table 1).

We first conducted an ANOVA for the scaled condition, with mean ratings as the dependent measure, problem type as a repeated factor, and order as an independent measure. This analysis showed a main effect of problem type [$F(4,62) = 33.17$, $p < .001$] and a significant problem type × order interaction [$F(4,62) = 4.17$, $p < .01$]. Post hoc analyses were performed using the Tukey test, with $p = .05$. An examination of the global effect of problem type showed that mean ratings were significantly higher when there were no disablers or exceptions ($M = 3.22$) relative to those for the two conditions with frequency of exceptions at 10% (combined $M = -0.01$), which were in turn higher than the ratings in the two conditions with frequency of exceptions at 50% (combined $M = -1.51$). No difference was observed as a function of number of disabling conditions, which exactly replicates the results of Geiger and Oberauer (2007).

In order to reexamine the effect of order in a clearer way, we performed an analysis on the four problem types with nonzero exceptions. We conducted an ANOVA with mean ratings as the dependent measure, frequency of exceptions and number of disablers as repeated factors, and order as an independent measure. This showed a main effect of frequency [$F(1,66) = 23.63$, $p < .001$] and a significant order × number of disablers interaction [$F(1,66) = 10.72$, $p < .01$]. Overall, ratings with 50%

Table 1
Percentage of Conclusions Judged to Be Logical in the Categorical Condition and Mean Ratings of Certainty in the Scaled Condition (Combined Over Modus Ponens and Modus Tollens Inferences) As a Function of Number of Disabling Conditions and Frequency of Exceptions

		Condition	
Number of Disablers	Frequency of Exceptions (%)	Categorical ($n = 70$)	Scaled ($n = 68$)
0	0	87.1	3.22
1	10	38.6	0.12
3	10	34.3	−0.14
1	50	32.1	−1.50
3	50	28.6	−1.52

exceptions were significantly lower than ratings with 10% exceptions. Post hoc analyses of the order × number of disablers interaction did not show any significant specific differences. Note that the ratings were somewhat lower with three disablers ($M = -1.30$) than with one disabler ($M = -0.47$) when the items were presented in the first order and that the difference was the opposite when the items were presented in the reverse order ($M = -0.89$ and $M = -0.42$, respectively).

We then conducted an ANOVA for the categorical condition, with mean number of logical judgments as the dependent measure, problem type as a repeated factor, and order as an independent measure. This showed only a main effect of problem type [$F(4,65) = 30.56$, $p < .001$]. Post hoc analyses showed that the conclusion was judged to be logical more often when there were no exceptions ($M = 87.1\%$) than for the other problem types. No significant differences were found among the other problem types.

As we did before, we performed an analysis on the four problem types with nonzero exceptions. We conducted an ANOVA with mean acceptance rate as the dependent measure, frequency of exceptions and number of disablers as repeated factors, and order as an independent measure. No significant difference was observed.

Discussion

The results of this study are clear. When a scaled response modality is used, the results exactly replicate those of Geiger and Oberauer (2007). These results show that there is a strong impact of frequency of exceptions and that there is a significant decrease in ratings between 0% and 10% and between 10% and 50% exceptions, with no effect of number of disablers. These results are also consistent with a probabilistic model.

On the other hand, when a categorical response modality was used, the results were very different. The results of the categorical condition show a large decrease between the 0% condition and all of the other four conditions. Although there was some variation among these conditions (suggestive of a combined effect of number and frequency), this variation is not significant. These results are consistent with a strong form of mental-model theory, which would suggest that any form of counterexample might be sufficient to include a p.nq token in the final representation of the premises and would result in denial of both the MP and the MT inferences.

These results show that changing the response modality changes the way that information is processed in what are otherwise identical inferential problems, which can, in turn, affect the interpretation of the results, despite the fact that the problems used here were identical in all other aspects. In this context, probabilistic theorists would take comfort from the results of the scaled condition, whereas mental-model theorists would take equal comfort from the results of the categorical condition.

Finally, note the similarity of these results to those of Markovits and Handley (2005). They provided participants with relative frequency information concerning potential alternatives to the antecedent and asked them to give either an explicit probabilistic evaluation of the AC inference or a categorical evaluation of this inference. Probabilistic evaluations were linearly related to the frequency of alternative antecedents, whereas categorical evaluations showed a steep decrease between no alternatives and at least one alternative, with little subsequent variation. This is the same pattern of variation found in the present study, and is consistent with the idea that the participants were processing the inferential problems in the scaled condition as if these required a probabilistic response.

