THINKING & REASONING, 2005, 11 (2), 123–150

Psychology Press
Taylor & Francis Group

Working memory and counterexample retrieval for causal conditionals

Wim De Neys, Walter Schaeken, and Géry d'Ydewalle
University of Leuven, Belgium

The present study is part of recent attempts to specify the characteristics of the counterexample retrieval process during causal conditional reasoning. The study tried to pinpoint whether the retrieval of stored counterexamples (alternative causes and disabling conditions) for a causal conditional is completely automatic in nature or whether the search process also demands executive working memory (WM) resources. In Experiment 1, participants were presented with a counterexample generation task and a measure of WM capacity. We found a positive relation between search efficiency, as measured by the number of generated counterexamples in limited time, and WM capacity. Experiment 2 examined the effects of a secondary WM load on the retrieval performance. As predicted, burdening WM with an attention-demanding secondary task decreased the retrieval efficiency. Both low and high spans were affected by the WM load but load effects were less pronounced for the most strongly associated counterexamples. Findings established that in addition to an automatic search component, the counterexample retrieval draws on WM resources.

The ability to think conditional "If, then" thoughts is considered one of the cornerstones of our mental equipment. When people reason with meaningful, causal conditionals in daily life (e.g., "If the brake is pushed, then the car slows down") they rely on stored background knowledge about the conditional. Most people will, for example, remember that cars can also slow down when the gear is shifted or that in the event of black ice the car might start slipping instead of slowing down. Retrieval of this kind of background knowledge, also known as counterexamples, has a profound impact on the inferences people draw. Over the last 20 years, research on the

Correspondence should be addressed to Wim De Neys, Lab Experimentele Psychologie, K.U. Leuven, Tiensestraat 102, 3000 Leuven, Belgium.
Email: Wim.Deneys@psy.kuleuven.ac.be

Preparation of the manuscript was supported by grants from the Fund for Scientific Research – Flanders (FWO – Vlaanderen). The authors would like to thank Jonathan Evans for his comments on an earlier version of this paper.

http://www.tandf.co.uk/journals/pp/13546783.html DOI: 10.1080/13546780442000105

impact of retrieved counterexamples has become one of the major research issues in the conditional reasoning field (Evans, 2002; Evans, Newstead, & Byrne, 1993; Manktelow, 1999).

Research has focused on the impact of two specific types of stored background knowledge or counterexamples: Alternative causes and disabling conditions. An alternative cause (alternative) is a possible cause that can also produce the effect mentioned in the conditional (e.g., shifting the gear down or driving uphill in the introductory example). A disabling condition (disabler) prevents the effect from occurring despite the presence of the cause (e.g., black ice or broken brake pads in the introductory example).

The effects of retrieved counterexamples on the reasoning process are well established. Numerous studies have shown, for example, that when people manage to retrieve an alternative they no longer tend to commit the infamous Denial of the Antecedent fallacy (e.g., Byrne, 1989; Cummins, 1995; Janveau-Brennan & Markovits, 1999; Quinn & Markovits, 1998; Thompson, 1994). That is, once people think of the possibility that "the gear might be shifted" they will be less inclined to conclude that the car will not slow down when they are told that the brake was not depressed. Likewise, when people retrieve a stored disabler they will tend to reject the Modus Ponens inference. That is, given the conditional "If the brake is pushed, then the car slows down" and the additional information that the brake has been pushed, people will be less inclined to conclude that the car will slow down if they think of the fact that "the brake can be broken" (e.g., Bonnefon & Hilton, 2002; Byrne, 1989; Cummins, 1995; De Neys, Schaeken, & d'Ydewalle, 2002b, 2003a; Thompson, 1994). The efficiency of the counterexample search directly determines the extent to which the inferences will be drawn. The more counterexamples that can be retrieved, the less likely it is that the different conclusions will be accepted (see De Neys, Schaeken, & d'Ydewalle, 2003b; Liu, Lo, & Wu, 1996).

Whereas the effects of retrieved counterexamples on the reasoning process are by now fairly well established, it is less clear how the counterexamples are actually retrieved. Current reasoning theories lack a specification of the background knowledge search process (Johnson-Laird & Byrne, 1994; Oaksford & Chater, 2001): A conditional reasoning researcher can nicely spell out what will happen with the reasoning process once you have retrieved a counterexample, but the same researcher faces a hard time explaining how you found that counterexample in the first place. The present study is part of recent research that has begun to focus on a specification of the processing characteristics of the counterexample search mechanism (e.g., Barrouillet, Markovits, & Quinn, 2001; De Neys et al., 2002b, 2003a, 2003b; Markovits & Barrouillet, 2002; Markovits, Fleury, Quinn, & Venet, 1998; Markovits & Quinn, 2002; Quinn & Markovits,

2002). We address a fundamental issue concerning the role of working memory in the search process. As in most studies, we thereby focus on counterexample retrieval for realistic, causal conditionals (i.e., the conditional expresses a familiar link between a cause and effect) that people typically reason with in everyday life.

Working memory (WM) is often conceived as a hierarchically organised system in which specific storage and maintenance components (i.e., the "slave" systems or short-term memory systems, see Engle, Tuholski, Laughlin, & Conway, 1999) subserve a central component responsible for the control of information processing (e.g., Baddeley, 1996; Cowan, 1995; Engle & Oransky, 1999). The crucial controlling component or "central executive" consists of a limited-capacity system that regulates the allocation of attentional resources. People's performance on standard working memory tests primarily reflects central executive capacity (Engle et al., 1999; Kane, Bleckley, Conway, & Engle, 2001). Our working memory investigation also focuses on the role of this central executive. Thus, the terms WM capacity and WM load always refer to the crucial central executive resources (e.g., Engle, 2002; Kane & Engle, 2002).

In reasoning research, the process where background knowledge or counterexamples are accessed is generally conceived as an undemanding, automatic mechanism (e.g., Cummins, 1995; Evans, 2002; Evans & Over, 1996; Newell & Simon, 1972; Stanovich & West, 2000). Influential dual process theories of reasoning (e.g., Evans & Over, 1996; Goel, 1995; Newell & Simon, 1972; Sloman, 1996; Stanovich & West, 2000) have typically attributed the impact of background knowledge and problem content to heuristic or System-1 processing. The dual process theories distinguish two types of reasoning systems (for a review see Evans & Over, 1996; Stanovich & West, 2000). In general, System-1 processes are characterised by a tendency towards an automatic contextualisation of a problem with prior knowledge, whereas System-2 processes decontextualise a problem and allow reasoning according to normative standards. The first, default system (System-1) is assumed to operate automatically whereas the operations of the second system would draw heavily on executive WM resources.

Numerous studies have established the role of WM in System-2 processing (e.g., Barrouillet & Lecas, 1999; Capon, Handley, & Dennis, 2003; Gilhooly, Logie, & Wynn, 1999; Klauer, Stegmaier, & Meiser, 1997; Stanovich & West, 2000, 1998a, 1998b). Stanovich and West (e.g., 1998a, 1998b) also established that in reasoning tasks where System-1 contextualisations trigger the correct answer, reasoning performance does not depend on cognitive ability or WM resources. Hence, this supports the assumed automatic nature of System-1 processing.

In the absence of clear evidence for a WM involvement, the current dominant characterisation of the counterexample search and related

System-1 operations as a purely automatic process might be preferred for reasons of parsimony. However, one should note that although memory studies have established that many forms of memory retrieval are indeed automatic and effortless, some forms do demand executive WM resources for their proper functioning (Conway & Engle, 1994; Kane & Engle, 2000; Moscovitch, 1994, 1995; Rosen & Engle, 1997). Moscovitch (1995) labelled these forms associative and strategic retrieval, respectively. Rosen and Engle posited that memory retrieval typically starts with an associative, automatic spreading of activation. In case of a real strategic search, WM resources would be used next for an active generation of cues to access new instances. The active cue generation would allow a much more efficient retrieval than the passive spreading of activation. Thus, contrary to the general assumption of dual process theories, it might be the case that by contributing to a more efficient retrieval of the necessary background knowledge, WM nevertheless plays a role in System-1 processing.

