Memory & Cognition
2003, 31 (4), 581-595

Inference suppression and semantic memory retrieval: Every counterexample counts

WIM DE NEYS, WALTER SCHAEKEN, and GÉRY D'YDEWALLE
University of Leuven, Leuven, Belgium

Reasoning with conditionals involving causal content is known to be affected by retrieval of counterexamples from semantic memory. In this study we examined the characteristics of this search process in everyday conditional reasoning. In Experiment 1 we manipulated the number (zero to four) of explicitly presented counterexamples (alternative causes or disabling conditions) for causal conditionals. In Experiment 2, using a generation pretest, we measured the number of counterexamples participants could retrieve for a set of causal conditionals. One month after the pretest, participants were presented a reasoning task with the same conditionals. The experiments indicated that acceptance of modus ponens linearly decreased with every additionally retrieved disabler, whereas affirmation of the consequent acceptance linearly decreased as a function of the number of retrieved alternatives. Results for denial of the antecedent and modus tollens were less clear. The findings show that the search process does not necessarily stop after retrieval of a single counterexample and that every additional counterexample has an impact on the inference acceptance.

Suppose that you are given the following information: "If the ignition key is turned, then the car starts. The car starts." When you are asked what you should infer from this information, you might conclude that "the ignition key has been turned." However, if you are reminded that the car might be hot-wired or started with a push button, you would likely be less prepared to conclude that the ignition key had been turned. Likewise, when you are told, "If the ignition key is turned, then the car starts. The ignition key is turned." You might conclude that "the car will start." However, if you are told that the car might have a dead battery or be out of fuel, you would be rather reluctant to infer that "the car will start."

Cognitive scientists have conducted a great deal of research to establish how people reason with these "if, then" sentences. Research has typically focused on people's performance on four kinds of conditional arguments: the just-illustrated modus ponens (MP, e.g., "If p then q, p therefore q) and affirmation of the consequent (AC, e.g., "If p then q, q therefore p") inferences, modus tollens (MT, e.g., "If p then q, not q, therefore not p"), and denial of the antecedent (DA, e.g., "If p then q, not p, therefore not q). The first (p) part of the conditional is called the antecedent and the second (q) part is called the consequent.

As the introductory examples make clear, additional knowledge about the conditional relation affects the inferences people are willing to draw. This impact of background knowledge on the reasoning process has long been acknowledged (e.g., Matalon, 1962; Staudenmayer, 1975). In the last few years it has even become one of the main focuses of interest in the conditional reasoning literature. In particular, the role of the availability of alternative causes and disabling conditions has attracted attention.

An alternative cause (alternative) is a condition, aside from the original antecedent, that can bring about the consequent (e.g., hot-wiring the car in the introductory example). A disabling condition (disabler) is a condition that prevents the antecedent from bringing about the consequent (e.g., having a dead battery in the introductory example). Further on, we adopt Byrne's (1989) terminology and refer to alternatives and disablers as counterexamples.

In a pioneering study, Rumain, Connell, and Braine (1983) showed that when a possible alternative was explicitly presented to participants, the AC and DA inferences were less endorsed. Byrne (1989) found a similar effect on MP and MT when a possible disabler was mentioned. These findings have come to be known as the suppression effect.[1]

Further studies established that the suppression effect arises even without explicit presentation of counterexamples (e.g., Cummins, 1995; Cummins, Lubart, Alksnis, & Rist, 1991; Markovits, 1986; Thompson, 1994, 1995). Cummins and colleagues examined the role of counterexample retrieval by looking at the effect of the number of possible alternatives and disablers of a conditional. In a pretest they

This research was supported by grants from the Fund for Scientific Research-Flanders (FWO). Parts of this study were presented at the Twenty-Fourth Annual Conference of the Cognitive Science Society, Washington, DC (2002), and the Phil Johnson-Laird workshop on deductive reasoning, Ghent (2002). We thank Phil Johnson-Laird for valuable discussions. The paper was greatly improved by comments from Evan Heit, Denise Cummins, Henry Markovits, and an anonymous reviewer. Correspondence should be addressed to W. De Neys, Lab Experimental Psychology, University of Leuven, Tiensestraat 102, B-3000 Leuven, Belgium (e-mail: wim.deneys@psy.kuleuven.ac.be).

identified conditionals for which participants generated many or few possible alternatives and disablers. These conditionals were then adopted for a reasoning task with a second group of participants.

Cummins (1995; Cummins et al., 1991) showed that people's acceptance of DA and AC inferences decreased for conditionals with many alternatives. In addition, the number of disablers affected the acceptance of the MP and MT inferences: If there were many conditions that could disable the relation between antecedent and consequent, people tended also to reject these valid inferences. The fact that alternatives and disablers were not explicitly presented indicated that the number of alternatives and disablers people can think of is a crucial factor in conditional reasoning. The findings implied that during a conditional reasoning task, people search their memory for stored counterexamples.

It is widely acknowledged that a theory of conditional reasoning cannot be complete without a full understanding of the counterexample retrieval process (e.g., Johnson-Laird & Byrne, 1994; Thompson, 1994). The vast amount of research in connection with the suppression effect has already resulted in a number of accounts (e.g., Byrne, Espino, & Santamaria, 1999; Oaksford & Chater, 1998; Politzer, in press; Thompson, 2000). These accounts try to explain how the retrieved information affects the reasoning process. However, the crucial question of how the information is retrieved has not yet been dealt with. The characteristics of the search process itself remain largely unknown (Johnson-Laird & Byrne, 1994; Oaksford & Chater, 2001). The present study focuses on this issue.

The recent work of Markovits and collaborators did start paying attention to a characterization of the search mechanism. This mechanism constitutes the core of the general model of conditional reasoning these researchers developed (see Janveau-Brennan & Markovits, 1999; Markovits, 2000; Markovits, Fleury, Quinn, & Venet, 1998; Markovits & Quinn, 2002; Quinn & Markovits, 1998).

The model assumes that as reasoners make conditional inferences, they will automatically access structures with relevant information in semantic memory (Markovits et al., 1998). Such a structure contains semantically or propositionally related elements. In conditional reasoning, the structures would consist of possible alternatives and disablers. According to many influential models of long-term memory (e.g., Anderson, 1983; Gillund & Shiffrin, 1984), the probability of retrieving at least one element from such a semantic memory structure will depend on the number of elements within the structure. Thus, the probability of retrieving at least one element from the structure storing alternatives will be higher for conditionals with many possible alternatives. Likewise, the probability of retrieving a disabler will be higher for conditionals with many possible disablers (Markovits et al., 1998; Markovits & Quinn, 2002; Vadeboncoeur & Markovits, 1999).

Markovits (2000; Markovits et al., 1998) stated that the outcome of the semantic search process will determine the kind of mental models (Johnson-Laird, 1983) a reasoner builds. It is assumed that when reasoners are confronted with a conditional, they will construct an initial internal model of the information the conditional contains. The initial model represents the fact that occurrence of the antecedent is linked with the occurrence of the consequent (e.g., ignition–start, for "If the ignition key is turned, then the car starts"). The initial model can be extended with additional models depending on the outcome of the memory search.

Successful retrieval of an alternative would lead to the construction of an extra model that represents the fact that the consequent can occur without occurrence of the antecedent (e.g., not ignition–start). With this model the AC and DA inferences will be suppressed (Markovits, 2000; Quinn & Markovits, 1998). Retrieval of a disabler would result in the construction of an additional model that makes it clear that it is possible that occurrence of the antecedent is not associated with the occurrence of the consequent (e.g., ignition–not start). This model no longer supports the MP and MT inferences (Markovits, 2000; Vadeboncoeur & Markovits, 1999). It is important to note that these models either do or do not license an inference (e.g., Johnson-Laird & Byrne, 1991). There are no intermediate or graded states of inference acceptance. Whenever a reasoner constructs the additional counterexample models, the inferences are completely rejected.

In Markovits's specification of the memory search process, the number of stored counterexamples is important because it determines the probability that at least one can be retrieved. This specification does not address the impact of additional counterexample retrieval. Indeed, in its present formulation the impact of counterexample retrieval on the inference acceptance is an all-or-nothing phenomenon. Retrieval of a counterexample results in additional model construction leading to the rejection of the otherwise accepted inferences. When there is no counterexample retrieved, the inferences would be accepted. Since an inference is already completely rejected when a single counterexample is retrieved, retrieving extra counterexamples can have no additional impact on the inference acceptance. Consequently, the search process is assumed to stop after the successful retrieval of a single counterexample.