STUDY 2

The results of Study 1 show that a simple change in response modality can produce differing interpretations of the same basic problem parameters. They also suggest that this difference might be the result of reasoners' interpreting of problems when a scaled response is used more as a form of probabilistic reasoning than when a categorical scale is used in these same problems. However, this latter conclusion remains speculative. Our aim in Study 2 was to examine this hypothesis more directly. Our basic method involved providing reasoners with an initial set of problems, patterned after those used in Study 1, except that in these problems, only relative probabilities of p.nq cases to p.q cases were presented, at three levels: 0%, approximately 10%, and approximately 50%. These problems were designed to allow the evaluation of the degree of certainty of conclusions in the scaled condition and the number of acceptances in the categorical condition. Following these problems, the reasoners were given a second set of abstract conditional reasoning problems. The abstract problems were designed to provide a measurement of the reasoners' level of abstract reasoning competence. It should be noted that previous studies have shown that presenting abstract reasoning problems before concrete problems has a deleterious effect on the latter (Markovits & Vachon, 1990), whereas presenting concrete problems before abstract problems has no effect on the level of abstract reasoning (Markovits & Lortie-Forgues, in press).

The logic of Study 2 was basically derived from the dual-process formulation of conditional reasoning proposed by Verschueren et al. (2005a, 2005b). This formulation assumes that reasoners have access to a form of probabilistic reasoning that is not resource demanding and to a more resource-demanding form of mental-model-based deductive reasoning. When given a problem for which a probabilistic interpretation is readily available, reasoners strongly tend to use the former strategy. Since the presentation of the problem parameters in the present studies was already designed to suggest a probabilistic format, using a scaled response in addition should very strongly activate a probabilistic strategy. Given the low-cost nature of any such strategy, its use should not vary with reasoners' level of abstract reasoning competence. Thus, we would predict that for the scaled response modality, there should be no

difference in ratings as a function of abstract reasoning performance.

However, when an alternative interpretation is potentially available, at least some reasoners will be able to deploy a mental-model strategy. Specifically, we assumed that using a categorical response should promote increased use of a mental-model-based strategy, at least among some reasoners. This should be more accentuated among reasoners whose level of competence is greater. The clearest a priori difference here concerns the difference between the 0% condition and the 10% condition. When making a primarily probabilistic response, reasoners should rate the 10% condition as relatively similar to the 0% condition. However, competent use of a mental-model-based strategy should result in a uniformly high level of denial of the MP inference at any level of exceptions. Such a strategy should be more often used with reasoners whose abstract reasoning performance is better. Thus, we can specifically predict that the mean number of acceptances in the 10% condition in the categorical condition will be lower for participants whose abstract reasoning performance is better.

We introduced one further manipulation. In half of the scaled responses, the instructions provided by DeNeys et al. (2003b) were used, in which simply an evaluation of a conclusion is asked for, whereas in the other half, more explicitly logical instructions are used. Both responses required indicating the participants' degree of certainty. Since we assume that the key determinate of reasoners' strategies is given by the response modality more than by the description of the task, we predicted no difference between these variables.

Finally, in order to ensure a robust measure of both forms of response, we gave the participants several problems at the 10% and 50% levels. In order to do this, we limited the problems to the MP inference. We also used levels of exceptions that were different but that were close to a given value (e.g., 520 or 500 exceptions). We did this in order to avoid problems associated with simple repetition of the same problem parameters. Since the variation among these examples is less than the precision of the scales used, they were roughly equivalent.

Method

Participants. A total of 172 college-level students (117 female, 54 male; average age = 18 years, 4 months) were randomly assigned to one of the conditions. All of the participants were French-speaking students from the same college in Montreal, Canada, and were volunteers.

Material. Six versions of a basic paper-and-pencil booklet were constructed. The format of these was identical to those used in Study 1, with the following exceptions. Each booklet presented the participants with an initial set of 12 situations presented in the same way as in the Study 1, each of which required evaluation of an MP inference. Each situation used conditional premises with nonsense antecedents and familiar consequents and provided a description of the number of p.nq and p.q cases observed out of a total of 1,000. Of these, 2 were situations with no exceptions; 5 were situations with exceptions close to 10% (100, 95, 90, 85, 80); and 5 were situations with exceptions close to 50% (500, 495, 490, 485, 480). We will refer to these as E0, E10, and E50, respectively. In half of the booklets, the situations were presented with the exceptions in the following order: 0, 100, 500, 0, 90, 495, 80, 490, 95, 480, 85, 485; in the other half, the reverse order was used. An example of one of these situations is

> A team of geologists on Kronus has discovered a new variety of rock, a *trolyte*. After a series of observations, they affirm that on Kronus, if a trolyte becomes wet, it changes color.
>
> Of the 1,000 last times that they observed trolytes, the geologists observed that 920 times trolytes became wet, and they changed color; 80 times trolytes became wet, and they did not change color.
>
> From this information, Jean reasoned in the following manner: The geologists have affirmed that if a trolyte becomes wet, it changes color.
>
> Observation: A trolyte becomes wet.
>
> Conclusion: It will change color.