Consistent with the memory studies, conflicting evidence for the Stanovich and West findings has been reported. One of Stanovich and West's central findings concerns the notorious deontic version of the selection task. In this task participants need to search cases that falsify a deontic rule (e.g., "If a person is drinking beer, the person must be over 18 years of age"). Contrary to standard version with abstract conditionals, our background knowledge about, for example, drinking regulations makes it readily clear that it is necessary to "check youngsters" here, and consequently reasoners perform very well (Evans et al., 1993; Manktlelow, 1999). Although Stanovich and West (1998b) showed that performance on this "System-1 task" was not associated with general cognitive ability, other studies failed to replicate the results and reported positive correlations (e.g., Dominowski & Dallob, 1991; Klaczynski, 2001; Newstead, Handley, Harley, Wright, & Farrelly, 2004).

In a number of studies one can also find somewhat more specific evidence that points to the strategic nature of the counterexample retrieval. For example, De Neys et al. (2003b) observed that the counterexample search was less extended (i.e., fewer counterexamples were retrieved) for inferences that contained negations in the minor premise (i.e., the "denial" inferences). Reasoning theories typically state that the negation processing for denial inferences increases the burden that is put on working memory (e.g., Barrouillet & Lecas, 1999; Braine & O'Brien, 1998; Johnson-Laird & Byrne, 1991). Although the WM burden was not directly measured, the fact that the efficiency of the search process was affected by the increased processing demands suggests that the search also draws on the same, limited WM resources.

In one of their experiments, Markovits and Quinn (2002) measured the time participants needed to retrieve an alternative as indicator of retrieval

efficiency. Interestingly, results showed that the retrieval times of strongly associated alternatives were less good predictors of the retrieval efficiency (as measured by the tendency to accept specific inferences) than the retrieval times of alternatives with a lower associative strength. The associative strength can be conceived as the strength of the connection between the mental representation of a conditional and a stored counterexample in long-term memory (De Neys et al., 2003a; Quinn & Markovits, 1998). Markovits and Quinn observed that the retrieval times for the strongly associated alternatives showed only small inter-individual variation. It is generally assumed that contrary to strategic, WM-dependent cognitive processes there is only little inter-individual variation in the efficiency of automatically operating cognitive processes (e.g., Evans & Over, 1996; Klaczynski, 2001; Sloman, 1996; Stanovich & West, 2000). Although inconclusive, such findings do indicate that a possible contribution of a strategic component to the counterexample retrieval should not be discarded *a priori*.

The current study attempts to settle the issue and presents a direct test of the hypothesis that retrieving stored counterexamples for a causal conditional draws on WM resources. This will pinpoint a fundamental characteristic of the retrieval process. In Experiment 1, we directly assessed the relation between the efficiency of the counterexample search process and WM capacity. If the retrieval of stored counterexamples draws at least partially on the limited WM resources, then individual differences in the amount of available resources should affect the search efficiency: The higher one's WM capacity, the more resources that can be allocated to the search, and the more successful the search should be. As in previous studies (e.g., De Neys et al., 2002b; Janveau-Brennan & Markovits, 1999), we measured search efficiency by looking at the number of counterexamples participants could generate for a set of conditionals in a limited time.

Experiment 2 adopted dual-task methodology to examine the causal nature of the relation between WM capacity and search efficiency. We studied the impact of a secondary executive task load on the generation efficiency. If WM is involved in the counterexample retrieval, burdening WM with a secondary task should reduce the efficacy of the search process.

EXPERIMENT 1

If allocation of WM resources plays a critical role in the retrieval of stored counterexamples from long-term memory, we expect a positive relation between search efficiency and WM capacity. The higher one's WM capacity, the more resources that can be allocated to the search, and the more successful the search should be. Participants in Experiment 1 were given a standard measure of WM capacity (Operation span task; see La Pointe & Engle, 1990) and a generation task in which participants tried to retrieve as

many counterexamples as possible for a set of conditionals in a limited time. In order to examine whether the relation was similar for both types of counterexamples, half of the participants generated disablers while the other half generated alternatives.

Method

Participants

A total of 104 undergraduate students (Mean age = 19.79, *SD* = 3.06) from the University of Leuven (Belgium) participated in the study. Half the students generated disablers in the counterexample generation task, whereas the other 52 participants generated alternatives (see Materials). Participants received course credit or 5 euro for their participation. All participants were native Dutch speakers.

Materials

Counterexample generation task. Participants were requested to generate counterexamples (alternatives or disablers) for a set of eight conditionals. All conditionals expressed familiar causal relations (see Appendix). Item format and instructions were adopted from De Neys et al. (2002b) and Cummins (1995). The following presents an example of the item format in the alternative generation task:

> Rule: If the air conditioner is turned on, then you feel cool.
>
> Fact: You feel cool, but the air conditioner was not turned on.
>
> Please give as many factors as you can that could make this situation possible.

The item format of the disabler generation task was similar except that under the heading "Fact:" would appear "The air conditioner was turned on, but you don't feel cool." Items like these were constructed for each conditional. Each item was presented for 30 s on a computer screen with a black background. The fixed headers ("Rule:", "Fact:", and "Please give ...") appeared in grey letters, the remaining text in yellow ones. After 30 s the background changed colour to red and participants saw the word "STOP" for 2 s. Finally, after the presentation of the text "NEXT ITEM" (white letters/ blue background) for 950 ms, the next item was presented. The items were presented in the same fixed order to all participants. Participants were instructed to say the retrieved counterexamples out loud and to stop generation when "STOP" appeared. The experimenter wrote down the generated counterexamples on a scoring sheet. Item presentation was paused after the fourth item until participants decided to continue.

A different set of conditionals was used for the alternative and disabler generation task. Half of the conditionals in each set were classified in previous generation studies (De Neys et al., 2002b; Dieussaert, Schaeken, & d'Ydewalle, 2002) as having many possible counterexamples (i.e., alternatives for the conditionals in the alternative generation task, disablers for the conditionals in the disabler generation task), while the other half had only a few possible counterexamples.

Task instructions stressed the importance of producing items that were reasonably realistic and different from each other. Participants were instructed that simple variations of the same idea (e.g., for the example above "taking off coat", "taking off shirt", "taking off sweater", "taking off tank top", "taking off cardigan", "taking off waistcoat", "taking off turtleneck") would be scored as a single item and should be avoided. We also told participants they could give brief responses that mentioned only the general core of the retrieved counterexamples (e.g., "shirt off" instead of "maybe it is possible that you feel cool because you took off your shirt ..."). When all items had been presented, participants were asked to comment on responses that could not be readily interpreted by the experimenter.

It will be clear that, as in all our studies, we only adopted familiar causal conditionals. We can assume that all participants (undergraduates) have a very similar counterexample knowledge base for these conditionals (see the appendix of De Neys et al., 2002b, for a complete overview of possible generated counterexamples). This issue is important since if higher spans had simply stored more counterexamples, the larger knowledge base, and not a better operating search process *per se*, could account for a higher number of retrieved counterexamples. Because of the familiar nature of the causal conditionals and limited search time (Rosen & Engle, 1997) such a confound is implausible in the present study. We have some data to illustrate the point. After the experiment, we presented two possible disablers generated in previous studies (i.e., "water not pure" and "pressure/height not normal") to participants who did not generate these counterexamples themselves in the experiment. Participants were simply asked to indicate whether they already knew the disabler or not. Of the 26 (out of 54) participants who did not generate "water not pure" as possible disabler for the conditional "If water is heated to 100°C, then it boils", 24 participants (92%) stated they knew the disabler. Out of 43 participants who did not generate "pressure/height not normal", 38 indicated that they recognised the disabler (88%). These figures are not surprising given that, for example, all our participants studied these disablers and were given experimental demonstrations in their high-school physics courses. Undergraduates have stored the possible counterexamples (i.e., the counterexamples are part of their long-term knowledge base)—the point is that not everyone will be able to retrieve them all on the spot.