The present study focused on an alternative specification of the semantic search process during conditional reasoning. We tested the assumption that the search process does not terminate after the retrieval of a single counterexample and that every retrieved counterexample has an additional impact on the reasoning process. Here, the number of stored counterexamples is important because it determines the number of counterexamples that can be retrieved, and this number determines the degree to which inferences will be accepted.

The alternative specification gains some credence from related studies. In the field of "uncertain" or probabilistic reasoning, the work of Liu, Lo, and Wu (1996) is especially relevant. Participants received three different conditionals that had previously been rated in terms of having

an antecedent with high (e.g., "If John lives in Canada, then he lives in the northern hemisphere"), medium (e.g., "If Mary moves, then she adds some furniture"), or low (e.g., "If Stan wears glasses, then he is intelligent") sufficiency. Liu et al. observed that MP and MT acceptance gradually decreased with decreasing sufficiency. With this realistic thematic material, conditionals with lower sufficiency levels presumably will have a higher number of possible disablers. Tentatively, one might suggest that the lower acceptance results from additional disabler retrieval: The more disablers are retrieved, the less the extent to which MP and MT will be accepted.

Likewise, De Neys, Schaeken, and d'Ydewalle (2002) compared inference latencies for conditionals with few and many alternatives or disablers. AC inferences took more time when many alternatives were available, whereas MP latencies increased when many disablers were available. De Neys et al. argued that the increased latencies reflected a time-consuming additional counterexample retrieval process. However, the additional retrieval hypothesis was not specifically tested.

The present study provides a more direct test of the characteristics of the counterexample search process by looking explicitly at the effect of the exact number of retrieved alternatives and disablers. This will allow a substantial and unambiguous claim.

Experiment 1 examined the effect of additional counterexample retrieval on conditional inference acceptance by explicitly providing possible counterexamples. As in traditional suppression studies (Byrne, 1989; Byrne et al., 1999; Rumain et al., 1983), we simulated the effect of successful counterexample retrieval by explicitly presenting the counterexamples to participants. The crucial manipulation was that we varied the number of presented counterexamples. Each participant received five different conditionals with the number of presented counterexamples ranging from zero to four. The proposed alternative specification of the search process predicts that there will be an additional suppression effect with every presented counterexample. In Markovits's view, a single counterexample should result in complete inference rejection. Therefore, one should see no additional effects of presenting more than one counterexample.

In Experiment 2 we tested the effect of additional counterexample retrieval without using an explicit presentation. A set of causal conditionals that varied in the number of possible disablers and alternatives (see Cummins, 1995; De Neys et al., 2002) was adopted. In a pretest we first assessed the number of alternatives or disablers a participant could retrieve for every conditional in the set. One month after the pretest, the same participants were invited back for a reasoning task with the conditionals from the pretest. We looked at participants' acceptance ratings of the MP, AC, DA, and MT inferences for each conditional as a function of the number of counterexamples they had been able to retrieve for that specific conditional. Determining whether or not there were graded effects on the inference acceptance as a function of the number of stored counterexamples enabled us to further extend and validate the findings of Experiment 1.

It should be specified that the present study focused on the counterexample search process during everyday reasoning. We adopted realistic causal conditionals and did not instruct participants to reason logically. Also, our adoption of an inference acceptance rating scale allowed participants to give a graded acceptance rating (e.g., see Evans, 2002). With Cummins (1995), one can assume this encourages participants to reason as they would in everyday situations. Recently, Markovits (2002; Quinn & Markovits, 2002) specified that his model primarily describes the retrieval process in a formal deductive reasoning task. There is some debate about whether the same processes account for daily life and more formal reasoning (Evans, 2002; Johnson-Laird & Byrne, 1991; Markovits, 2002; Oaksford, Chater, & Larkin, 2000). It should be noted, then, that as far as this distinction is maintained, the findings of the present study should not be conceived as a mere critique of Markovits's counterexample search characterization, but rather as an attempt to extend it to reasoning in everyday life.

EXPERIMENT 1

In Experiment 1 we sought to determine whether or not presenting more than one counterexample would have an additional effect on the inference acceptance. Traditional suppression studies have only examined the impact of a single presented counterexample. In the proposed alternative specification of the search process, every additional counterexample should have an impact on inference suppression.

Participants in Experiment 1 received five different causal conditionals, with the number of presented counterexamples ranging from zero to four. We presented disablers to half of the participants and alternatives to the other half.

Three consecutive issues are addressed: In order to examine the additional counterexample effect, we had to make sure that there was an effect of presenting one counterexample first. Therefore, we start by establishing whether we can replicate Byrne's (1989) standard findings with the present material and procedure. That is, presentation of a disabler should decrease MP and MT acceptance ratings, whereas an alternative should decrease AC and DA acceptance ratings. Then we address the crucial issue of whether increasing the number of presented counterexamples has an additional effect on the acceptance ratings. Finally, if we find an effect of additional counterexample retrieval, we will examine the precise trend in the data.

Method

Participants. A total of 178 1st-year students of the Educational Sciences Department of the University of Leuven voluntarily participated in the experiment. None of them had received formal logic training and they were all native Dutch speakers.

Materials. The materials were selected from previous pilot work (see De Neys et al., 2002) where 40 participants wrote down as many alternatives or disablers as they could for a set of 20 conditionals

(with 1.5-min generation time for each conditional). Two independent raters scored the generation protocols in order to eliminate unrealistic items and items that were variations of a single idea. The conditionals, item format, instruction, and scoring procedure for the pilot were based on Cummins (1995). For every conditional the mean number of generated counterexamples and the relative frequency of generation of every counterexample were recorded. For the present experiment, we selected five conditionals with many (above the group mean) possible disablers and five conditionals with many (above the group mean) possible alternatives.

The five conditionals with many disablers were used for the disabler presentation manipulation (disablers presentation group), and the other five were adopted for the alternatives presentation manipulation (alternatives presentation group). For every conditional we constructed five different counterexample versions by varying the number of presented counterexamples from zero to four. The counterexamples were taken from the pilot study (see below).

Each participant received a six-page booklet. Page 1 included the task instructions. On top of each of the next five pages the selected conditionals appeared in bold. One of them was presented without a possible counterexample, whereas for the others the versions with one, two, three, or four counterexamples were presented. Thus, each participant received five different conditionals, with the number of presented counterexamples ranging from zero to four. In every booklet, we varied which conditional was used in which counterexample version. We made sure that each of the five counterexample versions of the different conditionals was used equally often (i.e., in approximately one fifth of the booklets). The conditional without counterexample was always presented first, and the remaining conditionals appeared in random order.

The counterexamples were printed below the conditional. Each page also contained three inference problems. The conditionals for the disablers group were embedded in the MP, MT, and AC problems. In the alternatives group we presented AC, DA, and MP problems. The inferences always appeared in the same fixed order (MP, MT, AC, and AC, DA, MP). Below each inference problem was an 11-point rating scale. This resulted in the item format shown below.

The example shows a conditional from the alternatives presentation group with three presented counterexamples embedded in an AC inference. Except for the fact that possible disablers would be presented (e.g., "If the plants are dying, they will not grow quickly"), the item format for the conditionals in the disablers presentation group was completely similar.

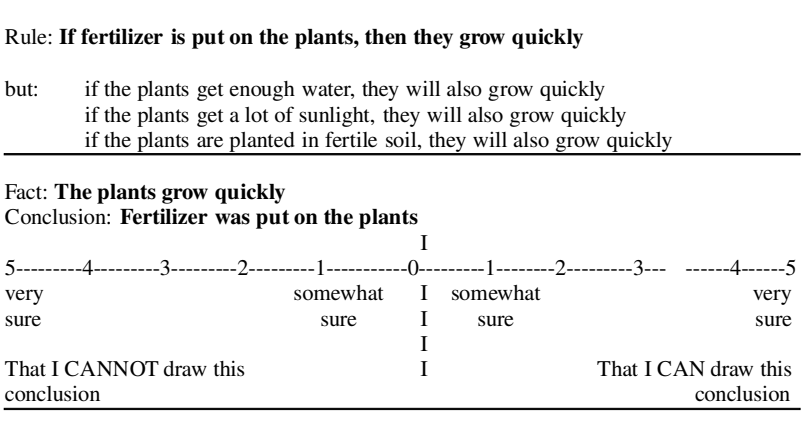
It is important to stress that in the construction and selection of the material, special care was taken to make the explicit presentation of the counterexamples as similar as possible to the actual retrieval. A first issue concerns the selection of the counterexamples. One should note that we did not artificially construct the presented counterexamples, but adopted the material that was generated by the pilot group. This guarantees that the presented counterexamples correspond to real stored background knowledge.