In half of the booklets, categorical responses that were identical to those used in Study 1 were used, whereas in the other half, scaled responses were used. These differed somewhat from those used in Study 1. In half of the scaled responses, a weak scaled form of the instructions, which was taken directly from DeNeys et al. (2003b), was used. In these instructions the participants were asked to consider the statements of the scientists and the observation and to indicate their evaluation of the conclusion. The instructions were followed by the certainty scale used in Study 1. In the other half, a strong scaled form of instructions was used, in which the participants were asked to indicate the degree of certainty with which they could affirm that the conclusion could be drawn logically from the information given. These instructions were followed by the same certainty scale as that used in the weak form.

Finally, at the end of each of these booklets, the participants received two pages, with a series of abstract conditional reasoning problems on each. On the top of the first page were the following instructions:

> Now, we will ask you to do a special exercise. You must respond to questions about fictitious things on the Planet Kronus, things that do not really exist. Even if these things do not exist, you must consider the statements about them to be true. Then, you must choose the response that follows logically from each of the statements.

The participants were then given a major premise that they were told to suppose was true. This was, "If a person morps, they will become plede."

Following this premise, four questions corresponding to the logical forms MT, AC, DA, and MP were presented . The following format was used:

> A person does not become plede. On can conclude that:
>
> (1) It is certain that this person has morped.
>
> (2) It is certain that this person has not morped.
>
> (3) One cannot be certain that this person has morped.

On the second page was presented the major premise: "If one frifines a bird, it will poite." This premise was followed by questions corresponding to the logical forms (AC, DA, MP, and MT).

Procedure. The booklets were distributed to all of the participants, who were told to read the instructions carefully and to take as much time as required to respond.

Results

We started by grouping together responses to the two E0 problems, the responses to the five E10 problems, and the responses to the five E50 problems. We first examined whether there was any difference in responding between

the two kinds of scaled instructions on the initial problem set. We calculated mean ratings for the three classes of exceptions, transforming them into a score from 0 to 1. The ratings for the weak scaled instructions (E0, $M = 0.91$; E10, $M = 0.67$; E50, $M = 0.31$) were very similar to those for the strong scaled instructions (E0, $M = 0.93$; E10, $M = 0.70$; E50, $M = 0.35$). We then performed an ANOVA with mean ratings for the E0, E10, and E50 problems as the dependent variable, level of exceptions as a repeated measure, and instruction (weak scaled, strong scaled) as an independent variable. This showed no significant effect of instruction [$F(1,83) < 1$] and no significant level of exceptions × instruction interaction [$F(1,82) < 1$]. Given the lack of any difference, we combined the two forms of scaled responses into a single scaled category.

We then calculated the mean number of acceptances of the conclusion in the initial problem set in the categorical condition for the E0 inferences ($M = 0.92$), the E10 inferences ($M = 0.45$), and the E50 inferences ($M = 0.09$). We also calculated the mean transformed ratings on the scaled condition (E0, $M = 0.92$; E10, $M = 0.69$; E50, $M = 0.33$). As can be seen from these means, the levels of response to the E0 inferences were uniformly high in both response modalities, as would be expected. The mean number of acceptances in the categorical condition were lower than mean ratings in the scaled condition for both the E10 and E50 problems, although a direct comparison is not really possible, given the difference in measures.

We then calculated the number of logically correct responses to each of the four inferences with the abstract premises (this gave scores between 0 and 2 for each inference form). We first examined whether receiving the initial problems in the scaled or categorical condition had an effect on the subsequent level of performance on the abstract problems. In order to look at this, we performed an ANOVA on the mean number of correct responses on each of the four inferences, with inference form as a repeated measure and response modality as an independent variable. This showed a significant effect of inference form [$F(3,167) = 30.83$, $p < .001$]. No other effects were significant. The overall mean number of correct responses (out of the eight total inferences) for the participants receiving the scaled problems ($M = 4.40$) was similar to that for those receiving the categorical problems ($M = 4.12$).