Operation span task (Ospan). Participants' working memory capacity was measured using the Ospan, in which they solved series of simple mathematical operations while attempting to remember a list of unrelated words (see La Pointe & Engle, 1990; De Neys, d'Ydewalle, Schaeken, & Vos, 2002a). Participants saw individual operation-word strings on the computer monitor. They read aloud and solved the maths problems, each of which was followed by a one-syllable, high-frequency (Dutch) word. After a set of operation-word strings (ranging from two to six items in length), they recalled the words. For example, a set of three strings might be:

IS (9 ÷ 3) + 2 = 5 ? JOB

IS (5 × 1) – 4 = 2 ? BALL

IS (3 × 4) – 5 = 8 ? MAN

Participants were instructed to begin reading the operation-word pair aloud as soon as it appeared. Pausing was not permitted. After reading the equation aloud, the participant verified whether the provided answer was correct and then read the word aloud. The next operation then immediately appeared. The participant read the next operation aloud and the sequence continued until three question marks (???) cued the participant to recall all of the words from that set. Participants wrote the words on an answer sheet in the order in which they had been presented.

The Ospan score is the sum of the recalled words for all sets recalled completely and in correct order. Three sets of each length (from two to six operation-word pairs) were tested, so possible scores ranged from 0 to 60. Set size varied in the same randomly chosen order for each participant. Thus, the participant could not know the number of words to be recalled until the question marks appeared.

Procedure

All participants were tested individually in a sound-attenuated testing room. The Ospan task was presented after the generation task. The generation protocols were scored by a rater in order to identify unrealistic items and items that were variations of a single idea. The list of accepted counterexamples as judged by the two raters in the study of De Neys et al. (2002b) was provided to clarify the rating task. To get a grasp of the rater's scoring criteria reliability we first asked whether or not the rater agreed with the previous judgements (e.g., should "taking shirt off" and "taking sweater off" be scored as a single variation of "taking clothes off"?). Judgements agreed on more than 93% of the classifications.

Results and discussion

Overall, 6.4% of the generated counterexamples were disallowed by the rater. On average participants generated a total number of 18.3 counterexamples ($SD = 4.56$) for the eight conditionals in the generation task (disablers = 18.65, $SD = 5.28$; alternatives = 17.94, $SD = 3.74$). All participants solved at least 85% of the Ospan operations correctly. Mean Ospan score was 15.94 words recalled correctly ($SD = 8.4$).

Participants generated counterexamples for eight different conditionals. The number of generated counterexamples for each conditional was used as a subscore to compute Cronbach's alpha as a measure of generation task reliability. The alpha coefficient reached .74. Engle et al. (1999) reported an alpha coefficient of .69 for the Ospan. In the present study we obtained a comparable alpha of .63 for the Ospan task. These figures show that the WM capacity and generation efficiency measures were reliable.

A positive correlation between the number of generated counterexamples and working memory capacity, $r = .25$, $n = 104$, $p < .015$, indicated that higher WM capacity was associated with a more efficient counterexample retrieval process. The pattern did not differ for both types of counterexamples (disablers, $r_1 = .24$, alternatives, $r_2 = .27$, $n_1 = 52$, $n_2 = 52$, $p > .43$).

To get a more specific picture we compared generation performance of participants in the upper (high spans) and bottom quartile (low spans) of the WM-capacity distribution. A total of 30 participants were classified as low spans (Ospan score 10 or lower), whereas 25 participants were classified as high spans (Ospan score 20 or higher). These cut-off values are similar to the typical Ospan-distribution quartile cut-offs across many studies (e.g., Kane et al., 2001; Kane & Engle, 2000, 2003). The mean Ospan score for the low spans was 7.77 ($SD = 2.11$) and 27.28 ($SD = 8.13$) for the high spans. Approximately half of the low and high spans had generated disablers (n low = 14; n high = 12) whereas the other half had generated alternatives (n low = 16; n high = 13). WM capacity (high or low) and type of generated counterexample (disabler or alternative) were entered as between-subjects factors in an ANOVA on the total number of generated counterexamples. Results showed a main effect of WM; high spans ($M = 20.09$) generated more counterexamples than low spans ($M = 17.31$), $F(1, 51) = 6.07$, $MSE = 17.29$, $p < .02$. Whether participants generated alternatives ($M = 17.9$, $SD = 3.92$) or disablers ($M = 19.27$, $SD = 4.76$) had no impact on the number of generated counterexamples, $F(1, 51) < 1$, and the difference between high and low spans was similar for participants who generated disablers and alternatives, $F(1, 51) < 1$.

EXPERIMENT 2

In Experiment 1 we found a positive relation between counterexample retrieval and WM capacity. Higher WM span was associated with the capacity to retrieve more alternatives and disablers. Such a pattern is consistent with the claim that in addition to a passive spreading of activation, the counterexample search also involves a more active component that draws on WM capacity: The more resources that can be allocated to the search, the more efficient the search will be. However, a mere correlation is not sufficient to conclude that WM is mediating the search. In Experiment 2 we therefore used dual-task methodology to examine the causal nature of the established relation. This allowed us to exclude possible confounds and to present a more conclusive test. We tested the crucial role of WM capacity in counterexample retrieval by examining the impact of a secondary task load on generation efficiency. If the counterexample search involves an active component that draws on WM capacity, then burdening WM with a secondary task should reduce the efficacy of the retrieval process. A purely automatic search should not be affected by the WM load.

The secondary tapping task was adopted from Kane and Engle (2000) and Moscovitch (1994). These studies showed that tapping a complex, novel sequence (e.g., index finger - ring finger - middle finger - pinkie) put a premium on efficient executive WM functioning, while tapping an often-habitual "cascade" sequence (e.g., pinkie - ring finger - middle finger - index finger) was less or not attention demanding. We therefore asked one group of participants to tap the complex sequence, whereas another group was instructed to tap the simple, cascade sequence.

We decided to use the participants of Experiment 1 as their own controls for the counterexample-generation performance under secondary task conditions. As compared to Experiment 1, we expected the number of generated counterexamples to decrease under complex tapping. The cascade tapping group served as a control group. Retrieval efficiency was expected to remain unaffected because of the non-demanding nature of the cascade tapping.

Based on Experiment 1 we hypothesised that the load effects would be similar for both counterexample types. However, previous studies have speculated on possible differences in the extent that retrieving alternatives and disablers would require WM resources (e.g., Verschueren, De Neys, Schaeken, & d'Ydewalle, 2002). By testing whether the load effects differed for both types of counterexamples, this issue could be tested explicitly.

Furthermore, we also wanted to test whether the expected WM-load effects differed for participants with high and low WM span. Rosen and Engle (1997) had already observed that only high spans were affected by a

dual task load in a category generation task (e.g., generating instances of the category ''animals''). They concluded that low spans relied mostly on an automatic retrieval mechanism in the category generation task. If this is also the case for the counterexample search, we expect that the WM-load effect will interact with span group.

Method

Design

Participants' performance on an initial counterexample generation task (see Experiment 1) served as the baseline for the effect of introducing a secondary WM load (complex or cascade tapping) on counterexample generation (either disablers or alternatives). This constitutes a 2 (secondary WM load or not, within-subjects) $\times$ 2 (counterexample type, between-subjects) $\times$ 2 (tapping type, between-subjects) design.

Participants

All 104 participants of Experiment 1 participated in the present experiment. Of these, 64 were assigned to the complex tapping group, whereas the remaining 40 participants were assigned to the control group that tapped the cascade pattern. Half of the participants in both groups completed the disablers generation task, whereas the other half completed the alternatives generation task. Participants received course credit or were paid for their participation.

Materials

Counterexample generation task. Apart from the fact that we used a different set of conditionals (see Appendix), the generation task was identical to the one used in Experiment 1. For both the disablers and alternatives generation tasks participants generated counterexamples for a set of eight ordinary, causal conditionals. The conditionals were once more selected from the pilot generation studies of De Neys et al. (2002b) and Dieussaert et al. (2002). We made sure that the mean number of counterexamples generated for the selected sets of conditionals used in the present experiment was comparable to the selected sets in Experiment 1 (typically about 24 counterexamples for a set, with 1.5 min generation time per conditional).