Furthermore, the order in which the counterexamples for a specific conditional were presented corresponded to their frequency of generation (i.e., the percentage of participants in the pilot that generated that specific counterexample). With this manipulation we tried to make sure that the order of presentation reflected the order in which the counterexamples would be actually retrieved. Frequency of generation is often used as an index of associative strength. This factor has been shown to affect counterexample retrieval (see De Neys, Schaeken, & d'Ydewalle, 2003; Quinn & Markovits, 1998). Furthermore, it is commonly assumed that the order in which items are retrieved from memory depends on their associative strength (e.g., Kahana & Loftus, 1999). Therefore, the most frequently generated counterexample (highest associative strength) was presented first, the second most frequently generated one was presented second, and so on. In general, this should guarantee that the presentation order corresponds to the retrieval order.

Finally, the counterexamples were presented as conditionals. This is important because a retrieved counterexample expresses a possible state of affairs and not a factual state of affairs. When we retrieve a counterexample we do not know whether the state of affairs it describes is effectively the case. For example, if you think of "getting enough water" as an alternative for plants growing quickly, you do not know whether or not it is actually the case that the plants got enough water; you only know that the possibility exists that they did so. Therefore, it is important to present the counterexamples in a conditional (e.g., "If the plants get enough water, they will grow quickly") and not in a categorical (e.g., "The plants got enough water") manner. Not taken into account, these issues might limit the contribution of an explicit counterexample manipulation to the examination of the retrieval process. The different conditionals with counterexamples are presented in the Appendix.

Procedure. The experiment was conducted during a regular psychology class. The booklets were randomly given out to students who agreed to participate in the experiment. The instruction page explained the specific item format of the task. Participants were told that the task was to indicate how certain they were that the presented conclusions could be drawn given the presented fact and rule. The instructions also stated that sometimes additional information would be presented that might be used for the judgment. The instruction page further showed an example problem with a copy of the rating scale. In the alternatives group, the example was a DA inference with one presented alternative. In the disablers group the example was an MT inference with one presented disabler.

Participants were instructed to place a mark on the scale number that best reflected their decision. Care was taken to make sure the participants understood the precise nature of the rating scale. Placing a mark on the left side of the scale indicated that they believed



that the conclusion could not be drawn; placing a mark on the right side of the scale indicated that they believed that the conclusion could be drawn. Marking the zero indicated that they could not tell one way or the other.

The participants were not explicitly told to accept the premises as always true or to endorse only conclusions that follow necessarily. Instead, participants were told to evaluate the conclusion by the criteria they personally judged to be relevant. This should encourage participants to reason as they would in everyday situations (Cummins, 1995).

Results

The data from 4 participants were discarded because they did not solve all the inferences. Of the remaining 174 participants, 88 had received booklets from the alternatives presentation group, and 86 participants had received booklets from the disablers presentation group.

The acceptance ratings corresponding to the numbers 5 to 1 on the left side of the 11-point rating scale were recoded and assigned the values −5 to −1 so that increasing numbers corresponded to increased acceptance.

The data in both counterexample groups were analyzed separately. This led to a 3 (inference type, within subjects) × 5 (number of counterexamples, within subjects) design in each group.

For every inference type in both counterexample presentation groups we performed separate multivariate analyses of variance (MANOVAs) on the acceptance ratings with the number of presented counterexamples as a within-subjects factor. In the analyses three consecutive issues are addressed. First, we tested whether there was an overall effect of the number of counterexamples factor. Then we examined the crucial issue of whether presenting more then one counterexample had an additional effect on the acceptance ratings. Third, the precise trend of an eventual additional retrieval effect was analyzed.

We always analyzed the data by participants as well as by materials. However, for each inference there were only five different conditionals. Therefore, we combined the materials analysis for MP and MT (in the disablers group) and DA and AC (in the alternatives group). This increased the n to 10 (see Stevenson & Over, 1995, for a similar approach).

Effect of number of alternatives. The mean acceptance ratings for the three inferences as a function of the presented number of alternatives are shown in Figure 1.

As expected, presentation of alternatives had a significant effect on AC [Rao $R(4,84) = 10.66, p < .0001$] and DA [Rao $R(4,84) = 9.93, p < .0001$] acceptance. Newman–Keuls tests showed that for every number of presented alternatives, AC and DA acceptance was lower than when no alternative was presented. These findings were confirmed by the combined materials analysis on AC and DA

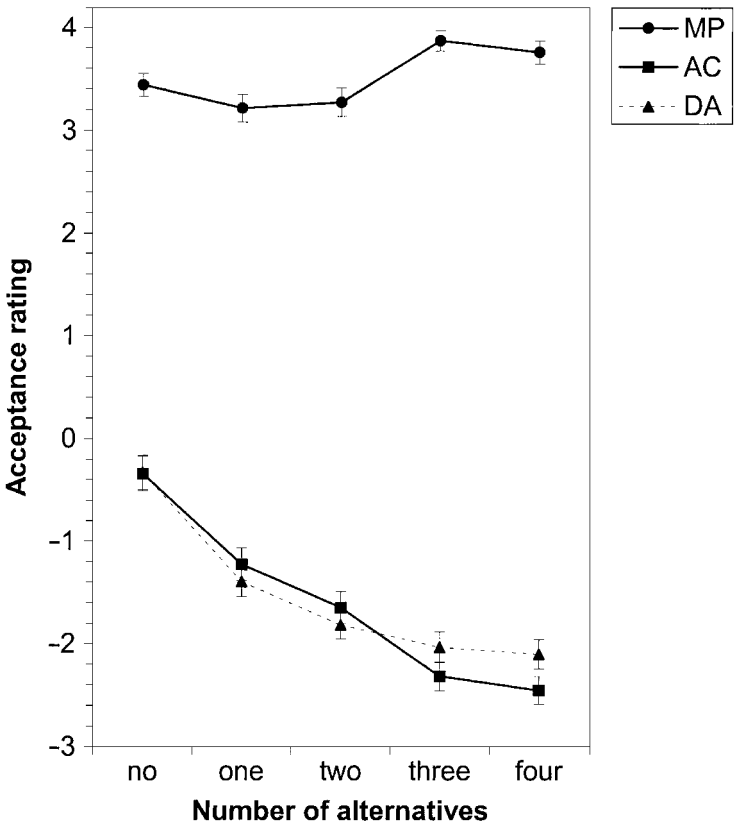


Figure 1. Inference acceptance as a function of the number of presented alternatives. The rating scale ranged from −5 (*very sure I cannot draw this conclusion*) to +5 (*very sure I can draw this conclusion*). Vertical lines depict standard errors of the means.

[Rao $R(4,6) = 60.96, p < .0001$]. Both participants and materials analyses indicated that the alternatives had no significant impact on MP. Although there were only five conditionals for the materials analysis on MP, it was clear that there were no meaningful trends in the data. These results replicate previous suppression findings.

In order to establish the crucial question of whether presenting more than one alternative has an additional effect on DA and AC acceptance, we examined whether there was still an effect of number of alternatives when only the levels with one, two, three, and four alternatives were compared. For AC this was indeed the case [Rao $R(3,85) = 7.01, p < .0003$]. Trend analysis showed that there was a significant negative linear trend [$F(1,87) = 20.84, MS_e = 83.35, p < .0001$], whereas higher order trends were not significant. This implies that every additional alternative further decreased AC acceptance. However, there was no clear effect of additional alternatives on DA [Rao $R(3,85) = 1.86, p < .15$].

The materials analysis established that there was a marginal effect of additional alternatives on combined DA and AC acceptance [Rao $R(3,7)=3.37, p < .09$] and that this effect was linear [$F(1,9) = 8.26, MS_e = 0.57, p < .02$]. These effects are depicted in Figure 1.

Effect of number of disablers. Figure 2 shows the mean acceptance ratings for the three inferences as a function of the presented number of disablers.