We then examined the relationship between performance on the abstract problems and responses to the initial problem set with scaled and categorical responses. In order to do this, we divided the participants into two competence levels by categorizing those whose total number of correct responses on the abstract problems was greater than the median (4) into a high-competence group (49 participants) and all others into a low-competence group (122 participants). We calculated the mean reasoning scores for the E0, E10, and E50 inferences as a function of competence level for the scaled problems and for the categorical problems (see Figure 1). For each of these, we performed an ANOVA with reasoning scores (mean ratings with the scaled problems and mean acceptance rate with the categorical problems) on the E0, E10, and E50 inferences as the dependent variable, level of exceptions as a repeated measure, and competence level as an independent variable.

For the categorical problems, this analysis showed significant effects of level of exceptions [$F(2,81) = 182.15$, $p < .001$] and competence level [$F(1,82) = 5.69$, $p < .02$] and a significant level of exceptions × competence level interaction [$F(2,81) = 6.04$, $p < .01$]. Post hoc analyses of the interaction were performed using t tests with a Bonferroni correction. These showed no difference in E0 performance between the high- and low-competence participants. However, the high-competence participants had significantly lower levels of acceptance of the MP inferences than did the low-competence participants at both the E10 (high competence, $M = .23$; low competence, $M = .54$) and E50 (high competence, $M = .00$; low competence, $M = .12$) levels.

For the scaled problems, there was a main effect of level of exceptions [$F(2,82) = 158.33$, $p < .001$]. There was no effect of competence level on responding with these problems.

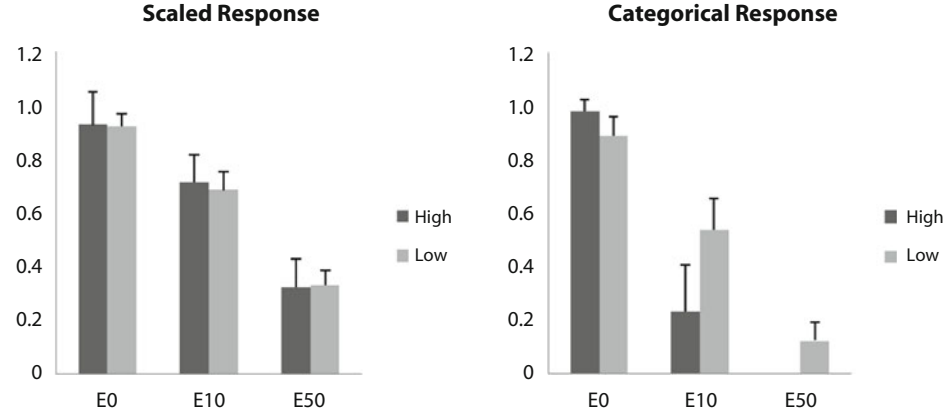

Figure 1. Mean responses to modus ponens inferences in the categorical and scaled conditions as a function of level of abstract reasoning competence (high vs. low). E0, situations with no exceptions; E10, situations with exceptions close to 10%; E50, situations with exceptions close to 50%.

Finally, note that developmental studies indicate that responses to the AC and DA inferences are particularly critical for distinguishing between abstract reasoning levels (Markovits & Vachon, 1990). We used combined AC and DA scores to replicate the above analysis. The results were very similar and will not be further reported.

Discussion

The results of Study 2 are completely consistent with our predictions. When given inferential problems using a scaled response of level of certainty, the participants responded in exactly the same way, irrespective of the exact instructions. More important, the participants who responded more logically to the abstract premises gave the same pattern of responses when using scaled responses as did the participants who responded less well to the abstract problems. In contrast, among the participants who were given problems with a categorical response, those who responded more logically to the abstract premises showed exactly the predicted difference in responses to the E10 problems (i.e., they accepted significantly fewer of these inferences than did the participants who responded less well to the abstract problems). Although the difference is less marked, they also accepted fewer of the E50 inferences.

Finally, although a direct comparison between the categorical and scaled problems is not reliable, it is worthwhile to note that the pattern of responses made by the low-competence reasoners to the categorical problems is quite similar in pattern to responses made on the scaled problems.

GENERAL DISCUSSION

The results of these two studies allow two levels of conclusion. The first, which is a general methodological point, is that the formal characteristics of an inferential problem are simply not sufficient to specify its nature, which is also affected by its surface characteristics. In other words, although all of the problems used in these studies required simple conditional inferences and required some evaluation of the logical status of conclusions, the response modality used clearly had an influence on the way that these problems were interpreted and processed. As both studies show, simply changing the form of response from an evaluation in which the participants were asked to indicate whether a putative conclusion could be logically drawn from provided information to a form in which the participants were asked to rate their level of certainty as to whether this conclusion could be logically drawn changes the pattern of variation in the results. This makes interpreting differences between studies in which inferential performance has been examined difficult if different problem parameters were used. Note also that categorical responses are less sensitive to variation in individual responses than are scaled responses, and the latter might also be less sensitive to group variations. It is possible that the difference in Study 1 might be due to this factor; however, this would not explain the results of Study 2, since in this case, the categorical response scale picked up variability not observed with the scaled responses.