Tapping task. A program executed by a second computer collected the finger-tapping data. All participants tapped on the ''V'', ''B'', ''N'', and ''M'' keys on the keyboard of the second computer.

Procedure

We tested each participant individually after a 2-minute break that followed Experiment 1. The first 40 disabler generators and first 40 alternative generators of Experiment 1 were randomly assigned to either the complex or cascade tapping condition. The remaining 24 participants all tapped the complex pattern. We tested a larger number of participants in the complex tapping group in order to have a more powerful test of the possible different complex tapping effects for high and low spans.

Participants generated the same type of counterexamples in Experiment 1 and 2. All participants tapped the pattern with their non-dominant hand (as indicated by self-report). In the cascade condition participants tapped the sequence pinkie/ ring finger/ middle finger/ index finger. Participants in the complex condition were asked to tap the sequence index finger/ ring finger/ middle finger/ pinkie.

The procedure closely paralleled Kane and Engle's (2000) tapping procedure. The experimenter first demonstrated the tapping sequence. Participants were instructed to repeatedly tap the sequence at a "comfortable and consistent rate". Only participants in the complex tapping condition were given strict instructions about tapping accuracy; cascade-tapping participants were told to keep tapping in a natural way.

All participants began with three 30-s practice trials of tapping. Complex-tapping participants received on-line accuracy feedback (a 300 ms, low-pitch tone followed every error) and were told that hearing many tones would indicate they should slow down. Then followed a 60-s practice trial where participants received both on-line accuracy feedback and response time feedback (600 ms, high-pitch tone). Participants received examples of the accuracy and response time tones and their different meaning was explained. The computer determined the feedback cut-off times for each participant individually: During the previous 30-s practice trial, the computer calculated the mean intertap interval and added 150 ms to it. This became the feedback cut-off for the 60-s practice trial. Thus, if any one intertap interval was more than 150 ms slower than the established mean from the prior practice trial, the computer immediately emitted a 600 ms tone.

Note that (unlike Kane & Engle, 2000) we presented participants in the complex tapping condition with accuracy feedback following the 60-s practice trial. Furthermore, pilot work for the present study also indicated that cascade-tapping participants preferred to let their fingers rest on the keyboard after tapping an individual finger and to lift all four fingers together after finishing a complete sequence (i.e., after tapping the index finger). Therefore, our computer program in the cascade condition only recorded the number of times the first finger of the pattern was tapped (to calculate the taps per second we multiplied this number by four). This

allowed participants to tap the cascade sequence in its most natural, habituated form. The pilot work made it clear that participants tapped this pattern rather effortlessly and without errors. Therefore, following Kane and Engle, we did not give accuracy feedback in the cascade condition. In order to force people to keep tapping at a consistent rate we did give response time feedback starting from the 60-s practice trial in the cascade condition. As cut-off we calculated the average time needed to tap the complete sequence in the last 30-s practice trial. If the sequence was tapped more than 600 ms slower than the average from this prior practice trial, the computer emitted a 600 ms tone.

During all tapping practice trials participants were instructed to focus on a fixation cross presented at the centre of a computer screen placed in front of them. Thus, participants could not watch their fingers while tapping.

After a final 30-s practice trial with response time (cascade and complex tapping) and accuracy feedback (only for complex tapping) participants received the instructions for the counterexample generation task. The experimenter orally repeated the instructions of the first generation task. The experimenter then explained that the primary job in the upcoming generation task was to maintain practice tapping speeds throughout, and that tapping should not be compromised to improve retrieval performance.

The counterexample generation task began with a "BEGIN TAPPING" instruction screen. This "baseline tapping" signal remained onscreen for 15 s, during which participants tapped with response time (cascade and complex tapping) and accuracy feedback (for complex tapping only). The feedback cut-off time was calculated from the immediately preceding 30-s trial, i.e., mean preceding intertap interval + 150 ms for complex tapping, and mean preceding sequence interval + 600 ms for cascade tapping. From this point onwards participants always tapped with this feedback cut-off time.

Following the 15-s baseline tapping a signal ("NEXT ITEM", presented for 1 s on a blue background) announced the beginning of the generation task. The generation task was similar to the one used in Experiment 1. Participants continuously tapped the sequence until, after the fourth generation item, the generation task was paused. After the break participants started with 15-s baseline tapping after which the warning signal announced the beginning of the last part of the generation task.

The generation protocols were scored by the same rater as in Experiment 1.

Results

Counterexample generation task

Overall, 6.5% of the generated counterexamples were disallowed by the rater.

Participants were assigned to one of the four groups (tapping type × counterexample type) in the experiment. A first control analysis established that the mean Ospan score of the participants in the different groups did not differ significantly, $F(3, 100) = 2.01$, $MSE = 68.47$.

We ran a 2 (secondary WM load or not) × 2 (counterexample type) × 2 (tapping type) ANOVA on the number of retrieved counterexamples with the secondary WM load as within-subjects factor and tapping type (complex or cascade), and type of generated counterexample (alternatives or disablers) as between-subjects factors. Results are shown in Figure 1.

We found a main effect of the secondary task burden, $F(1, 100) = 38.71$, $MSE = 6.64$, $p < .0001$. As expected, the impact of this secondary task burden depended on the nature of the tapping task, $F(1, 100) = 35.39$, $MSE = 6.64$, $p < .0001$. Tapping the complex sequence resulted in a significant decrease in the number of retrieved counterexamples, $F(1, 100) = 96.29$, $MSE = 6.64$, $p < .0001$. Tapping the habitual cascade sequence, however, did not affect the number of retrieved counterexamples, $F(1, 100) < 1$.[1] The fact that retrieval performance was not affected in the cascade tapping condition establishes that the decline in retrieval efficiency for the complex tapping should not be attributed to specific characteristics of the conditionals in the first and second generation task or to a lack of motivation to generate counterexamples a second time.

The remaining factors and their interactions did not reach significance, all $F(1, 100) < 1$. Thus, whether participants generated alternatives or disablers had no impact on retrieval performance and the WM-load effects did not differ for both types of counterexamples. We consequently dropped the type of counterexample factor in the subsequent analyses.

The next analysis checked whether the complex tapping effects on counterexample generation differed for high and low spans. Following the quartile criteria of Experiment 1, 19 out of the 64 participants who tapped the complex sequence could be classified as high span (Ospan score 20 or higher) and 16 as low span (Ospan score 10 or lower). An ANOVA on the number of generated counterexamples with WM load (no tap vs. complex) as within-subject factor and span group (high vs. low) as between-subjects factor showed a marginal main effect of span group, $F(1, 33) = 3.23$,

[1]To check that the load findings were not affected by the higher power in the complex (vs. cascade) tapping group (64 vs. 40 participants) we repeated the analysis with the first 40 participants who tapped the complex pattern. Results were not affected. There was a main secondary task load effect, $F(1, 76) = 41.15$, $MSE = 5.66$, $p < .0001$, that interacted with tapping type, $F(1, 76) = 37.81$, $MSE = 5.66$, $p < .0001$. The load effect was significant for complex tapping, $F(1,76) = 78.92$, $MSE = 5.66$, $p < .0001$, but not for cascade tapping, $F(1, 76) < 1$. Other effects and interactions were not significant, all $F(1, 76) < 1$.