On both MP [Rao $R(4,82) = 10.23, p < .0001$] and MT [Rao $R(4,82) = 6.61, p < .0001$] we obtained the expected effect of disabler presentation [combined material analysis, Rao $R(4,6) = 13.44, p < .005$]. A Newman–Keuls test made it clear that for every number of presented disablers, MP acceptance was lower than when no disabler was presented. Except for the difference between no and one presented disabler, this was also the case for acceptance of MT.

Presentation of disablers also affected AC [Rao $R(4,82) = 4.67, p < .002$]. Contrary to the the results for MP and MT, presentation of disablers led to a higher AC acceptance. For the material analysis on AC, only five conditionals were available. Although the effect did not reach significance, a similar trend as in the participants analysis was observed. For both participants and materials, Newman–Keuls tests showed that for any number of presented disablers, AC acceptance was higher than when no disablers were present.

The crucial manipulation of presenting more than one counterexample had a significant effect on MP

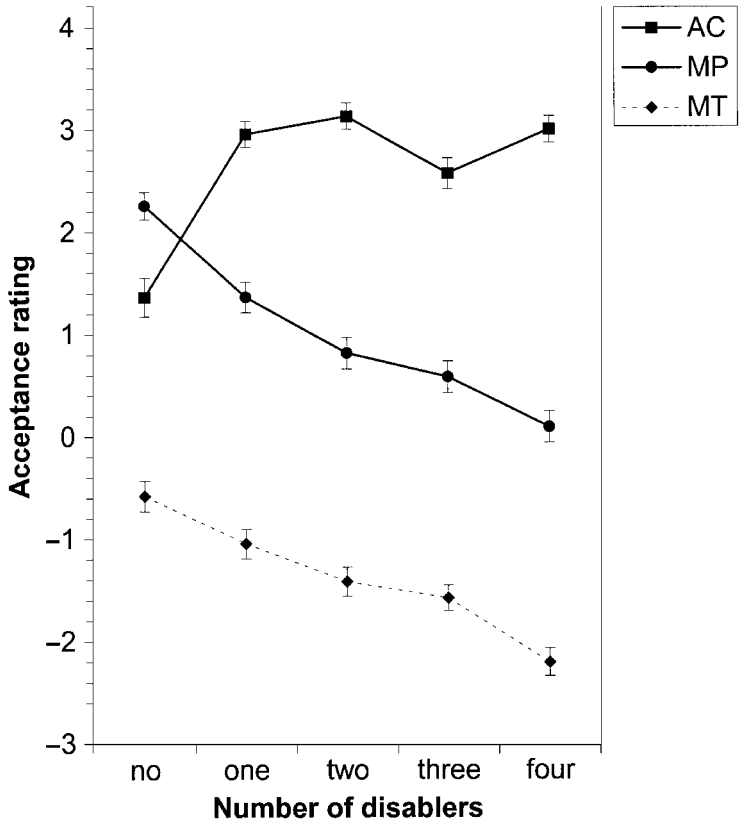


Figure 2. Inference acceptance as a function of the number of presented disablers. The rating scale ranged from −5 (*very sure I cannot draw this conclusion*) to +5 (*very sure I can draw this conclusion*). Vertical lines depict standard errors of the means.

[Rao $R(3,83) = 6.77, p < .0005$] and MT [Rao $R(3,83) = 5.56, p < .002$] acceptance. Trend analysis showed that for both MP [$F(1,85) = 20.46, MS_e = 3.42, p < .0001$] and MT [$F(1,85) = 16.02, MS_e = 3.58, p < .0002$] there was a significant negative linear trend in the acceptance data, whereas higher order trends were not significant. Thus, every disabler that is retrieved in addition to the first one will result in a further decrease of MP and MT acceptance ratings. These findings were confirmed by the combined materials analysis [Rao $R(4,6) = 13.44, p < .004$; significant linear trend, $F(1,9) = 10.18, MS_e = 0.68, p < .015$; no higher order trends].

Both the participants and materials analysis clearly established that presenting more than one disabler had no further effect on AC acceptance. Thus, although disabler presentation led to a higher AC acceptance, the number of additionally presented disablers had no further impact.

Discussion

By showing that explicit presentation of an alternative decreased AC and DA acceptance, whereas presentation of disablers resulted in lower MP and MT acceptance, we replicated previous suppression findings (e.g., Byrne, 1989; Byrne et al., 1999; Rumain et al., 1983).

The traditional observations were extended by the finding that suppression is affected by the number of presented alternatives and disablers. MP and MT acceptance linearly decreased with every additionally presented disabler, whereas AC acceptance ratings showed a similar linear decrease for every additionally presented alternative. The effect of additional alternatives on DA was less clear. We will come back to this issue later on.

Presentation of a disabler also resulted in higher AC acceptance. A similar effect of disablers on AC (and DA) acceptance has already been reported (e.g., De Neys et al., 2002; Liu et al., 1996; Markovits & Potvin, 2001). De Neys et al. (2002) argued that retrieval of disablers would have priority over alternative retrieval. Due to the resource-limited nature of the memory retrieval process, retrieval of disablers would thereby hinder subsequent alternative retrieval. Thus, by affecting the efficiency of the alternative search process, disabler retrieval can result in higher AC and DA acceptance. Because of the priority of the disabler search, retrieval of alternatives would not bias the disabler search. A similar mechanism could account for the present AC observation. Although participants do not have to search the disablers themselves, the load caused by the processing of a presented disabler (e.g., incorporation into a mental model) might result in a less efficient search for alternatives.

The results of Experiment 1 support the alternative specification of the counterexample search process. As predicted, inference acceptance decreased with every additionally available counterexample. This implies that inference suppression is not an all-or-nothing phenomenon but depends on the number of available counterexamples.

Although the results establish an important characteristic of the suppression effect, the implications for establishing the characteristics of the counterexample retrieval process can be debated. Although special care was taken to make the counterexample presentation as similar as possible to the actual retrieval, one could argue that adopting an additional counterexample when it is presented is not the same thing as searching it oneself. The present results do show that people will use additional counterexamples when they are available. Nevertheless, the findings do not necessarily imply that people will search for additional counterexamples themselves. Thus, in order to specify the crucial search characteristic of the retrieval process we needed an additional test without explicit counterexample presentation.

EXPERIMENT 2

In Experiment 1 inference suppression linearly increased with every presented counterexample. Because of the explicit presentation procedure, we cannot conclude that the actual search process retrieves additional counterexamples. However, the Experiment 1 findings do imply that if people would indeed search and retrieve additional counterexamples themselves, we should see a similar linear decreasing acceptance pattern. In Experiment 2 we looked at participants' inference acceptance as a function of the number of counterexamples they could retrieve for a conditional. We sought to validate the findings of Experiment 1 by checking whether the same graded trends would be observed.

We adopted a set of causal conditionals that varied in the number of possible disablers and alternatives (see Cummins, 1995; De Neys et al., 2002). In a pretest we first assessed the number of alternatives or disablers a participant could retrieve for every conditional in the set. One month after the pretest, the same participants were invited back for a reasoning task with the conditionals from the pretest. We looked at participants' acceptance ratings of the MP, AC, DA, and MT inferences for each conditional as a function of the number of counterexamples they had been able to retrieve for that specific conditional.

Markovits's specification of the search process predicts that up to a certain number of available counterexamples, inferences will tend to be accepted. After successful retrieval the inferences will be rejected and additionally available counterexamples will not affect inference acceptance any further. On the basis of this specification we expect a stepwise trend in the acceptance ratings as a function of the number of counterexamples one has stored. The alternative specification we propose should result in gradually decreasing acceptance ratings with every additionally available counterexample.

It is crucial to stress the within-subjects nature of the analyses in the present study. The number of stored counterexamples is of course directly associated with the probability of retrieving a single counterexample. Thus, if we compared different groups of participants (e.g., groups that retrieved one, two, three, or more counterexamples for a specific conditional), we could not attribute a graded effect to additional disabler retrieval. Indeed, it could simply be claimed that there will be a larger number of par-

ticipants that retrieve a single counterexample in the successive groups. Therefore, we always compared the inference acceptance of the same participants for conditionals for which they retrieved a different number of disablers or alternatives.

Likewise, the experiment's crucial contribution lies in the examination of the nature of the acceptance rating trends. Previous studies (e.g., Thompson, 1995, 2000) have shown a correlation between a conditional's number of possible counterexamples and the degree of inference acceptance. However, a mere correlation does not allow us to address the present additional retrieval issue since it is consistent with different trends. Therefore, the present analyses focus on the actual pattern in the acceptance ratings.