The second conclusion, which is most clearly supported by the results of Study 2, is that the processes preferentially activated by the two response modalities used in these studies are different. Specifically, they support Verschueren et al.'s (2005a) dual-process formulation of conditional reasoning, which postulates a low-cost form of probabilistic evaluation of putative conclusions and a higher cost form of a mental-model-based deductive process. Within this perspective, the results of both studies are consistent with the idea that when given inferential problems using a scaled response of certainty, reasoners will preferentially deploy a probabilistic reasoning strategy, whereas using a categorical response format will activate greater use of the mental-model-based strategy.

These results are also consistent with others that suggest the usefulness of distinguishing between probabilistic and categorical inferential processes (Markovits & Thompson, 2008; Verschueren et al., 2005a). This distinction could also explain the fact that mental-model theory allied with a dual-process version (Verschueren et al., 2005b) was found to be a better predictor of inferential performance with categorical judgments than was a probabilistic theory (Oberauer, 2006), whereas probabilistic theory was a better predictor of inferential performance with scaled judgments (Geiger & Oberauer, 2007).

Finally, although the results of these studies are quite clear, note that they are limited in scope. In future studies, the role of these factors should be examined with more naturalistic materials and with respect to alternative antecedents that are related to performance on the AC and DA inferences.

AUTHOR NOTE

Preparation of the manuscript was supported by a grant from the Natural Sciences and Engineering Research Council of Canada (NSERC) to H.M. and a graduate scholarship from NSERC to H.L.F. Correspondence concerning this article should be addressed to H. Markovits, Department of Psychology, Université du Québec à Montréal, C. P. 8888, Succ Centre-Ville, Montréal, Québec, H3C 3P8 Canada (e-mail: henrymarkovits@gmail.com or markovits.henry@uqam.ca).

REFERENCES

Beller, S., & Kuhnmünch, G. (2007). What causal conditional reasoning tells us about people's understanding of causality. *Thinking & Reasoning*, **13**, 426-460.

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

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

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

De Neys, 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.

De Neys, W., Schaeken, W., & d'Ydewalle, G. (2003a). Causal conditional reasoning and strength of association: The disabling condition case. *European Journal of Cognitive Psychology*, **15**, 161-176.

De Neys, W., Schaeken, W., & d'Ydewalle, G. (2003b). Inference suppression and semantic memory retrieval: Every counterexample counts. *Memory & Cognition*, **31**, 581-595.

Evans, J. St. B. T., Handley, S. J., Neilens, H., & Over, D. [E.]

(2007). Thinking about conditionals: A study of individual differences. *Memory & Cognition*, **35**, 1772-1784.

EVANS, J. ST. B. T., HANDLEY, S. J., & OVER, D. E. (2003). Conditionals and conditional probability. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **29**, 321-335.

EVANS, J. ST. B. T., & OVER, D. E. (2004). *If*. New York: Oxford University Press.

GEIGER, S. M., & OBERAUER, K. (2007). Reasoning with conditionals: Does every counterexample count? It's frequency that counts. *Memory & Cognition*, **35**, 2060-2074.

JOHNSON-LAIRD, P. N., & BYRNE, R. M. J. (1991). *Deduction*. Hove, U.K.: Erlbaum.

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

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

MARKOVITS, H., & LORTIE FORGUES, H. (in press). Conditional reasoning with false premises facilitates the transition between familiar and abstract reasoning. *Child Development*.

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**, 736-744.

MARKOVITS, H., & THOMPSON, V. (2008). Different developmental patterns of simple deductive and probabilistic inferential reasoning. *Memory & Cognition*, **36**, 1066-1078.

MARKOVITS, H., & VACHON, R. (1990). Conditional reasoning, representation and level of abstraction. *Developmental Psychology*, **26**, 942-951.

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

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

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.

VERSCHUEREN, N., SCHAEKEN, W., & D'YDEWALLE, G. (2005a). A dual-process specification of causal conditional reasoning. *Thinking & Reasoning*, **11**, 239-278.

VERSCHUEREN, N., SCHAEKEN, W., & D'YDEWALLE, G. (2005b). Everyday conditional reasoning: A working memory-dependent tradeoff between counterexample and likelihood use. *Memory & Cognition*, **33**, 107-119.

(Manuscript received May 4, 2009;
revision accepted for publication November 1, 2009.)