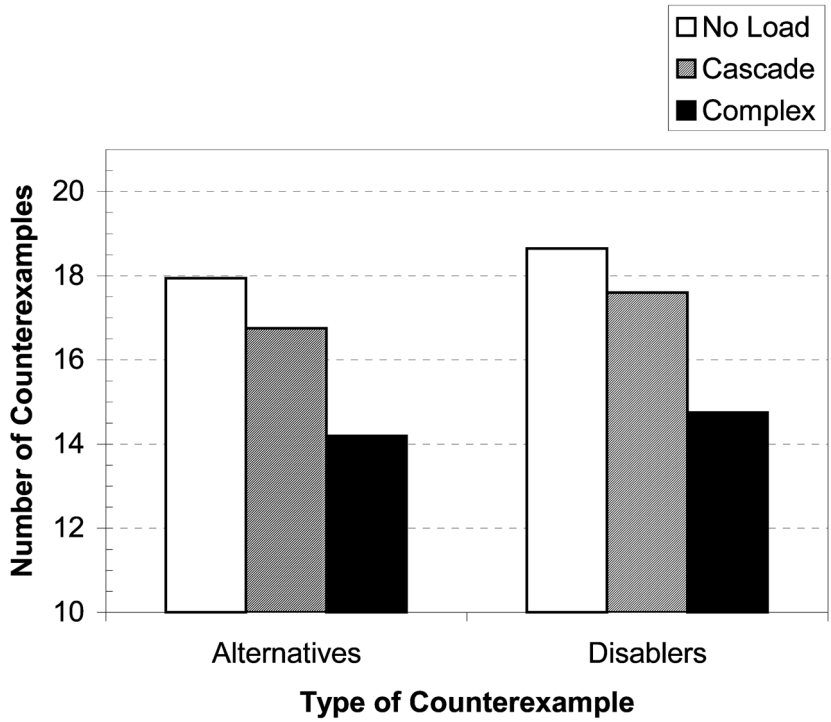

Figure 1. Mean number of generated alternatives and disablers for the eight conditionals in the generation tasks when participants concurrently tapped a cascade or complex finger pattern. The "No Load" condition presents the mean generation performance of participants in the Cascade and Complex condition when there was no dual task imposed.

$MSE = 37.58$, $p < .085$.[2] As Figure 2 indicates, under both no-load and load conditions high spans tended to generate more counterexamples than low spans. However, when WM was burdened by the complex tapping task, the retrieval performance of both span groups declined, $F(1, 33) = 343.85$, $MSE = 7.86$, $p < .0001$. Indeed, there was no sign of an interaction between span group and WM load, $F(1, 33) < 1$.

It should be clear that the emphasis on the involvement of a strategic search component in the present research does not imply that we argue against the contribution of an automatic mechanism. Indeed, following Rosen and Engle (1997) and Markovits and Barrouillet (2002), we assume

[2]Note that the present analysis focused on the crucial complex tapping group. Therefore, only 35 high and low spans were included in the analysis (vs. 55 in Experiment 1). When we looked at the generation performance under secondary task load for the complete sample we replicated the association with WM capacity, $r = .19$, $n = 104$, $p < .05$ found in Experiment 1.

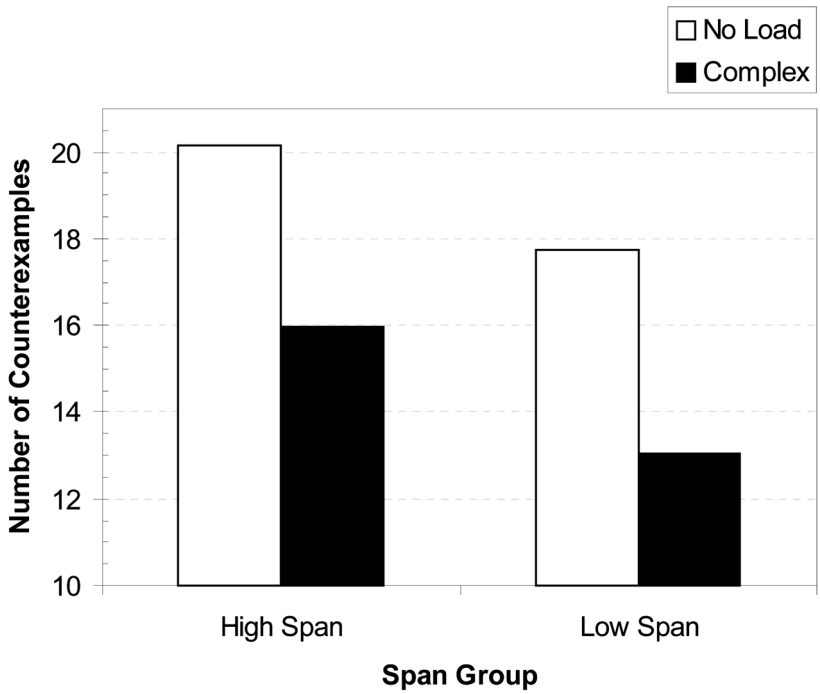

Figure 2. Mean counterexample generation performance of participants classified as High or Low Span when there was no secondary task imposed ("No Load") and when concurrently tapping the complex finger pattern.

that the retrieval starts with an automatic, passive spreading of activation. The point is simply that, except in specific instances, the automatic process will not be very successful for causal conditionals. The allocation of WM resources allows a more efficient retrieval. Note that the specific instances where the passive spreading of activation might be successful will most likely be counterexamples with the highest associative strength (or the lowest activation threshold, e.g., Markovits & Quinn, 2002). The associative strength can be conceived as the strength of the connection between the mental representation of a conditional and a stored counterexample in long-term memory. The higher the associative strength, the easier a counter-example will be retrieved (De Neys et al., 2003a; Quinn & Markovits, 1998). To illustrate this point we specifically looked at the WM-load impact on strongly associated counterexamples. Since the automatic spreading of activation can presumably suffice to activate the most strongly associated counterexamples, we can expect that retrieval of these counterexamples will be less affected by a WM load.

De Neys et al. (2002b) have already reported associative strength values (based on generation frequency) for the counterexamples of the conditionals in the present study. Based on these values a counterexample was classified as weakly or strongly associated. Following De Neys et al., counterexamples with an associative strength smaller than 50% (AS −50%, i.e., the counterexample was generated by less than 50% of the participants in the De Neys et al. sample) were considered weakly associated. A first analysis classified all counterexamples with an associative strength of 50% or higher (AS + 50%) as strongly associated. In a second analysis, we used the more stringent 75% associative strength level (AS + 75%) as criterion (see De Neys et al., 2002b).

We simply compared the number of generated weakly (AS −50%) and strongly (AS + 50% or + 75%) associated counterexamples under no load and complex tapping load. As Figure 3 shows, the load impact was less

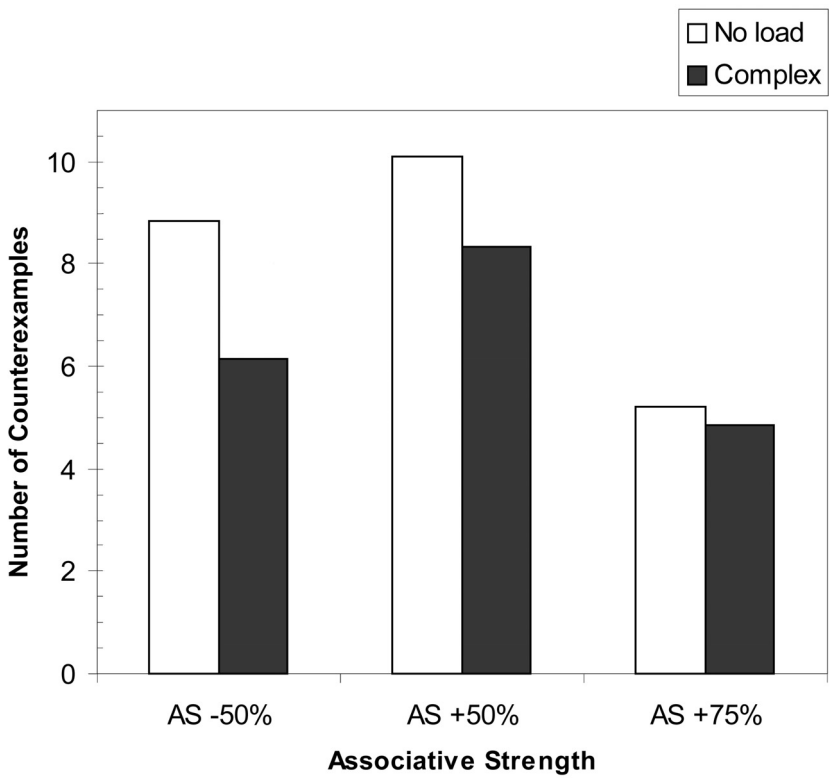

Figure 3. Effects of concurrently tapping the complex finger pattern on the mean number of generated weakly (AS −50%) and strongly (AS + 50% and + 75%) associated counterexamples.

pronounced for the more strongly associated counterexamples. For the weakly associated counterexamples the load resulted in a decrease of 2.7 retrieved counterexamples (compared to the no-load total, a 31% decrease), $F(1, 63) = 46.66$, $MSE = 5.01$, $p < .0001$. For the strongly associated counterexamples of at least 50% AS the load resulted only in a 17% decrease, but the impact remained significant, $F(1, 63) = 34.6.66$, $MSE = 2.9$, $p < .0001$ [AS −50% and + 50% load interaction, $F(1, 63) = 3.53$, $MSE = 3.98$, $p < .065$]. For the strongly associated counterexamples of at least 75% AS, the load impact, a 7% decrease, was no longer significant, $F(1, 63) = 2.12$, $MSE = 2.12$, $p > .15$ [AS −50% and + 75% load interaction, $F(1, 63) = 26.51$, $MSE = 3.27$, $p < .0001$]. As one might expect, these findings indicate that the active, WM-mediated search becomes less important for the more strongly associated counterexamples.