Method

Pretest. A set of 20 conditionals (based on Cummins, 1995) that varied in the number of possible alternatives and disablers was adopted for the pretest. Participants were asked to write down as many alternatives or disablers as possible for each conditional (with 1.5-min generation time for each conditional).

Two independent raters scored the generation protocols in order to eliminate unrealistic items and items that were variations of a single idea. Item format, instructions, and scoring procedure were similar to those in Cummins (1995). For each participant we recorded the number of alternatives or disablers she/he retrieved for every conditional.[2]

Participants. Forty 1st-year psychology students participated in the experiment. None of them had received formal logic training and they were all native Dutch speakers. Twenty participants generated alternatives in the pretest, and the other half generated disablers.

Materials. Sixteen conditionals from the pretest were selected for the reasoning task. The conditionals constituted a 2 (few/many) × 2 (alternatives/disablers) design with four items per cell (see De Neys et al., 2002). The 16 conditionals were embedded in the four (MP, DA, MT, and DA) inference types, producing a total of 64 inferences for each participant to evaluate.

The experiment was run on computer. Each argument was presented on screen together with a 7-point rating scale and accompanying statements. This resulted in the following format shown below.

Each of the 64 arguments was presented in this way. The premises, conclusion, and typed number were always presented in yellow. The remaining text appeared in white on a black background.

Procedure. Participants were run in groups of 2 to 8. Approximately 1 month (28 to 35 days) after the pretest participants were called in for the reasoning task. Instructions for the reasoning task were presented verbally and on screen. They showed an example item that explained the specific task format. Participants were told that the task was to decide whether or not they could accept the conclusions. Care was taken to make sure participants understood the precise nature of the rating scale.

Participants used the keypad to type the number reflecting their decision. The 64 items were presented in random order. The experimental session was preceded by one practice trial. As in Experiment 1, participants could evaluate the conclusions by the criteria they personally judged relevant.

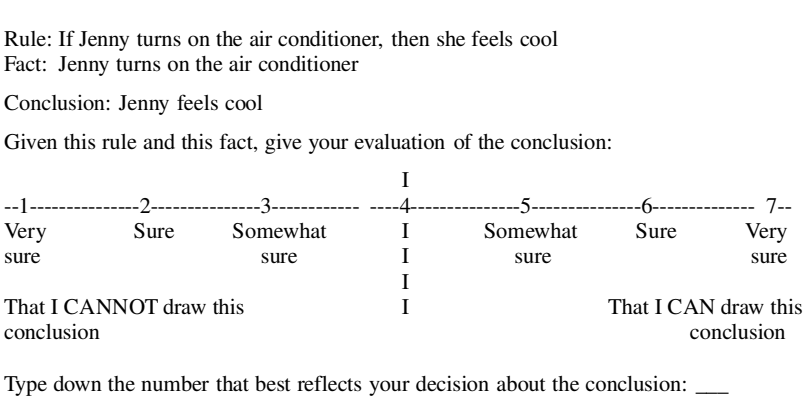
Results

Three participants could not be contacted for the reasoning task. This resulted in a total of 19 participants in the disablers retrieval group and 18 participants in the alternatives retrieval group.

A first control analysis[3] established that the inferences were not affected by the specific generation of alternatives or disablers 1 month earlier: There were no significant differences in the inference performance of participants that were asked to produce disablers and those that were asked to produce alternatives.

For the main analysis we grouped all conditionals for which a participant could retrieve no or one, two, three, or four or more counterexamples. Since the majority of participants generated at least one counterexample for every conditional we combined the no and one groups. Likewise, since rarely more than four counterexamples were generated, these conditionals were combined with the four group. On average, participants generated no or one, two, three, or four or more alternatives for 3.22 (SD = 1.9), 3.11 (SD = 1.75), 3.67 (SD = 1.78), and 6.00 (SD = 2.81) conditionals, respectively. The average number of conditionals in the successive number of disablers retrieval groups was 1.53 (SD = 1.07), 3.53 (SD = 1.35), 5.37 (SD = 1.71), and 5.58 (SD = 1.80), respectively. For every participant we calculated the mean inference acceptance for the different conditionals in every number of counterexamples group. For every inference type, these means were subjected to a MANOVA with the number of retrieved alternatives or disablers as a within-subjects factor.

Missing observations (e.g., a participant had no conditionals for which two alternatives were retrieved) were set



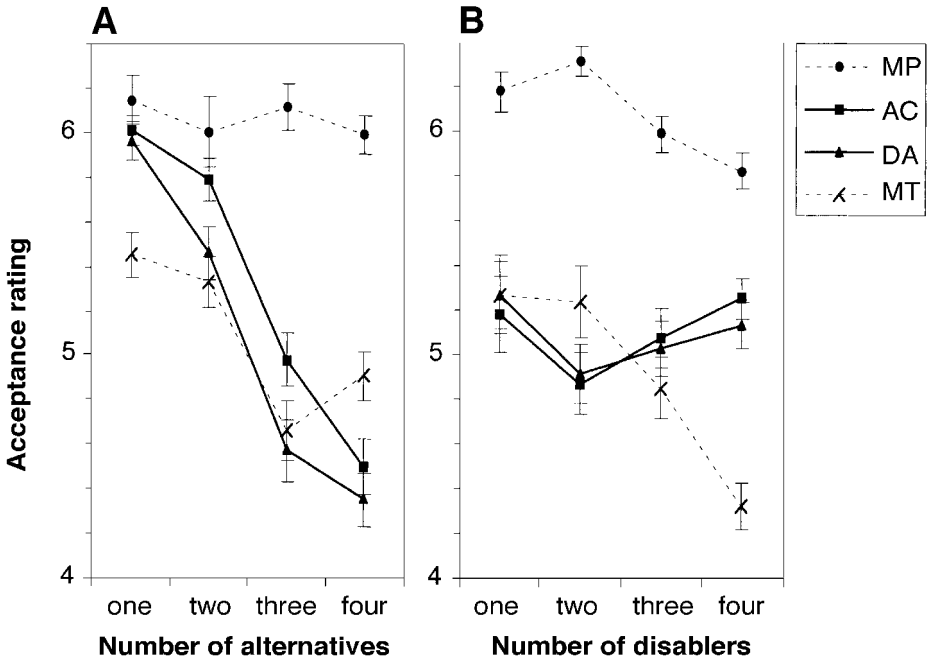


Figure 3. Inference acceptance as a function of the number of alternatives (A) or disablers (B) participants could retrieve for a conditional. The ratings scale ranged from 1 (*very sure I cannot draw this conclusion*) to 7 (*very sure I can draw this conclusion*). Vertical lines depict standard errors of the means.

to the overall mean. In both the alternative and disabler retrieval groups this affected less than 4% of the observations.

The MANOVA indicated that the number of available alternatives affected AC [Rao $R(3,15) = 18.15, p < .001$] and DA [Rao $R(3,15) = 16.08, p < .001$] acceptance, and the disablers affected both MP [Rao $R(3,16) = 3.08, p < .06$] and MT [Rao $R(3,16) = 3.88, p < .03$].

As Figure 3A makes clear, the AC and DA ratings did not show a stepwise trend as a function of the number of retrieved alternatives. Trend analyses established that for both AC [$F(1,17) = 59.62, MS_e = 0.43, p < 0.001$] and DA [$F(1,17) = 47.36, MS_e = 0.62, p < .001$], there was a significant negative linear trend, and that higher order trends were not significant. Likewise, acceptance of MP [$F(1,18) = 7, MS_e = 0.27, p < .02$] and MT [$F(1,18) = 6.26, MS_e = 1.59, p < .03$] also linearly decreased with every retrieved disabler (see Figure 3B). Higher order trends were not significant. These observations are in line with the findings of Experiment 1.

We also examined the individual acceptance patterns for every participant. This was necessary to eliminate further interpretation complications. It could for example be the case that different participants have a very different retrieval threshold. That is, the number of stored counterexamples that is sufficient for successful retrieval of a single counterexample during the reasoning task may vary extremely between participants. If this is the case, then it might be claimed that the graded inference acceptance effects are the result of aggregating individual stepwise trends instead of reflecting additional counterexample retrieval. Thus, the individual patterns would all show a stepwise trend, but the steps would be located at different positions. In order to eliminate such a confound, we looked at the individual acceptance patterns.