Tapping task

For the tapping task we analysed the mean number of correct taps per second across two relevant tapping periods: The "baseline" period represents the average tapping performance during the two 15-s periods that preceded the presentation of the first and last four items of the generation task. The "counterexample" period represents the average tapping performance during the 240 s (8 items × 30 s) that participants tapped while generating counterexamples.

The two top lines in Table 1 present the tapping performance of participants who tapped the cascade or complex tapping sequence. A 2 (period, within-subjects) × 2 (tapping type, between-subjects) ANOVA showed that cascade tapping was indeed easier than complex tapping, $F(1, 102) = 78.06$, $MSE = 4.11$, $p < .0001$. Comparing the tapping performance

Table 1
Mean number and standard deviations of correct taps per second during baseline-tapping and during concurrent counterexample retrieval

	Period			
	Baseline		*Counterexample*	
Condition	*M*	*SD*	*M*	*SD*
Cascade (n = 40)	5.05	2.14	4.93	2.14
Complex (n = 64)	2.76	0.84	2.13	0.72
Complex				
High spans (n = 19)	2.86	0.90	2.19	0.74
Low spans (n = 16)	2.30	0.55	1.89	0.68

in the baseline and counterexample period further showed that concurrent counterexample generation led to a decline in complex tapping performance, $F(1, 102) = 80.19$, $MSE = 0.15$, $p < .0001$, while cascade tapping was not affected; period × tapping type interaction, $F(1, 102) = 19.78$, $MSE = 0.15$, $p < .0001$.

The two bottom lines in Table 1 compare the complex tapping performance of high- and low-span participants. A 2 (period, within-subjects) × 2 (span group, between-subjects) ANOVA indicated that high spans tended to tap slightly better than low spans, $F(1, 33) = 3.43$, $MSE = 0.94$, $p < .075$. The WM load caused by the concurrent generation of counterexamples resulted in a similar decrease in tapping performance for both high and low spans; main effect of period, $F(1, 33) = 33.39$, $MSE = 0.15$, $p < .0001$; Period × Span interaction, $F(1, 33) = 2.01$, $MSE = 0.15$, $p > .16$. This indicates that high and low spans were not differentially trading-off retrieval and tapping performance.

Discussion

Experiment 1 showed that people with more WM capacity were better at retrieving counterexamples. The present experiment established that retrieval efficiency declines when WM is burdened with an executive attention demanding task. This implies that WM capacity is important for retrieving stored counterexamples.

Results clearly showed that the load effects of complex tapping on the number of generated alternatives and disablers did not differ. Thus, the present findings suggest that retrieving disablers and alternatives is equally demanding for WM.

Putting a load on WM affected retrieval performance of both high- and low-span participants. This implies that, contrary to Rosen and Engle's (1997) findings for category exemplar generation, even for low spans the counterexample retrieval is not completely automatic in nature.

Rosen and Engle (1997) proposed an interesting component model of memory retrieval. They stated that long-term memory retrieval would start with an automatic spreading of activation from a retrieval cue. This component requires little in the way of executive attention and is important for people with both low and high WM span. Both span groups would then use their WM resources to monitor the automatic retrieval to prevent errors and re-access of previously retrieved category instances. When additional WM resources are available these would be used for an active generation of cues to access new instances. The active cue generation results in a more efficient retrieval than the passive spreading of activation.

Rosen and Engle (1997) claimed that in their category generation task only high spans had enough resources to both monitor the retrieval and

generate new cues. Low spans needed all their WM resources for the monitor component. Therefore low spans relied on the more passive spreading of activation for the actual retrieval and were thus less affected by a WM load.

It should be stressed that the present counterexample search results do fit within the general Rosen and Engle (1997) retrieval model. The finding that WM load also affects low spans' counterexample retrieval is not surprising if one takes the different demands of the monitor component in both retrieval tasks into account. Rosen and Engle's participants generated category instances for 10 minutes, but the different load effects were already apparent after the first minute of retrieval. Within the first minute even low spans typically retrieve about 20 different instances for a category like "animal". For the causal conditionals in the present study, however, even high spans are rarely able to retrieve more than five counterexamples for a conditional. Thus, the monitoring component will be much less demanding in the counterexample retrieval case (e.g., keeping track of 5 vs. 20 items). Therefore, both high and low spans will be able to use WM resources to actively generate cues. Because of the higher level of available resources, high spans will nevertheless be more successful in the cue generation.

The importance of the active cue generation for successful counterexample retrieval is also apparent if one looks at the nature of the counterexamples for the causal conditionals adopted in this study. As Markovits and Barrouillet (2002; see also Oaksford & Stenning, 1992) noted, counterexamples for causal conditionals resemble what Barsalou (1983) called "ad hoc" categories. In contrast to common categories (e.g., "animal names") ad hoc categories are less well established in memory. Whereas instances of a common category can be accessed relatively directly, an ad hoc category (e.g., "things to take on a trip" or "things that can stop a car") needs to be reconstructed on-line. An active generation of retrieval cues is more important here for successful retrieval than with common categories.

Although there is clear evidence for the involvement of WM in the retrieval of stored counterexamples for causal conditionals, this does not imply that there is no role for the automatic retrieval process. Our data indicated that the WM-load effects were less pronounced for strongly associated counterexamples. The strongly associated counterexamples have a lower activation threshold and are more easily accessed. Although the passive spreading of activation will not be very successful for causal conditionals in general, the automatic search can suffice for the most strongly associated counterexamples. Consequently, it makes sense that retrieval of these counterexamples is less hindered by a WM load.

GENERAL DISCUSSION

Our findings showed that searching stored counterexamples for a causal conditional is strategic in nature. The correlation and dual-task analyses indicated that the search efficiency is directly mediated by the available WM resources: Participants with higher WM capacity managed to retrieve more counterexamples in a limited time, and fewer counterexamples were retrieved when WM was burdened by an executive attention demanding secondary task.

The present study allows us to extend Rosen and Engle's (1997) memory retrieval model to sketch the elementary components of the counterexample retrieval process: Following Markovits and Barrouillet (2002) we can assume that when drawing conditional inferences, reasoners construct and maintain a mental representation of the premises in working memory. This representation serves as a retrieval cue from which activation will automatically start to spread towards associated counterexamples in long-term memory (Anderson, 1993; Cowan, 2001). The stored counterexamples can be conceived as nodes in a semantic network (Anderson, 1983). A counterexample will be retrieved when a node's activation level crosses a critical threshold. More strongly associated counterexamples have lower retrieval thresholds and will be retrieved more easily. The spreading of activation requires little in the way of executive attention and can suffice to activate the most strongly associated counterexamples. In addition to the passive spreading of activation, executive WM resources will be recruited to monitor the automatic retrieval in order to prevent errors and re-access of previously retrieved counterexamples. Available WM resources will be used next for an active, strategic search to access new counterexamples.