The individual acceptance patterns were classified in three groups. If a participant gave three or four successive decreasing ratings, her/his acceptance pattern was classified as "graded." If there was a clear single step in the pattern, it was classified as "stepwise." A rather liberal criterion was adopted: The step had to be larger in size than the differences between the hypothesized equal ratings. For example, an acceptance rating pattern of 4, 5, 2, 3 for conditionals with, respectively, one, two, three, and four counterexamples would be classified as a stepwise pattern with the retrieval threshold at three counterexamples. Likewise,

Table 1
Percentage of Participants Whose Acceptance Rating Pattern Showed a Graded, Stepwise, or Other Trend as a Function of the Number of Disablers (MP, MT) or Stored Alternatives (AC, DA)

	Classification		
Inference Type	Graded	Stepwise	Other
Number of Disablers (n = 19)			
MP	63	0	37
MT	68	11	21
Number of Alternatives (n = 18)			
AC	78	17	5
DA	56	39	5

a pattern like 6, 2, 3, 2 would be classified as a stepwise pattern with the retrieval threshold at two counterexamples. Patterns that could not be classified in these two categories were labeled "other" (e.g., a pattern like 5, 3, 4, 2). Table 1 shows the classification results.

Table 1 shows that the graded trends in Figure 3 cannot be attributed to the aggregation of individual stepwise trends. For every inference type the acceptance rating for the majority of participants showed a graded acceptance trend. For the participants that did show a stepwise trend, the step or "threshold" was always at two (MT, DA) or three (AC, DA) stored counterexamples. It is interesting to note that both for the disablers [0% MP vs. 11% MT; $n = 19$, $p < .08$] and for the alternatives [17% AC vs. 39% DA; $n = 18$, $p < .08$], the stepwise trends seemed to pop up especially for the "denial" inferences (DA and MT).

For completeness, we report that the number of alternatives also affected MT [Rao $R(3,15) = 9.14$, $p < .002$] acceptance. As for DA and AC, the MT trend was linear [$F(1,17) = 22.89, MS_e = 0.21, p < .001$]. MP also tended to decrease with the number of alternatives, but the trend was not significant. Likewise, AC and DA acceptance showed an opposite trend, with increasing acceptance when two and more disablers were available, but the effect did not reach significance.

Discussion

The results of Experiment 2 imply that every alternative or disabler that can be retrieved has an impact on inference acceptance. Every retrieved alternative decreased AC and DA acceptance, whereas every retrieved disabler resulted in lower MP and MT acceptance. These graded effects of up to four different numbers of available counterexamples cannot be explained if the semantic search process during conditional reasoning would stop after successful retrieval of a single counterexample.

The classification of the individual acceptance rating patterns established that the findings cannot be attributed to an aggregation confound. Most participants showed a graded acceptance trend. However, the individual classification also indicated an increase in stepwise acceptance patterns on the DA and MT inferences. Thus, there does seem to be a tendency to stop the search process after retrieval of a single counterexample for these "denial" inferences.

In Experiment 1 we did not observe an effect of additional alternatives on DA, either. Similarly, the evidence for additional counterexample retrieval in the latency findings of De Neys et al. (2002) was clear only for the MP and AC inferences.

These findings might indicate that the semantic search process during conditional reasoning is affected by inference complexity. DA and MT are more complex inferences than AC and MP. DA and MT involve negations (thus "denial" inferences), and reasoning theories typically state that these demand more cognitive (working memory) resources (Braine & O'Brien, 1998; Johnson-Laird & Byrne, 1991; Oaksford et al., 2000). Now, semantic memory retrieval is known to be a (working memory) resource-demanding process (e.g., Rosen & Engle, 1997). By burdening the available resources, the additional need to process negations could thus affect the extent of the search process for DA and MT. Due to a lack of resources, it will be less likely that additional counterexamples will be searched.

We primarily focused on the standard (e.g., Byrne, 1989; Byrne et al., 1999; Cummins, 1995) effects of disablers on MP and MT acceptance and alternatives on AC and DA acceptance, but there were also signs of extra trends in the data: MP, and especially MT, acceptance tended to go down with increasing number of alternatives, whereas there was some indication of an opposite trend for the effect of number of disablers on AC and DA acceptance. Similar effects have been reported previously (see De Neys et al., 2002, for a detailed discussion). The cause of the MP and MT trends seems to lie in the fact that in the set of conditionals we adopted, the numbers of alternatives and disablers were positively correlated ($r_s = .37$, n.s.; see De Neys et al., 2002). Thus, conditionals with more alternatives will also have somewhat more disablers. Since more disablers will become available, MP and MT acceptance will tend to go down with increasing number of alternatives. On the other hand, as reported in Experiment 1, De Neys et al. (2002) argued that disabler retrieval may affect the efficiency of the search for alternatives. Such a mechanism would explain the trend toward higher AC and DA acceptance when more disablers become available.

One might object that the retrieval pretest in Experiment 2 showed us only the number of counterexamples a participant had stored for the different conditionals. Obviously, there is no direct evidence that these stored counterexamples were actually retrieved during the reasoning task. Here, it is crucial to stress the relation with the findings of Experiment 1. The explicit presentation illustrated the kind of effect the different number of counterexamples should have on inference acceptance. The fact that (at least for MP and AC) the same linear trends were observed in both experiments supports the additional counterexample retrieval hypothesis.

A final objection concerns the fact that even in our individual, within-subjects analysis we still aggregated over several items (i.e., the mean inference acceptance rating in every number of counterexamples group was calculated over approximately four conditionals). Hence, one might argue that the individual graded acceptance patterns resulted from averaging across items or conditionals. That is, in the successive counterexample groups there would simply be more conditionals for which a single counterexample is retrieved. We believe that this alternative explanation is implausible. It implies, for example, that frequently a participant might retrieve three or more counterexamples for a conditional in the generation task but nevertheless that the same participant would not retrieve a single counterexample for the conditional during reasoning.[4]

With respect to the possible procedural complications, it is interesting that there is also converging evidence for the present findings. In a recent thinking-aloud study

(Verschueren, Schaeken, De Neys, & d'Ydewalle, 2003) without a generation pretest or explicit counterexample presentation, we observed for example that participants spontaneously produced two, three, or more counterexamples in the evaluation of a single MP or AC argument. Such a result would be hard to explain if people stop the search after retrieval of a single counterexample. The consistent results in these experiments indicate that the additional retrieval findings should not be attributed to a procedural artifact.

GENERAL DISCUSSION

Manipulating the number of presented counterexamples in Experiment 1 showed that inference suppression is not an all-or-nothing phenomenon but depends on the number of available counterexamples. Experiment 2 extended these findings by showing that the same effects are observed when we look at counterexamples that participants retrieve themselves. Taken together, the findings of Experiments 1 and 2 support the alternative specification of the counterexample search process during conditional reasoning: After successful retrieval of a counterexample, the search process will continue, and every additionally retrieved counterexample will further decrease inference acceptance.

The present findings also indicate that the counterexample search process is not occurring in complete cognitive isolation. For DA and MT, the additional retrieval findings were less clear. In line with previous findings (e.g., De Neys et al., 2002) it is suggested that the additional processing requirements for these inferences burden the counterexample search process. Thus, due to a higher cognitive load, searching additional counterexamples after successful retrieval would be less likely for DA and MT.

In considering the statement that "every counterexample counts," one should further bear in mind that we looked at retrieval only up to four counterexamples. It is thus possible that after four items the impact of subsequently retrieved counterexamples will taper off. Note however that in Experiment 2 people rarely generated more than four counterexamples. If retrieving counterexamples is indeed resource demanding, retrieving more than four counterexamples while reasoning should also be rather rare. In this sense our generalization is not entirely unwarranted.

The results of this study are relevant to a number of issues in the conditional reasoning domain. We discuss the implications for Markovits's reasoning model, probabilistic reasoning theories, and the debate on the nature of the suppression effect.

Markovits's Reasoning Model

Markovits's (e.g., Janveau-Brennan & Markovits, 1999; Markovits, 2000; Markovits et al., 1998; Markovits & Quinn, 2002) original specification of the counterexample search process does not address the impact of additional counterexample retrieval. The search process is assumed to stop after the successful retrieval of a single counterexample. The clear additional retrieval effects on AC and MP show that this is not the case. In order to account for these effects the initial model needs to be revised.

It should be specified that Markovits and Barrouillet (2002) recently acknowledged the possibility of a continued search process and a resulting additional counterexample retrieval. However, the framework does not yet take count of the impact of additionally retrieved counterexamples: Additional retrieval is "allowed," but whether this can affect the inference acceptance is not addressed.

The framework's main problem with respect to the additional counterexample findings seems to lie in its incorporation of the standard mental models theory (Johnson-Laird, 1983). Markovits stated that the outcome of the semantic search process will affect the kinds of mental models a reasoner builds. A standard mental model either licenses an inference or does not. There are no intermediate or graded states of inference acceptance. Whenever a reasoner constructs the additional counterexample model, the inferences are completely rejected. Therefore, it has been argued that it is hard to explain graded inference effects in standard mental model terms (e.g., George, 1997; Stevenson & Over, 1995).

However, the recent extension of the mental models theory toward extensional reasoning (Johnson-Laird, 1994; Johnson-Laird, Legrenzi, Girotto, Legrenzi, & Caverni, 1999) offers an interesting revision approach. Consider for example the conditional "If Jenny turns on the air conditioner, then she feels cool." On the basis of Johnson-Laird et al. it could be argued that a reasoner will construct a specific model for every retrieved alternative (e.g., clothes off–cool, window open–cool, shower–cool . . .) instead of immediately building a more general model after retrieval of a single alternative (e.g., not air conditioner–cool). The proportion of constructed "counterexample models" would then determine the extent to which inferences will be accepted (see George, 1997; Stevenson & Over, 1995, for related suggestions).

Although such a revision could in theory account for the additional counterexample findings, it faces an important problem (e.g., George, 1997). A basic assumption of mental models theory is that every constructed model puts a load on working memory. Consequently, reasoning with more than three different models has been shown to be extremely difficult (e.g., Barrouillet & Lecas, 1999; Johnson-Laird & Byrne, 1991). In the present study we observed graded acceptance effects of up to four different numbers of retrieved counterexamples. Together with the initial model this would call for the construction of five different models for an AC or MP inference. Therefore, the computational complexity of the suggested mental models revision would exceed people's cognitive abilities.

The recent work of Schaeken, Vander Henst, and Schroyens (in press) on isomeric mental models might, however, provide a solution to the "computational complexity" caveat. The authors argued that people can con-

struct more economical mental models than traditionally assumed. They showed that when two models are redundant in that they share the same information, people can combine them into a single "isomeric" model. Results indicated that with indeterminate relational inferences (e.g., "Bart is to the left of Mark. Mark is to the right of Jan. Is Bart to the right of Jan?"), instead of constructing two possible specific models (e.g., Jan → Bart → Mark and Bart → Jan → Mark), people rather constructed a single isomeric model (e.g., Bart ↔ Jan → Mark) that represented the same crucial indeterminacy. The basic idea is that people will avoid building a model of a piece of information that is already represented. This idea can be extended to the present situation.

Indeed, all the specific models that represent the different alternatives, for example, refer to the same consequent term (e.g., "Jenny feels cool"). One could suggest then that people will combine the different specific models into a single "isomeric" model. The resulting model would not be specific since the concrete counterexamples would not be individually represented. On the other hand it would not be general in the sense that it would keep track of the crucial number of retrieved counterexamples. This would allow a considerable decrease in working memory load while the crucial number information would nevertheless be maintained. Though interesting, the proposal is of course speculative and needs to be tested properly.

Daily Life or Formal Deductive Reasoning?

Our study did not examine the counterexample retrieval process in a formal, deductive reasoning task, but rather in a situation closer to everyday life reasoning. Participants were not specifically instructed to reason logically and were allowed to give a graded acceptance rating. Therefore, if a sharp distinction is maintained between formal and daily life reasoning, the present findings should not be immediately generalized to reasoning in a formal, deductive reasoning task. Again, note that Markovits developed his model in the context of formal, deductive reasoning. Hence, the present findings do not necessarily refute Markovits's original search process specification. That is, it might be the case that in a formal deductive reasoning task, people do stop the search and take only one counterexample into account. The present work can be best characterized as an adaptation and extension of Markovits's search process specification to everyday life reasoning. This extension is nevertheless crucial for the final evaluation of Markovits's model. Accounting for people's daily life reasoning behavior is considered the ultimate goal of any reasoning model (Johnson-Laird, 1983; Oaksford & Chater, 1998). The present findings do indicate that the model will need to be fine-tuned to encompass daily life reasoning.

Probabilistic Reasoning Models

According to the probabilistic approach toward human reasoning (e.g., Liu et al, 1996; Oaksford & Chater, 1998, 2001; Oaksford et al., 2000), reasoning is essentially probabilistic in nature. The MP inference for example would require participants to calculate the value of an "exceptions parameter" (i.e., the probability of "not-q given p"; see also Stevenson & Over, 1995). This parameter represents the probability that "exceptions" (disablers) will occur. The higher the exceptions value, the less likely that MP will be accepted.

However, a major problem for this approach is that it is not clear how people would derive the necessary probabilities. Indeed, probabilistic approaches toward human reasoning have typically focused on the computational level of explanation (i.e., "what" is computed, not "how"; see Oaksford & Chater, 1998, 2001). The finding that the number of retrieved counterexamples determines the degree of inference acceptance allows the probabilistic frameworks to specify that (as Oaksford and Chater suggested) it is the outcome of the counterexample retrieval process that determines the crucial probabilities. The higher the number of retrieved disablers, for example, the higher the exceptions parameter will become and the less likely that MP will be accepted. As such, the characterization of the counterexample retrieval process can contribute to a more fine grained, algorithmic level specification of the probabilistic reasoning accounts.

The Nature of Inference Suppression

Finally, we note the relevance of the present findings to the debate on the nature of the suppression effect (see Byrne et al., 1999, for an overview). Byrne has maintained that the suppression effect arises because whenever a counterexample is available and explicitly represented, certain inferences are no longer supported. The inference is thus suppressed, but the status of the conditional itself remains unaffected. However, from the findings on reasoning with uncertain conditionals (e.g., George, 1997; Liu et al., 1996; Stevenson & Over, 1995), it has been argued that suppression arises because the counterexample may lead people to doubt the conditional. Conditionals would be interpreted probabilistically, and a counterexample would directly lower the certainty status of the conditional itself. Traditionally, the graded suppression effects of manipulating $P(q/p)$ on MP and MT acceptance in these studies have been interpreted as support for the "conditional doubt" position.

In line with Byrne, the present findings indicate that graded suppression can be explained without altering the certainty status of the conditional: Graded suppression can simply express the number of retrieved counterexamples. We have already argued that accounting for the additional counterexample retrieval would require an extension of standard mental models theory. Nevertheless, the point is that our findings indicate that (at least with realistic, causal conditionals) it is not necessary to assume that the conditional itself is doubted to explain graded suppression effects.

Conclusion

This study has supplemented traditional reasoning studies by establishing the characteristics of the coun-

terexample search process during everyday conditional reasoning. We complemented Markovits's first specification of the search process by showing that when the cognitive system is not burdened by negation processing, the search continues after retrieval of a single counterexample. Thereby, every additionally retrieved counterexample will have an additional impact on the inference acceptance.

REFERENCES

Anderson, J. R. (1983). *The architecture of cognition.* Cambridge, MA: Harvard University Press.

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

Braine, M. D. S., & O'Brien, D. P. (Eds.) (1998). *Mental logic.* Mahwah, NJ: Erlbaum.

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

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

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. (2003). Causal conditional reasoning and strength of association: The disabling condition case. *European Journal of Cognitive Psychology*, **15**, 161-176.

Dieussaert, K., Schaeken, W., Schroyens, W., & d'Ydewalle, G. (2000). Strategies during complex conditional inferences. *Thinking & Reasoning*, **6**, 125-160.

Evans, J. St. B. T. (2002). Logic and human reasoning: An assessment of the deduction paradigm. *Psychological Bulletin*, **128**, 978-996.

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

Gillund, G., & Shiffrin, R. M. (1984). A retrieval model for both recognition and recall. *Psychological Review*, **91**, 1-67.

Janveau-Brennan, G., & Markovits, H. (1999). The development of reasoning with causal conditionals. *Developmental Psychology*, **35**, 904-911.

Johnson-Laird, P. N. (1983). *Mental models: Towards a cognitive science of language, inference, and consciousness.* New York: Cambridge University Press.

Johnson-Laird, P. N. (1994). Mental models and probabilistic thinking. *Cognition*, **50**, 189-209.