In the present study we looked at the role of WM in an explicit counterexample search task. This allowed us to study the retrieval process directly. One might question, however, whether the present findings can be generalised to counterexample retrieval during actual reasoning. One could claim, for example, that during reasoning people will make no extra effort to allocate WM resources to the search and will stick to the automatic component. Our findings did indeed indicate that the automatic spreading of activation can suffice to retrieve the most strongly associated counter-examples. However, one should take into account that only a limited number of the counterexamples of a conditional are strongly associated (e.g., on average a conditional has less than one counterexample with an AS of + 75%, see De Neys et al., 2002b). It has already been established that during reasoning people typically search and retrieve multiple counter-examples for a conditional (De Neys et al., 2003b). In one of their experiments De Neys et al. (2003b) first measured the exact number of counterexamples participants could retrieve for each conditional in a set and

later invited the participants for an inference task with the same conditionals. Results showed that inference acceptance linearly decreased with every additional counterexample that was available. De Neys et al. (2003b) concluded that in a conditional inference task reasoners do not end the search after retrieval of a single counterexample but take up to four different counterexamples into account. The final reasoning judgement is determined by the exact number of retrieved counterexamples. The present findings imply that such an extended search will draw on WM.

Furthermore, given that we found that search efficiency decreases under executive WM load, we may also expect similar secondary task effects during a conditional reasoning task. Remember that the extent to which a reasoner accepts, for example, the Denial of the Antecedent (DA) and Affirmation of the Consequent (AC) inferences depends on the number of alternatives one can retrieve. The more alternatives that can be retrieved, the less DA and AC are accepted. A less efficient alternative retrieval should thus result in a higher DA and AC acceptance. Based on the present findings we can claim that an executive WM load reduces retrieval efficiency. Therefore, one can expect that imposing an executive attention demanding secondary task during everyday reasoning will result in higher DA and AC acceptance. In a number of unpublished experiments De Neys (2003) did indeed observe such effects.

We wanted to clarify that WM resources are recruited for the retrieval of causal counterexamples. This does not imply that the role of WM in reasoning stops here. It should be clear that there are also other reasoning components that can or will draw on WM resources. For example, the premises of a reasoning problem will have to be processed and mentally stored, and when a counterexample is retrieved an additional representation will need to be maintained. Previous studies have already demonstrated that such storage processes draw on WM (e.g., Barrouillet & Lecas, 1999; Gilhooly et al., 1999; Kyllonen & Christal, 1990; Markovits, Doyon, & Simoneau, 2002; Meiser, Klauer, & Naumer, 2001; Toms, Morris, & Ward, 1993). However, note that these studies typically examined reasoning with abstract material (e.g., ''If square, then circle'', but see Markovits et al., 2002). With abstract material one sidesteps the background knowledge retrieval problem. We are interested in everyday causal reasoning and its central characteristic, the retrieval of stored background knowledge. This component has not yet been studied (or rather, has been excluded from study). With the present experiments we tried to clarify a previously neglected role of WM: WM resources will not only be recruited for model or information maintenance but also for the retrieval of stored counter-examples. Since reasoning theories make reference to WM capacity, this process should be taken into account. Besides storage purposes, sound reasoning can also require WM resources for retrieval.

Interestingly, the present study even points to a further possible WM role. The secondary task results indicated that the most strongly associated counterexamples will automatically be activated. Now, when the automatic search process activates a disabler this will result in the rejection of the Modus Ponens (MP) and Modus Tollens (MT) inferences in a deductive reasoning task. However, in standard logic MP and MT are valid and should be accepted. In the case of such a "belief–logic" conflict, WM resources may be recruited for an active inhibition of the background knowledge (e.g., De Neys, 2003; Gilinsky & Judd, 1994; Kokis, Macpherson, Toplak, West, & Stanovich, 2002). This makes sense from a WM perspective, since the inhibition of information deemed inappropriate is considered as one of the key executive functions (e.g., Baddeley, 1996; Engle et al., 1999). Such a proposal remains speculative of course,[3] and will need to be tested properly.

Our findings question the basic assumptions of dual process theories. The involvement of WM resources in the retrieval of relevant background knowledge about causal counterexamples conflicts with the general characterisation of System-1 processing as automatic and effortless. The point is that any automaticity claim will need to take the nature of the retrieved information into account. Dual process theories need to differentiate among different types of background knowledge retrieval and pragmatic reasoning processes. System-1 cannot simply be conceptualised as a residual system that unitarily handles all content and belief processing. It was argued that the WM involvement in the counterexample retrieval for the popular causal conditionals is tied to the "ad hoc" (Barsalou, 1983) nature of the retrieved information. In case of retrieval of more common categories such as class-based information (e.g., "If an animal is a cow, then it has four legs", see Markovits & Barrouillet, 2002) the System-1 mediation might be completely automatic. The issue is also apparent, for example, in studies of the believability effect in syllogistic reasoning. Two conclusions might both be empirically false and equally unbelievable (e.g., "All boys are females" and "All Canadian towns lie in the southern hemisphere") but assessing the empirical falsity might merely involve a direct category mismatch detection in one case, whereas in the other it might involve a more active geographical knowledge retrieval. Dual process theories (and reasoning studies in general) need to take this possible diversity into account.

Finally, we note that our work makes no claims about the nature of the elementary representation of a conditional and the further inferential

[3]A possible problem is that the status of standard logic as a normative system for everyday, causal conditional reasoning is heavily debated (e.g., Edgington, 1995; Evans, 2002). If one were to adhere to a different normative system there might simply be no conflict and consequently no ground for an inhibition process.

processes that will operate on it (e.g., these might amount to additional mental model construction, adjusting probabilistic parameters, or specific rule selection). Our research respects a reasoning theory neutrality here. Reasoning theories like mental models theory (Johnson-Laird & Byrne, 1991, 2002), mental logic (Braine & O'Brien, 1998; Politzer & Bourmaud, 2002; Rips, 1994), and the probabilistic approach (Oaksford & Chater, 1998, 2001; Oaksford, Chater, & Larkin, 2000) all specify (to some degree) what happens with a retrieved counterexample. However, so far, no theory has paid attention to the actual retrieval process. Our research programme focuses on this general shortcoming and takes a neutral stance concerning the traditional debate about the nature of the inferential processes. Basically, a specification of the search process characteristics can be incorporated in all the different reasoning theories. Indeed, all theorists should benefit from the present results for a further development of their accounts.

Manuscript received 6 November 2003
Revised manuscript received 7 May 2004

REFERENCES

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

Anderson, J. R. (1993). *Rules of the mind*. Hillsdale, NJ: Lawrence Erlbaum Associates Inc.

Baddeley, A. D. (1996). Exploring the central executive. *Quarterly Journal of Experimental Psychology*, *49*A, 5–28.

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

Barrouillet, P., Markovits, H., & Quinn, S. (2001). Developmental and content effects in reasoning with causal conditionals. *Journal of Experimental Child Psychology*, *81*, 235–248.

Barsalou, L. W. (1983). Ad hoc categories. *Memory & Cognition*, *11*, 211–227.

Bonnefon, J-F., & Hilton, D. J. (2002). The suppression of modus ponens as a case of pragmatic preconditional reasoning. *Thinking and Reasoning*, *8*, 21–40.

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

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

Capon, A., Handley, S., & Dennis, I. (2003). Working memory and reasoning: An individual differences perspective. *Thinking and Reasoning*, *9*, 203–244.

Conway, A. R. A., & Engle, R. W. (1994). Working memory and retrieval: A resource-dependent inhibition model. *Journal of Experimental Psychology: General*, *123*, 354–373.

Cowan, N. (1995). *Attention and memory: An integrated framework*. New York: Oxford University Press.

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences*, *24*, 87–185.

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

De Neys, W. (2003). *In search of counterexamples: A specification of the memory search process for stored counterexamples during conditional reasoning*. Unpublished doctoral dissertation, University of Leuven, Belgium.

De Neys, W., d'Ydewalle, G., Schaeken, W., & Vos, G. (2002a). A Dutch, computerized, and group administrable adaptation of the operation span test. *Psychologica Belgica*, *42*, 177–190.

De Neys, W., Schaeken, W., & d'Ydewalle, G. (2002b). 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*, *42*, 177–190.

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

Dieussaert, K., Schaeken, W., & d'Ydewalle, G. (2002). The relative contribution of content and context factors on the interpretation of conditionals. *Experimental Psychology*, *49*, 181–195.