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

Johnson-Laird, P. N., & Byrne, R. M. J. (1994). Models, necessity, and the search for counterexamples. *Behavioral & Brain Sciences*, **17**, 775-778.

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

Kahana, M., & Loftus, G. (1999). Response time versus accuracy in human memory. In R. J. Sternberg (Ed.), *The nature of cognition* (pp. 323-340). Cambridge, MA: MIT Press.

Liu, I., Lo, K., & Wu, J. (1996). A probabilistic interpretation of "if–then." *Quarterly Journal of Experimental Psychology*, **49A**, 828-844.

Markovits, H. (1986). Familiarity effects in conditional reasoning. *Journal of Educational Psychology*, **78**, 492-494.

Markovits, H. (2000). A mental model analysis of young children's conditional reasoning with meaningful premises. *Thinking & Reasoning*, **6**, 335-347.

Markovits, H. (2002). *Is inferential reasoning probabilistic?* Manuscript submitted for publication.

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

Markovits, H., Fleury, M., Quinn, S., & Venet, M. (1998). The development of conditional reasoning and the structure of semantic memory. *Child Development*, **69**, 742-755.

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

Matalon, B. (1962). Etude génétique de l'implication. In E. W. Beth, J. B. Grize, R. Martin, B. Matalon, A. Naess, & J. Piaget (Eds.), *Implication, formalisation et logique naturelle* (Études d'Epistémologie Génétique, Vol. 16, pp. 69-93). Paris: Presses Universitaires de France.

Oaksford, M., & Chater, N. (1998). *Rationality in an uncertain world: Essays on the cognitive science of human reasoning.* Hove, U.K.: Psychology Press.

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

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

Politzer, G. (in press). Premise interpretation in conditional reasoning. In D. Hardman & L. Macchi (Eds.), *Reasoning and decision making.* New York: Wiley.

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

Quinn, S., & Markovits, H. (2002). Conditional reasoning with causal premises: Evidence for a retrieval model. *Thinking & Reasoning*, **8**, 179-191.

Rosen, V. M., & Engle, R. W. (1997). The role of working memory capacity in retrieval. *Journal of Experimental Psychology: General*, **126**, 211-227.

Rumain, B., Connell, J., & Braine, M. D. S. (1983). Conversational comprehension processes are responsible for reasoning fallacies in children as well as adults. *Developmental Psychology*, **19**, 471-481.

Schaeken, W., Vander Henst, J. B., & Schroyens, W. (in press). The mental models theory of relational reasoning: Premise relevance, conclusion phrasing and cognitive economy. In W. Schaeken, A. Vandierendonck, W. Schroyens, & G. d'Ydewalle (Eds.), *The mental models theory of reasoning: Extensions and refinements.* Mahwah, NJ: Erlbaum.

Staudenmayer, H. (1975). Understanding conditional reasoning with meaningful propositions. In R. J. Falmagne (Ed.), *Reasoning: Representation and process in children and adults* (pp. 55-79). Hillsdale, NJ: Erlbaum.

Stevenson, R. J., & Over, D. E. (1995). Deduction from uncertain premises. *Quarterly Journal of Experimental Psychology*, **48A**, 613-643.

Thompson, V. A. (1994). Interpretational factors in conditional reasoning. *Memory & Cognition*, **22**, 742-758.

Thompson, V. A. (1995). Conditional reasoning: The necessary and sufficient conditions. *Canadian Journal of Experimental Psychology*, **49**, 1-60.

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

Vadeboncoeur, I., & Markovits, H. (1999). The effect of instruction and information retrieval on accepting the premises in a conditional reasoning task. *Thinking & Reasoning*, **5**, 97-113.

Verschueren, N., Schaeken, W., De Neys, W., & d'Ydewalle, G. (2003). *The difference between generating counterexamples and using them during reasoning.* Manuscript submitted for publication.

NOTES

1. In line with previous research we use the term *suppression effect* to refer to the effect of disabler and alternative retrieval on inference acceptance. However, see Dieussaert, Schaeken, Schroyens, and d'Ydewalle (2000) for a critique of the label *suppression.*

2. Precise materials, procedures, and results of the pretest have previously been reported in De Neys et al. (2002, Experiment 1). The materials for Experiment 1 were taken from the same study.

3. Since the number of possible alternatives and disablers of the conditionals in the reasoning task varied systematically, the data could be analyzed as a 2 (few/many) × 2 (alternatives/disablers) × 4 (inference type) within-subjects design (see Cummins, 1995). The kind of generated counterexample (disablers or alternatives) was entered as a between-subjects factor in this design. An ANOVA showed that neither the kind of generated counterexample factor nor any of its interactions with the other factors was significant.

4. Note also that our additional retrieval hypothesis predicts that a participant's acceptance ratings of inferences based on conditionals with an equal number of available counterexamples will differ only because of a random error. This random error deviation should be equal in all number of counterexample (CE) groups. A strict reading of the alternative explanation implies that there would be systematic differences in the acceptance rating deviations across the different CE groups: In the one counterexample group most inferences should tend to be accepted, whereas in the next groups there should be an increasing number of inferences that will be rejected (e.g., rating 7 for all conditionals in the one CE group, one conditional with rating 1 in the two CE group, two conditionals with rating 1 in the three CE group, etc.). This should result in some systematic differences in the standard deviation of the means in the different groups. We calculated the standard deviation of the mean inference acceptance in every CE group for every participant. As expected, a MANOVA showed that the MP rating deviations did not significantly differ for the conditionals with one, two, three, or four disablers. Likewise, AC rating deviations did not differ in the successive number of alternatives groups. This supports the additional retrieval explanation of the graded AC and MP trends. However, we did observe differences for the MT [Rao $R(3,13) = 5.17$, $p < .015$] and DA [Rao $R(3,12) = 2.85$, $p < .09$] inferences. The mere fact that there are rating deviations is consistent with the alternative explanation of the DA and MT trends.

APPENDIX
The Conditionals and Counterexamples Adopted for Experiment 1 (Translated From Dutch)

Alternatives

1. If An turns on the air conditioner, then she feels cool.
But,
If An takes off some clothes, she will also feel cool
If An opens a window, she will also feel cool
If An takes a shower, she will also feel cool
If An turns on the fan, she will also feel cool

2. If fertilizer is put on plants, then they grow quickly
But,
If the plants are well watered, they will also grow quickly
If the plants get enough sunlight, they will also grow quickly
If the plants are put in a fertile soil, they will also grow quickly
If the plants are naturally fast growers, they will also grow quickly

3. If Mark studies hard, then he does well on the test
But,
If Mark is cribbing, he will also do well on the test
If the test is easy, he will also do well on the test
If Mark is lucky, he will also do well on the test
If Mark is very smart, he will also do well on the test

4. If the brake is depressed, then the car slows down
But,
If the car is driving uphill, the car will also slow down
If you take your foot of the accelerator, the car will also slow down
If you run out of gas, the car will also slow down
If the car is involved in a collision, the car will also slow down

5. If water is poured on the campfire, then the fire goes out
But,
If the fire dies out, the fire will also go out
If the fire is smothered with sand, the fire will also go out
If it rains, the fire will also go out
If there's a lot of wind, the fire will also go out

APPENDIX (Continued)

Disablers

1. If John studies hard, then he does well on the test
But,
If the test is very hard, he will not do well on the test
If John does not concentrate, he will not do well on the test
If John is not smart enough, he will not do well on the test
If John studied the wrong subject, he will not do well on the test

2. If the match is struck, then it lights
But,
If the match is wet, the match will not light
If the match is not struck hard enough, the match will not light
If the matchbox pad is worn, the match will not light
If the match was already used, the match will not light

3. If Jenny turns on the air conditioner, then she feels cool
But,
If the air conditioner is broken, then she will not feel cool
If Jenny has a fever, then she will not feel cool
If the heating is on, then she will not feel cool
If it is very hot weather, then she will not feel cool

4. If fertilizer is put on plants, then they grow quickly
But,
If the plants are not getting enough water, they will not grow quickly
If the plants are dying, they will not grow quickly
If the plants are not getting enough sunlight, they will not grow quickly
If the wrong type of fertilizer is applied, the plants will not grow quickly

5. If the ignition key is turned, then the car starts
But,
If the engine is broken, the car will not start
If the wrong key is used, the car will not start
If the fuel tank is empty, the car will not start
If the key is not turned far enough, the car will not start

(Manuscript received May 29, 2002;
revision accepted for publication January 10, 2003.)