Dominowski, R. L., & Dallob, P. I. (1991, September). *Reasoning abilities, individual differences, and the four card problem*. Paper presented to the British Psychological Society Cognitive Section Conference, Oxford, UK.

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

Engle, R. W. (2002). Working memory capacity as executive attention. *Current Directions in Psychological Science*, *11*, 19–23.

Engle, R. W., & Oransky, N. (1999). Multi-store versus dynamic models of temporary storage in memory. In R. J. Sternberg (Ed.), *The nature of cognition*. Cambridge, MA: MIT Press.

Engle, R. W., Tuholski, S. W., Laughlin, J. E., & Conway, A. R. A. (1999). Working memory, short-term memory, and general fluid intelligence: A latent-variable approach. *Journal of Experimental Psychology: General*, *128*, 309–331.

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

Evans, J. St. B. T., Newstead, S. E., & Byrne, R. M. J. (1993). *Human reasoning: The psychology of deduction*. Hove, UK: Lawrence Erlbaum Associates Ltd.

Evans, J. St. B. T., & Over, D. E. (1996). *Rationality and reasoning*. Hove, UK: Psychology Press.

Gilhooly, K. J., Logie, R. H., & Wynn, V. (1999). Syllogistic reasoning tasks, working memory, and skill. *European Journal of Cognitive Psychology*, *11*, 473–498.

Gilinsky, A., & Judd, B. B. (1994). Working memory and bias in reasoning across the life span. *Psychology and Aging*, *9*, 356–371.

Goel, V. (1995). *Sketches of thought*. Cambridge, MA: MIT Press.

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

Johnson-Laird, P. N., & Byrne, R. M. J. (1991). *Deduction*. Hove, UK: Lawrence Erlbaum Associates Ltd.

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

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

Kane, M. J., Bleckley, M. K., Conway, A. R. A., & Engle, R. W. (2001). A controlled-attention view of working memory capacity. *Journal of Experimental Psychology: General*, *130*, 169–183.

Kane, M. J., & Engle, R. W. (2000). Working-memory capacity, proactive interference, and divided attention: Limits on long-term memory retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*, 336–358.

Kane, M. J., & Engle, R. W. (2002). The role of the prefrontal cortex in working-memory capacity, executive attention, and general fluid intelligence: An individual-differences perspective. *Psychonomic Bulletin & Review*, *9*, 637–671.

Kane, M. J., & Engle, R. W. (2003). Working-memory capacity and the control of attention: The contributions of goal neglect, response competition, and task set to Stroop interference. *Journal of Experimental Psychology: General*, *132*, 47–70.

Klaczynski, P. A. (2001). Analytic and heuristic processing influences on adolescent reasoning and decision making. *Child Development*, *72*, 844–861.

Klauer, K. C., Stegmaier, R., & Meiser, T. (1997). Working memory involvement in propositional and spatial reasoning. *Thinking and Reasoning*, *3*, 9–47.

Kokis, J. V., Macpherson, R., Toplak, M. E., West, R. F., & Stanovich, K. E. (2002). Heuristic and analytic processing: Age trends and associations with cognitive ability and cognitive styles. *Journal of Experimental Child Psychology*, *83*, 26–52.

Kyllonen, P. C., & Christal, R. E. (1990). Reasoning ability is (little more than) working memory capacity. *Intelligence*, *14*, 389–433.

La Pointe, L. B., & Engle, R. W. (1990). Simple and complex word spans as measures of working memory capacity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *16*, 1118–1133.

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

Manktelow, K. I. (1999). *Reasoning and thinking*. Hove, UK: Psychology Press.

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

Markovits, H., Doyon, C., & Simoneau, M. (2002). Individual differences in working memory and conditional reasoning with concrete and abstract content. *Thinking and Reasoning*, *8*, 97–107.

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

Meiser, T., Klauer, K. C., & Naumer, B. (2001). Propositional reasoning and working memory: The role of prior training and pragmatic content. *Acta Psychologica*, *106*, 303–327.

Moscovitch, M. (1994). Cognitive resources and dual-task interference effects at retrieval in normal people: The role of the frontal lobes and medial temporal cortex. *Neuropsychology*, *8*, 524–534.

Moscovitch, M. (1995). Models of consciousness and memory. In M. S. Gazzaniga (Ed.), *The cognitive neurosciences* (pp. 1341–1356). Cambridge, MA: MIT Press.

Newell, A., & Simon, H. A. (1972). *Human problem solving*. Englewood Cliffs, NJ: Prentice-Hall.

Newstead, S. E., Handley, S. J., Harley, C., Wright, H., & Farrelly, D. (2004). Individual differences in deductive reasoning. *Quarterly Journal of Experimental Psychology*, *57*A, 33–60.

Oaksford, M., & Chater, N. (1998). *Rationality in an uncertain world: Essays on the cognitive science of human reasoning*. Hove, UK: 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, and Cognition*, *26*, 883–899.

Oaksford, M., & Stenning, K. (1992). Reasoning with conditionals containing negated constituents. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *18*, 835–854.

Politzer, G., & Bourmaud, G. (2002). Deductive reasoning from uncertain conditionals. *British Journal of Psychology*, *93*, 345–381.

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 and Reasoning*, *8*, 179–191.

Rips, L. J. (1994). *The psychology of proof: Deductive reasoning in human thinking*. Cambridge, MA: MIT press.

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

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

Stanovich, K. E., & West, R. F. (1998a). Individual differences in framing and conjunction effects. *Thinking and Reasoning*, *4*, 289–317.

Stanovich, K. E., & West, R. F. (1998b). Cognitive ability and variation in selection task performance. *Thinking and Reasoning*, *4*, 193–230.

Stanovich, K. E., & West, R. F. (2000). Individual differences in reasoning: Implications for the rationality debate. *Behavioral and Brain Sciences*, *23*, 645–726.

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

Toms, M., Morris, N., & Ward, D. (1993). Working memory and conditional reasoning. *Quarterly Journal of Experimental Psychology*, *46*A, 679–699.

Verschueren, N., De Neys, W., Schaeken, W., & d'Ydewalle, G. (2002). Working memory capacity and the nature of generated counterexamples. *Proceedings of the Annual Conference of the Cognitive Science Society*, *24*, 914–999.

APPENDIX

Table A1
The conditionals for the counterexample generation tasks of Experiments 1 and 2

ALTERNATIVES GENERATION TASK

Experiment 1:
If water is heated to 100°C, then it boils.
If Mark reads without his glasses, then he gets a headache.
If Ben frequently inhales the smoke of cigarettes, then he gets lung cancer.
If Steven goes in for sports, then he loses weight.
If the brake is depressed, then the car slows down.
If the trigger is pulled, then the gun fires.
If water is poured on the campfire, then the fire goes out.
If Jan consumes alcohol, then he gets drunk.

Experiment 2:
If fertilizer is put on plants, then they grow quickly.
If the gong is struck, then it sounds.
If Jenny turns on the air conditioner, then she feels cool.
If Tom grasps the glass with his bare hands, then his fingerprints are on it.
If John studies hard, then he does well on the test.
If the match is struck, then it lights.
If Bart's food goes down the wrong way, then he has to cough.
If the ignition key is turned, then the car starts.

DISABLERS GENERATION TASK

Experiment 1:
If Jenny turns on the air conditioner, then she feels cool.
If water is heated to 100°C, then it boils.
If Jan consumes alcohol, then he gets drunk.
If John studies hard, then he does well on the test.
If Marry jumps into the swimming pool, then she gets wet.
If the match is struck, then it lights.
If the car is out of gas, then it stalls.
If the gong is struck, then it sounds.

Experiment 2:
If fertilizer is put on plants, then they grow quickly.
If Bart's food goes down the wrong way, then he has to cough.
If the trigger is pulled, then the gun fires.
If Tom grasps the glass with his bare hands, then his fingerprints are on it.
If Andy eats a lot of candy, then he gets cavities.
If the apples are ripe, then they fall from the tree.
If the ignition key is turned then the car starts.
If water is poured on the campfire, then the fire goes out.